Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie

Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki
Katedra Automatyki

PRACA MAGISTERSKA

MARIUSZ SASKO

LINGWISTYCZNY SYSTEM DEFINICYJNY WYKORZYSTUJĄCY KORPUSY TEKSTÓW ORAZ ZASOBY INTERNETOWE.

PROMOTOR:
dr Adrian Horzyk

Kraków 2010

University of Science and Technology in Krakow

Faculty of Electrical Engineering, Automatics, Computer Science and Electronics
Department of Automatics

MASTER OF SCIENCE THESIS

MARIUSZ SASKO

LINGUISTIC DEFINING SYSTEM USING TEXT CORPUSES AND INTERNET SOURCES.

SUPERVISOR:
Adrian Horzyk Ph.D

Krakow 2010

OŚWIADCZENIE AUTORA PRACY

Oświadczam, świadomy odpowiedzialności karnej za poświadczenie nieprawdy, że niniejszą pracę dyplomową wykonałem osobiście i samodzielnie, i nie korzystałem ze źródeł innych niż wymienione w pracy.

. . . . . .

Serdecznie dziękuję dr Adrianowi Horzykowi za pomoc w napisaniu pracy oraz udostępnienie materiałów.

Spis treści

1. Wstęp

Geneza

Powstanie i dynamiczny rozwój Internetu w XX wieku oraz wszechobecna komputeryzacja bardzo przyspieszyły rozwój społeczeństwa informacyjnego, w którym głównym towarem mającym największą wartość jest informacja. Wraz z rozwojem Internetu powstał ogólnodostępny otwarty system informacyjny o nazwie **World Wide Web**, którego podstawowym zadaniem jest publikowanie informacji w postaci dokumentow HTML. Pozwala on każdemu bez graniczeń zamieszczać nowe strony w sieci, które automatycznie stają się dostępne dla wszystkich użytkowników. Spowodowało to bardzo szybki wzrost popularności Internetu oraz dynamiczny przyrost informacji dostępnych w sieci.

Dzięki rozwojowi technologii informacyjnej, (eng. Information Technology) człowiek uzyskał dostęp do narzędzi, za pomocą których może szybko pozyskiwać nowe informacje, selekcjonować je, analizować, przetwarzać, zarządzać nimi oraz przekazywać je innym ludziom na bardzo duże odległości w bardzo krótkim czasie.

Ogromna ilość informacji zgromadzonych w Internecie spowodowała konieczność stworzenia wyszukiwarek, które umożliwiają dostęp do pewnych stron, bez znajomości ich adresów, na podstawie wprowadzonej kwerendy. W 2008 roku inżynierowie firmy Google ogłosili, że ich wyszukiwarka zawiera adresy około 1 biliona unikalnych stron internetowych. Ciągle jednak prowadzone są badania nad szybszymi sposobami dostepu do informacji uzyskanymi z sieci Internet.

Założenia pracy

Celem niniejszej pracy jest skonstruowanie systemu posiadającego wyspecjalizowany algorytm ekstrakcji informacji z tekstu w celu stworzenia definicji pewnego znaczenia. Tworzony system w założeniu powinien samodzielnie wyszukiwać w Internecie teksty zawierające istotne dane w postaci ciągów wyrazów, budować z nich korpus, a następnie ekstrahować z niego informacje i tworzyć bazę wiedzy. Kolejno na tej podstawie

modelować sensowną definicję. Praca jest wykonywana dla języka polskiego.

Najważniejszym elementem pracy jest baza wiedzy, którą system ma sam rozbudowywać. Informacje mają zostać zapisane w postaci specjalnych asocjacji, czyli połączeń między słowami, które będą niosły dodatkowe informacje. Taka forma ma pozwalać na nieustanne rozbudowywanie wiedzy o kolejne słowa. W efekcie ma powstać rozbudowana sieć połączeń, powiązanych logicznie oraz syntaktycznie ze sobą. Kolejno na bazie powstałej sieci ma powstać opis znaczenia, zbudowany przez specjalistyczny algorytm.

Warto również tutaj wspomnieć, że praca jest próbą przeniesienia wyszukiwania informacji z Internetu, na kolejny znacznie wyższy poziom. Oczywiście należy pamiętać, że pewnych problemów z pewnością nie uda się tutaj rozwiązać, bo jest to zagadnienie bardzo skomplikowane i porusza wiele problemów lingwistycznych.

Zawartość pracy

Niniejsza praca zawiera opis metodologii stosowanej do rozwiązania powyższego problemu. Całość została podzielona na cztery rozdziały.

- **Rozdział 1** to wstęp, opisuje założenia niniejszej pracy.
- **Rozdział 2** zawiera wprowadzenie w tematykę pracy i opowiada o lingwistyce komputerowej.
- **Rozdział 3** prezentuje narzędzia niezbędne do tworzenia algorytmów lingwistycznych, oraz opisuje obecnie stosowane metody ekstrakcji informacji.
- **Rozdział 4** to opis algorytmu, który powstał w ramach niniejszej pracy. Zawiera on także prezentację i opis powstałej aplikacji.
- **Rozdział 5** zawiera opis oraz porównanie uzyskanych wyników.
- **Rozdział 6** to zakończenie i podsumowanie.

Praca powstała pod nadzorem promotora dr Adriana Horzyka, który podsunął wiele ciekawych pomysłów, które okazały się bardzo pomocne w realizowaniu niniejszego tematu.

2. Wprowadzenie w problematykę pracy

2.1. Lingwistyka

Lingwistyka czyli językoznawstwo, jest to dział nauk humanistycznych, podejmujący zagadnienia dotyczące budowy, istoty oraz przekształceń zachodzących w języku. Obejmuje ona także rozważania nad powstawaniem, rozwojem i funkcjonowaniem języka, bada zależności między poszczególnymi językami. Dla konkretnych kierunków przedmiot rozważań lingwistycznych zawęża się do bardziej szczegółowych aspektów. Wśród jej licznych działów wymienić można przede wszystkim gramatykę, semantykę, składnię, fonetykę oraz fonologię.

Celem językoznawstwa jest poznawanie języka jako metody opisu otaczającej rzeczywistości. Prócz wartości poznawczych, dla lingwistów ważne jest to, co z taką wiedzą można również zrobić prócz jej posiadania.

Badania lingwistyczne dotyczą zarówno języka mówionego jak i pisanego. Jednak tam, gdzie wykorzystywane są metody lingwistyki informatycznej i korpusowej, język zapisany w postaci tekstu, jest często o wiele wygodniejszy, bo pozwala na przetwarzanie bardzo dużej ilości danych. Ciężko jest przecież znaleźć albo zbudować wielkie korpusy języka mówionego. Są one zazwyczaj sporządzane na piśmie. Przetwarzanie języka mówionego jest dodatkowo o wiele trudniejsze, bo komputer, poza samą analizą wypowiedzi, musi wcześniej przekonwertować ją do postaci akceptowalnej i zrozumiałej dla siebie, co jest zadaniem wieloetapowym i bardzo skomplikowanym [10][19].

Na pograniczu lingwistyki wykształciły się także dziedziny interdyscyplinarne, do których należy zaliczyć nowoczesny kierunek, którym jest **lingwistyka komputerowa**.

2.1.1. Język naturalny

Jest to jedno z najważniejszych pojęć używanych w pracy. Jego definicję można przedstawić następująco:

Język naturalny jest sposobem opisu otaczającego świata i służący do komunikacji pomiędzy ludźmi. Powstał na drodze historycznych przemian, rozwoju oraz ewolucji i jest charakterystyczny dla danej grupy etnicznej lub narodowej.

Język naturalny opisywany jest za pomocą dwóch struktur:

- **semantyka leksykalna** - czyli zbiór słów, które są stosowane w danym języku lub które mają jakieś znaczenie,
- **gramatyka** - znormalizowany system zasad, opisujący sposoby tworzenia słów z liter alfabetu, zdań i wypowiedzi.

2.1.2. Korpus tekstu

"Korpus to dowolny zbiór tekstów, w którym czegoś szukamy."[26] Do celów niniejszej pracy, za korpus tekstu będziemy uważać zbiór zdań, zgromadzonych w bazie danych, niosących pewną informację. Z takiego korpusu powstała aplikacja, będzie ekstrahować słowa tematycznie powiązane z definiowanym pojęciem.

2.1.3. Lingwistyka Komputerowa oraz Natural Language Procesing(NLP)

Dynamiczny rozwój informatyki, wszechobecna komputeryzacja i automatyzacja procesów, spowodowały potrzebę rozwoju badań nad językiem naturalnym, przy wykorzystaniu komputerów. Człowiek zaczął podejmować próby zbudowania inteligentnych robotów, wzorowanych na swoje podobieństwo i wykorzystujące język naturalny do porozumiewania się.

Te działania spowodowały powstanie nowego, interdyscyplinarnego kierunku badań z pogranicza informatyki, lingwistyki i matematyki nazwanego **lingwistyką komputerową** lub **inżynierią lingwistyczną**.

Jest to dziedzina naukowa zajmująca się badaniem zagadnień językowych z informatycznego punktu widzenia. Lingwistyka komputerowa zajmuje się dostarczaniem modeli informatycznych, związanych z różnymi rodzajami zjawisk językowych.

Często jest ona utożsamiana z NLP, czyli przetwarzaniem języka naturalnego. Ogólnie można przyjąć, że jest to podejście właściwe, jednak należy pamiętać, że nie chodzi tu wyłącznie o implementację algorytmów. Dla NLP istotne są także procesy tworzenia gramatyk formalnych. Analiza języka naturalnego może zajmować się zarówno

tekstem, jak i mową, ale prace nad syntezą mowy rozwinęły się jako oddzielny dział.

Główne problemy, jakie podejmują lingwistyka komputerowa oraz NLP, pokrywają się ze sobą i obejmują: automatyzację analizy, rozumienia, tłumaczenia i generowania języka naturalnego przez komputer.

Stwierdzenie, co powinno znajdować się pod definicją "rozumienia", stanowi ciekawy problemem filozoficzny. Właściwe wydaje się założenie, że umiejętność ta wymaga obszernej wiedzy o świecie zewnętrznym i zaawansowanych zdolności do przekształcania go. Problem rozumienia języka naturalnego nazywany jest często problemem AI-complete.

Zadania realizowane przez lingwistykę komputerową

Lingwistyka komputarowa podejmuje następujące problemy:

- tworzenie formalizmów modelujących różne aspekty języków naturalnych,
- automatyczne budowanie i udostępnianie wiedzy o poszczególnych językach,
- tworzenie algorytmów i metod do przetwarzania wypowiedzi językowych,
- ewaluacja systemów lingwistycznych,
- testowanie hipotez dotyczących mowy i języka,
- rozpoznawanie możliwości opisowych modeli i formalizmów,
- badanie możliwości automatycznej nauki i kategoryzacji podzbiorów językowych,
- transformacji języka pisanego i mówionego.

Naukowcy zajmujący się tą dziedziną wiedzy prowadzą badania nad budową systemów komputerowych, które będą intensywnie wykorzystywać język naturalny. Sprzęt komputerowy nieustannie się rozwija, komputery posiadają coraz to większą moc obliczeniową, stale powstają nowe narzędzia, a to pozwala na budowę bardziej skompilowanych algorytmów.

Zastosowania lingwistyki komputerowej:

Można sobie wyobrażać, jak bardzo zmieniłoby się postrzeganie komputerów, gdyby zaczęły one wykorzystywać i w pełni potrafiły interpretować język naturalny. Nie ma wątpliwości, że wpłynęłoby to bardzo pozytywnie na podejście do komputerów i informatyki, bo drastycznie zwiększyłoby komfort pracy z maszynami.

Osiągnięcia lingwistyki komputerowej mogą w przyszłości zostać wykorzystane do tworzenia zaawansowanych systemów:

- opracowanie tłumaczy maszynowych oraz słowników kontekstowych,
- systemy do automatycznej korekcji tekstów,
- budowa automatycznych systemów dialogowych, umożliwiających komunikację pomiędzy człowiekiem a maszyną,
- budowa systemów wyszukujących i przetwarzających informacje, ekstrakcja informacji z dużych nieuporządkowanych źródeł np. Internet, archiwa elektroniczne itp.,
- budowa systemów generujących tekst, a także rozpoznających i syntetyzujących mowę ciągłą,
- budowa internetowych systemów konwersacyjnych np. rezerwacja miejsc przez rozmowę z automatem,
- budowa interaktywnych systemów multimedialnych, interfejsy językowe do systemów informatycznych,
- budowa automatycznych systemów publikacyjnych.

Niestety, większości nie udało się jeszcze zbudować, a te które już działają, wymagają jeszcze wielu poprawek i udoskonaleń. Wynika to z faktu, że proces komunikacji jakim posługują się ludzie, jest w istocie bardzo złożony. Skalę trudności można docenić dopiero podczas próby zbudowania algorytmów komputerowych, które miałyby symulować ten proces. Jednak potrzeba wdrażania elementów naturalnej komunikacji pomiędzy człowiekiem i komputerem staje się na tyle silna, że warto podejmować działania w tym kierunku.

2.1.4. Potrzeba rozwoju metod komunikacji

Produkowane obecnie oprogramowanie, wykorzystuje model komunikacji człowieka z komputerem, oparty na interfejsie graficznym. Zakłada on potrzebę wprowadzania danych przez użytkownika, za pomocą bardzo niewygodnych formularzy, tabel i specjalnie do tego przygotowanych kontrolek. Ten sposób komunikacji jest w wielu przypadkach bardzo uciążliwy i prowadzi do licznych błędów. Z kolei dla producenta tego typu oprogramowana, jest on kosztowny i mało uniwersalny.

Powyższe czynniki powodują podejmowanie działań, które mają na celu stworzenie interfejsu wykorzystującego język naturalny, aby komunikacja z komputerem przypominała tę, która zachodzi pomiędzy ludźmi. To podejście wymaga jeszcze wielu badań prowadzonych na styku filologi, matematyki oraz informatyki [1].

2.1.5. Potrzeba badań nad językiem polskim

Język polski stanowi od lat bardzo trudne zagadnienie dla lingwistyki komputerowej. Bardzo bogata fleksja polszczyzny, niełatwo poddaje się formalnym opracowaniom. W

ostatnich latach, opis taki staje się coraz bardziej potrzebny.

Należy tutaj zwrócić szczególną uwagę na konieczność prowadzenia badań dla języka polskiego. Jak podkreślił prof. Lubaszewski we wstępie do książki: *"Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu"* ***"w badaniach nad komputerowym przetwarzaniem języka ojczystego nikt nas nie zastąpi"*** [1]. Większość współczesnych badań prowadzonych jest dla języka angielskiego. Niestety, język polski posiada różne osobliwości, które powodują, że nie można dla niego stosować formalizmów, modeli i algorytmów skonstruowanych dla innych języków. Dodatkowo, język naturalny jest bardzo silnie związany z narodowością, a nawet folklorem i kulturą jego użytkowników. Te cechy języka, powodują konieczność badań rodzimych naukowców nad językiem narodowym [1].

Lingwistyka komputerowa w Polsce

W Polsce prowadzone są liczne badania z wieloma sukcesami w tej dziedzinie. Najważniejsze ośrodki znajdują się przy wyższych uczelniach.

- **Kraków:** Grupa Lingwistyki Komputerowej AGH, Katedra Lingwistyki Komputerowej UJ - prof. Wiesław Lubaszewski
- **Łódź:** PELCRA (Polish and English Language Corpora for Research and Applications) - prof. Barbara Lewandowska-Tomaszczyk
- **Poznań:** Zakład Lingwistyki Informatycznej i Sztucznej Inteligencji - prof. Zygmunt Vetulani,
  Zakład Angielskiego Językoznawstwa Komputerowego - prof. Włodzimierz Sobkowiak
- **Warszawa:** Zakład Inżynierii Lingwistycznej - dr Adam Przepiórkowski,
  Zakład Językoznawstwa Komputerowego - prof. Marek Świdziński,
  Zakład Systemów Informacyjnych - dr Piotr Gawrysiak
- **Wrocław:** Pracownia Dygitalizacji Tekstów i Lingwistyki Kwantytatywnej - dr Adam Pawłowski

2.1.6. Główne problemy lingwistyki komputerowej

Największym problemem jest wieloznaczność słów i wyrażeń na wszystkich poziomach systemu językowego. Pisał o tym mgr Marcin Gadamer w swojej pracy dyplomowej [2][9]. Występuje ona w:

- fonetyce, np. homofonia (tożsamość różnych znaków językowych),
- morfologii, np. problem analizy części mowy, homonimia (wyrażanie różnych znaczeń za pomocą identycznych form językowych),

- składni, np. problem analizy części zdania,
- semantyce, np. polisemia (wyraz ma więcej niż jedno znaczenie w zależności od kontekstu),
- pragmatyce, np. metafory (wyrazy stojące w określonym porządku obok siebie mają inne znaczenie).

Konstrukcje językowe, poza powyższymi cechami, mogą zawierać również niedopowiedzenia - zdania mogą być urwane i niedokończone. Mogą być one także niepoprawne gramatycznie. Pomimo tego, człowiek w większości przypadków potrafi bez problemu poradzić sobie ze zrozumieniem takich wypowiedzi.

2.2. Matematyczny opis języka - lingwistyka matematyczna

Wielu badaczy zajmujących się lingwistyką komputerową uważa, że nie da się skonstruować maszyny posługującej się, czy też działającej w oparciu o język naturalny, bez opracowania jego matematycznego modelu. Wyróżnić można tutaj dwa kierunki badań, które znacząco różnią się od siebie, a które posiadają zarówno zalety jak i wady.

2.2.1. Modelowanie języka

Podejścia obecne w dotychczasowych pracach nad językiem naturalnym:

- **Modele statystyczne** - bazują na statystyce, nie można za ich pomocą badać poprawności gramatycznej każdego zdania, ale pozwalają na statystyczne przetwarzanie tekstu i ekstrakcję informacji.
- **Modele gramatyczne** - stosują bardzo szczegółowy, matematyczny opis gramatyki języka, jego składni i konstrukcji słów. Wyróżnić należy tutaj prace Noama Chomsky'ego, któremu należy przypisać powstanie teorii gramatyk formalnych oraz rozwój teorii automatów.

2.2.2. Podział języków

Wszystkie języki możemy podzielić na dwie grupy:

- **języki naturalne** - np. język polski, niemiecki, rosyjski - nie ma znanych modeli gramatycznych dla tych języków, czyli inaczej mówiąc, nie da się za pomocą dostępnych modeli, zbudować każdego zdania w języku naturalnym,
- **języki formalne** - czyli języki sztuczne, języki programowania np: C++,Java, ADA itd. - posiadają zdefiniowane modele gramatyczne, powszechnie wykorzystywane przez parsery tych języków.

2.2.3. Opis języka

Opis języka, jego semiotyka, według podziału wprowadzonego przez Charlesa W. Morrisa, składa się z trzech części:

- **syntaktyki** - opis składni, funkcje syntaktyczne, relacje, które zachodzą między wyrażeniami wewnątrz języka,
- **semantyka** - znaczenie słów, bada ponadto relacje między znaczeniem podstawowym wyrazu, a jego znaczeniem w konkretnym kontekście,
- **pragmatyka** - użyteczność, stosowalność języka. Inaczej mówiąc, pragmatyka bada sposoby posługiwania się mową oraz bada mechanizmy jej rozumienia.

Lingwistyka matematyczna zajmuje się składnią języka. Bada zasady budowy poprawnych wypowiedzi, natomiast nie rozstrzyga, czy dana wypowiedź jest prawdziwa, czy też fałszywa.

2.3. System definicyjny wykorzystujący zasoby internetowe

Zasoby internetowe są obecnie bardzo obszerne tak, że zawierają informacje praktycznie na każdy temat. To spowodowało, że zaczęły rodzić się pomysły, aby wykorzystać je do automatycznego budowania wiedzy na pewne tematy.

Rozwój tego typu badań, ma przede wszystkim na celu ułatwienie człowiekowi dostępu do zgromadzonych w sieci informacji. Jest to także ważne zagadnienie z punktu widzenia budowy systemów uczących się.

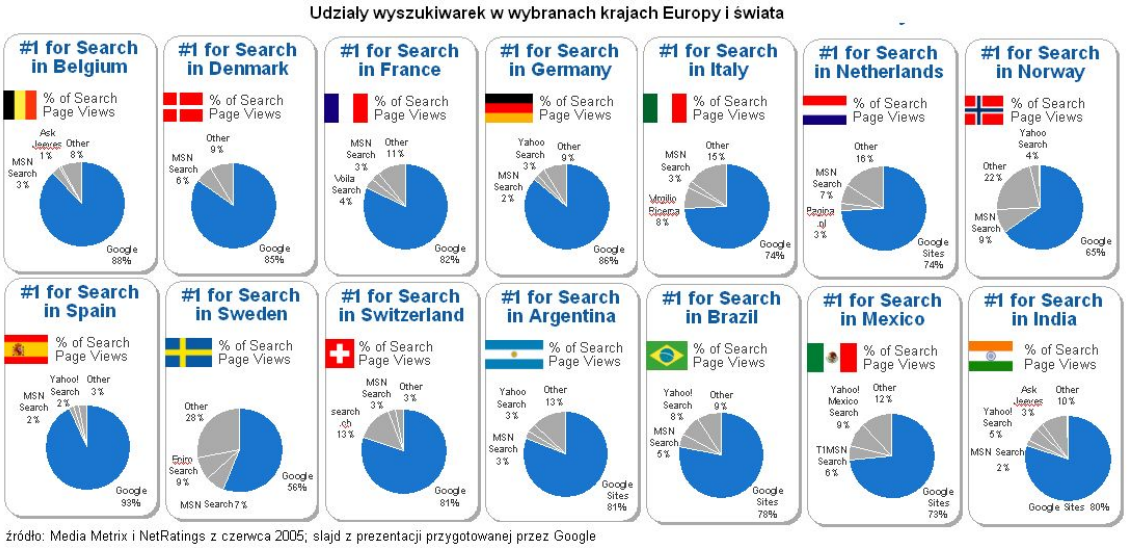
2.3.1. Współczesne wyszukiwarki

Przeszukiwanie Internetu w celu znalezienia interesujących nas danych, polega obecnie na wyszukiwaniu stron internetowych, które prawdopodobnie w większym lub mniejszym stopniu poruszają zadaną tematykę.

Rynek wyszukiwarek zdominowała w ostatnich latach firma Google, której produkt obecnie jest najpopularniejszy nie tylko w Polsce, ale także w większości krajów na świecie, czego dowodzą badania przeprowadzone przez firmy Media Metrix i NetRatings. Oczywiście wyszukiwarki te nie są idealne i posiadają swoje wady, jednak obecnie stanowią najlepsze źródło pozyskiwania informacji z Internetu, w relatywnie krótkim czasie.

Wady współczesnych wyszukiwarek:

- zasoby Internetu są tak rozległe, że czasami nie sposób przejrzeć wszystkie strony, traktujące o interesujących użytkownika zagadnieniach,



Rysunek 2.1. Popularność wyszukiwarek internetowych w różnych krajach na świecie.

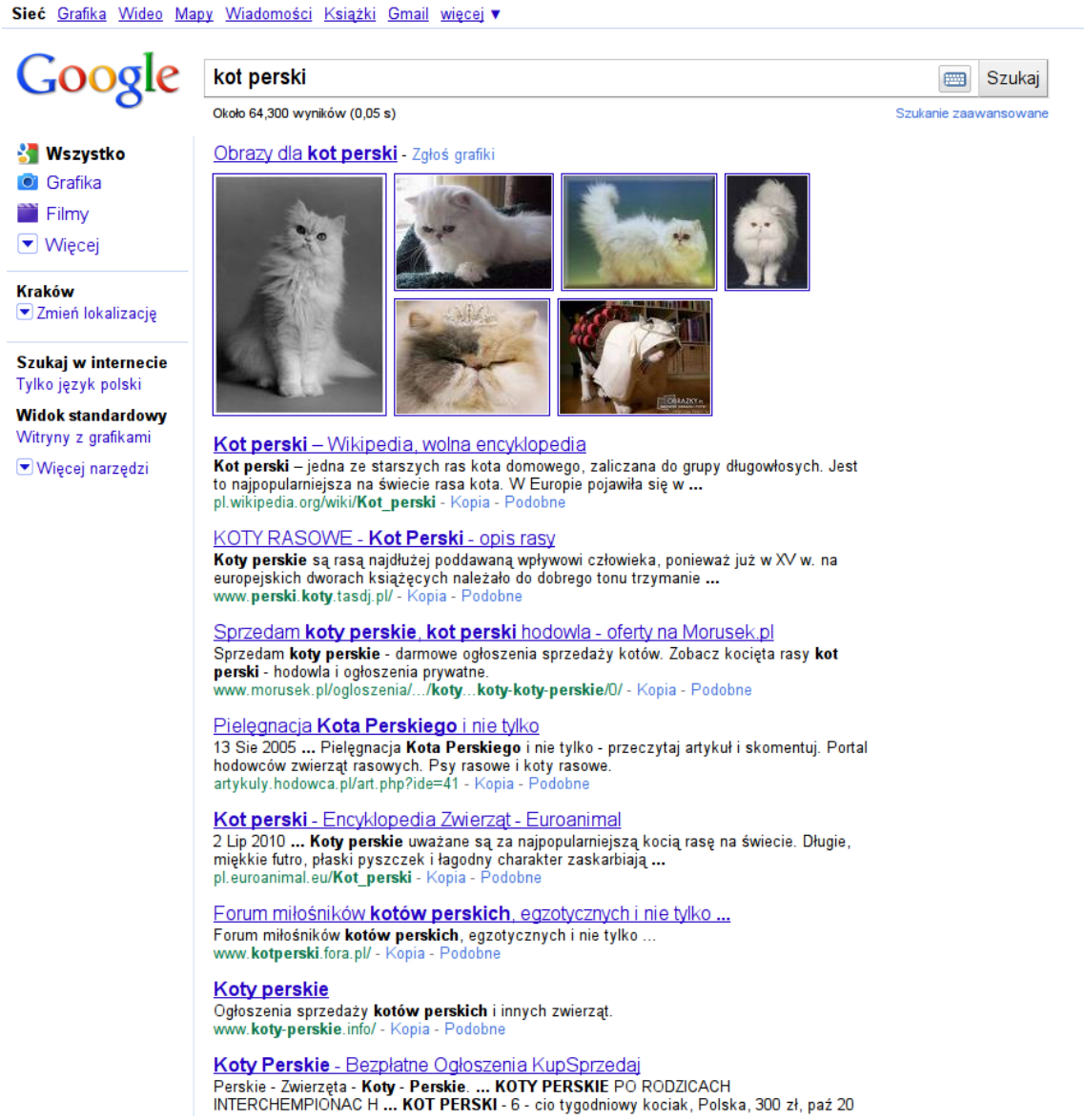
- często musimy wielokrotnie przetwarzać te same informacje, które znaleźliśmy na innej stronie, aby rozbudować naszą wiedzę w dodatkowe szczegóły na wybrany temat,
- nie wiemy ile potrzebnych informacji znajdziemy na danej stronie, żeby się o tym przekonać musimy przejrzeć jej zawartość,
- nie możemy być pewni, że znalezione dane są wiarygodne, bo każdy może publikować artykuły w Internecie i mogą być one błędne, chyba, że dane źródło jest nam znane i jest dla nas wiarygodne,

Zalety inteligentnych wyszukiwarek lingwistycznych:

Inteligentne wyszukiwarki w znacznym stopniu poprawiłyby komfort wyszukiwania informacji, niwelując istotne wady klasycznych wyszukiwarek. Główne zalety wyszukiwarek analizujących tekst to przede wszystkim:

- opis będący syntezą znalezionych w Internecie informacji - baza wiedzy na pewien temat,
- nie trzeba przeglądać wielu stron internetowych, szczegółowa wiedza publikowana w całości na jednej stronie,
- oszczędność czasu zapewniona przez szybszy dostęp do informacji,
- informacje bardziej wiarygodne, bo zostały statystycznie zweryfikowane,
- informacja wyszczególniona tylko jeden raz, nie trzeba przetwarzać wiele razy tych samych danych - robi to za nas specjalistyczny algorytm.

Nie ma wątpliwości, że wyszukiwarki interpretujące język naturalny, byłyby milowym krokiem w dziedzinie pozyskania informacji i budowania wiedzy. Dlatego prace naukowe prowadzone nad tym zagadnieniem są bardzo potrzebne.



Rysunek 2.2. Rezultat zwrócony przez najpopularniejszą w Polsce wyszukiwarkę internetową google.com dla zapytania: kot perski

2.3.2. Wykorzystanie budowanego systemu

System definicyjny, posiadający mechanizmy odpowiadające za naukę i pozyskiwanie wiedzy, posiadałby bardzo dużo zastosowań do budowy systemów lingwistycznych i robotów komputerowych. Najważniejszymi z nich z pewnością byłyby:

- budowa inteligentnych wyszukiwarek lingwistycznych interpretujących język naturalny,
- automatyczne generatory baz wiedzy dla inteligentnych robotów komputerowych,
- wspomaganie tworzenia i rozwijania słowników, tezaurusów i innych dużych zbiorów językowych,
- wspomaganie badań lingwistycznych (analiza dużych zbiorów tekstów, weryfikacja hipotez lingwistycznych),
- automatyczne generatory treści,
- budowa systemów służących do korekcji tekstów, gdzie wymagane jest rozpoznanie kontekstu wypowiedzi,
- budowa systemów służących do tłumaczenia treści,
- inteligentne systemy dialogowe.

Budowa systemów interpretujacych język naturalny, potrafiących wybrać z niego ważne informacje, jest zagadnieniem bardzo skomplikowanym. Składa się na to przede wszystkim, sam charakter jezyka oraz zasada działania współczesnych komputerów, wykorzystywanych do realizacji tych celów, która jest inna od zasady działania mózgu człowieka, na którym wzoruje się wiele algorytmów i rozwiązań.

3. Narzędzia i metody służące do ekstrakcji informacji

3.1. Narzędzia

Systemy lingwistyczne, przeprowadzające pewne automatyczne operacje na języku naturalnym i posiadające pewne cele do opracowania, są generalnie bardzo trudne do skonstruowania. Najprościej można sobie uzmysłowić rząd wielkości problemu, gdy spróbujemy przeanalizować jakiś dokument napisany w języku obcym, całkowicie dla nas niezrozumiałym. Nie posiadając żadnych dodatkowych informacji o tekście, rozróżniając jedynie symbole, możemy dokonać prostej analizy statystycznej i stwierdzić, które wyrazy powtarzają się najczęściej.

Niestety informacje statystyczne nie są wystarczające do stwierdzenia, o czym jest dany tekst. Wynika to z tego, że pewne wyrazy, mimo iż nie niosą ważnych informacji, to w danym języku mogą być szczególnie często używane, z racji na stosowane konstrukcje gramatyczne lub charakter.

Prowadząc takie proste rozważania, można dojść do wniosku, że konieczne jest dostarczenie dodatkowych informacji, które algorytmy będą umiały należycie wykorzystać. Można posłużyć się tutaj analogią do zachowania się człowieka w takiej sytuacji. Jeżeli nie rozumiemy pewnych słów, to zaczynamy szukać o nich informacji w słownikach językowych. Podobne rozwiązanie wydaje się być konieczne w przypadku komputerowych metod przetwarzania tekstu. Okazuje się, że słowniki komputerowe, stanowią nieodzowną pomoc przy budowie algorytmów lingwistycznych, ponieważ są jedynym źródłem dostarczającym informacje o przetwarzanych słowach.

3.1.1. Znaczenie wyrazu

Wyraz, to najmniejszy znaczący element języka służący porozumiewaniu się, zdolny spełniać pewne funkcje gramatyczne. Może samodzielnie lub w połączeniu z innym wyrazem, stanowić człon wypowiedzenia lub wypowiedzenie.

Często to, co nazywamy wyrazem, zależy od specyfiki języka oraz tradycji językowych. Dla języka pisanego wyrazem, bardzo często nazywamy ciąg liter pomiędzy dwoma spacjami i traktujemy go identycznie jak słowo.

Wyraz natomiast, może być pojedynczym słowem lub zestawem słów, które posiadają pewne znaczenie, nie dające się podzielić na mniejsze wyrazy. Są to wyrazy złożone. Podana definicja zakłada, że niektóre słowa mogą nie być wyrazami jak: *na, w, oraz.*

Jednym z najtrudniejszych problemów stawianych przed lingwistyką komputerową, jest opis znaczenia wyrazu oraz budowa metod służących do rozpoznawania znaczeń, które będą mogły być w pełni wykorzystywane przez maszyny.

Każdy wyraz w tekście reprezentowany jest przez zbiór znaków, a komputer tylko takie dane posiada we wstępnej analizie tekstu. Dopiero wykorzystanie inteligencji oraz dodatkowej wiedzy, pozwala rekonstruować znaczenie wyrazu.

Zadanie to relatywnie łatwe dla człowieka, staje się ogromnym problemem dla komputera. Ludzie gromadzą obszerną wiedzę z przeróżnych dziedzin, będącą efektem wielu życiowych doświadczeń i potrafią ją wykorzystać w dowolnej chwili w zależności od potrzeby.

Umysł człowieka jest bardzo skomplikowany. Mimo, iż pewne zachowania wydają się nam oczywiste i łatwe, to są bardzo trudne do przeniesienia na algorytmy komputerowe. Mózg człowieka działa na zasadzie analogii. Potrafi błyskawicznie kojarzyć pewne fakty i wykorzystywać je, podczas gdy komputer jest jedynie maszyną do przetwarzania danych. Rozwiązania, które dla człowieka od razu wydają się błędne, komputer musi poddać identycznym testom jak te prawidłowe, aby uzyskać rozwiązanie. Mózg potrafi wykorzystać pamięć o przeszłych doświadczeniach, aby przyspieszyć swoją pracę i szybciej kierować się w stronę rozwiązania.

Analizę znaczeń bardzo utrudnia zjawisko homonimii, czyli możliwość zapisu różnych znaczeń, za pomocą tej samej słownej reprezentacji. Ten sam zbiór wyrazów, w różnym kontekście, może oznaczać zupełnie co innego. Jedynie analiza całego kontekstu wypowiedzi, pozwala na odgadnięcie poprawnego znaczenia. Niestety jest to zadanie bardzo trudne, bo może się ono wiązać z koniecznością powrotu do informacji z wcześniejszego fragmentu tekstu, czy też przeanalizowania dalszej jego części. Co gorsza, takie informacje wcale nie muszą znajdować się w bezpośrednim sąsiedztwie analizowanego ciągu.

Niestety, bardzo często jedynym rozwiązaniem takiego problemu, jest zawężenie zbioru potencjalnych znaczeń do pewnych szczególnych, które mogą pojawić się w analizowanym tekście.

Opis znaczenia wyrazu

Opis znaczenia wyrazu ma służyć do odróżnienia znaczeń wyrazów występujących w tekście, a to warunkuje poprawność działania algorytmów interpretacji niesionej przez tekst informacji.

Skonstruowanie wyczerpującego opisu dowolnego znaczenia jest zadaniem trudnym. Na samym początku warto uzmysłowić sobie pewne podstawowe właściwości tekstu. Temat został bardzo ciekawie opisany w pozycji [1], z której zaczerpnięty jest poniższy opis relacji, zachodzących miedzy słowami.

Możemy intuicyjnie stwierdzić, że pomiędzy wyrazami: *pies w znaczeniu "zwierze domowe"* oraz *"szczeka"*, zachodzi pewna relacja. Dodatkowo możemy stwierdzić, że taka relacja nie zachodzi pomiędzy wyrazami: *pies w znaczeniu "samiec lisa i borsuka"* oraz *"szczeka"* [1].

Relacje takie jednoznacznie definiują znaczenia, niestety automatyczne generowanie takich zależności jest bardzo trudne. Analizując pewien tekst, w wiekszości przypadków, musimy sami zdecydować, o które znaczenie chodzi.

Relacje między wyrazami

W pozycji [1] wyróżniane są dwa podstawowe typy relacji:

1. **Relacje ogólne** - ten typ znajduje zastosowanie przy opisie każdego znaczenia. Wyróżniamy tutaj następujące relacje szczegółowe:
   - **określające podobieństwo** - pewne relacje są wspólne, a pewne różne dla danych znaczeń,
   - **określające identyczność** - synonimy - pozostałe relacje opisujące oba znaczenia są identyczne,
   - **określające przynależność do kategorii semantycznej** np: *Animal*, *Human*,
2. **Relacje swoiste** - ten typ relacji służy do opisywania specjalnych cech danego znaczenia. Relacje takie są różne, dla różnych grup znaczeniowych.

3.1.2. Słownik

Słownikiem nazywamy zbiór wyrazów opracowanych według pewnej zasady, zwykle objaśnianych pod względem znaczeniowym. Standardowo hasła, w klasycznych słowni-

kach książkowych, poukładane są w kolejności alfabetycznej, rzadziej tematycznej. Typ słownika określa sposób budowy opisu tematycznego oraz rodzaj informacji, jakie on zawiera.

Najpopularniejsze rodzaje słowników :

- **ortograficzny** - zawiera poprawną pisownię wyrazów oraz ich form i służy do jej sprawdzania,
- **języka np. polskiego** - zawiera zbiór słów wraz z opisem ich znaczenia, zbiór form fleksyjnych dla języka fleksyjnego oraz przykłady stosowania danych form,
- **dwujęzyczny** - zawiera tłumaczenia wyrazów dla dwóch języków,
- **wyrazów bliskoznacznych** - zawiera wyrazy o podobnym znaczeniu,
- **terminologiczny** - zawiera słowa wraz z ich objaśnieniami, charakterystyczne dla pewnej dziedziny np. medycyny.
- **encyklopedyczny** - zbudowany jest ze zbioru haseł, wraz z artykułami objaśniającymi ich znaczenie uporządkowane.

Bardzo popularne stały się słowniki komputerowe, zwłaszcza internetowe, które posiadają wiele zalet nad klasycznymi słownikami papierowymi:

- szybszy dostęp do informacji, dzięki zastosowaniu wyszukiwarek komputerowych,
- darmowe słowniki internetowe pozwalają na natychmiastowy dostęp do informacji wszystkim tym, którzy posiadają dostęp do Internetu,
- słowniki komputerowe są łatwiejsze do modyfikacji i rozbudowywania, a zmiany są natychmiastowo udostępniane użytkownikowi,
- edytory tekstu zostają wzbogacane o słowniki ortograficzne oraz tezaurusy, podpowiadające użytkownikowi treść.

Automatyczna analiza języka naturalnego jest problemem skomplikowanym, a jedynymi narzędziami, które mogą pomóc przy budowie algorytmów, są właśnie słowniki komputerowe, skonstruowane w postaci bibliotek programistycznych.

Słownik języka polskiego

Słownik języka polskiego należy do grupy słowników jednojęzycznych ogólnych. Podstawowe zadanie jakie on spełnia, to miedzy innymi, dostarczenie wiedzy o znaczeniu ujętych w nim słów, poprzez opisowe wyjaśnienie. Często zawiera elementy fleksji, ortografii, etymologii oraz przykłady użycia. Hasła w takim słowniku są uszeregowane alfabetycznie, zgodnie z alfabetem.

Ze względu na dynamiczny rozwoju techniki, w szczególności łatwemu dostępowi komputerów oraz Internetu, coraz bardziej popularne są słowniki języka polskiego w postaci elektronicznej, dostępne lokalnie lub zdalnie.

Słownik fleksyjny języka polskiego

Słownik ten należy do słowników jednojęzykowych. Zawiera on zbiór form odmiany wyrazu, dla pewnej formy podstawowej. Język polski jest językiem fleksyjnym, a to oznacza, że każdy wyraz reprezentowany jest w tekście przez jedną z form fleksyjnych [1]. Często ten typ słownika występuje w połączeniu z innymi słownikami.

Opis hasła składa się z następujących części:

- **forma hasłowa** - jedna z form fleksyjnych wyrazu, zwyczajowo reprezentująca wyraz w słownikach (np. mianownik liczby pojedynczej dla rzeczowników).
- **opis gramatyczny** - zwiera części mowy dla wszystkich wyrazów i zbiór końcówek,
- **lista form** - para $opis\_formy$ - $forma$, gdzie $opis\_formy$ dla np. rzeczownika to przypadek i liczba, dla przymiotnika - przypadek, liczba i rodzaj, a dla czasownika: osoba, czas, tryb i rodzaj.

Tworzenie słowników fleksyjnych dla języka polskiego stało się, w ostatnim dziesięcioleciu, dość popularnym tematem projektów i badań. Efektem tych działań jest powstanie słowników fleksyjnych:

- **Grupa Lingwistyki Komputerowej - Kraków** - Biblioteka CLP. Głównym jej elementem jest baza danych składająca się ze 120 tysięcy rekordów i obejmująca praktycznie cały zasób wyrazów pospolitych.
- **Firma POLENG** - słownik fleksyjny “Dylemat” - składa się z 223 tys. form bazowych wraz z formami fleksyjnymi. Jest to również narzędzie do lematyzacji, czyli określania poprawnej gramatycznie formy podstawowej słowa oraz biblioteka dla języka C++.

Słownik fleksyjny języka polskiego - biblioteka CLP

W niniejszej pracy wykorzystany został słownik CLP w wersji 2.0 [7]. Jest to biblioteka programistyczna napisana w języku C. Jest ona dostępna jedynie w środowisku akademickim i działa wyłącznie pod systemem Linux. Aby można jej było używać z innymi językami programowania, należy napisać specjalny interfejs. Podlega ona ochronie praw autorskich, co wynika z licencji.

Podstawowymi funkcjami realizowanymi przez bibliotekę CLP zaczerpniętymi bezpośrednio z dokumentacji są:

- rozpoznawanie wyrazu na podstawie jego dowolnej formy fleksyjnej,
- dostarczanie informacji o słowie,
- wygenerowanie form fleksyjnych dla danego słowa,
- dostarczanie informacji o formie słowa.

Głównym elementem biblioteki jest słownik fleksyjny języka polskiego. Podczas inicjalizacji, cała baza słownika jest ładowana do pamięci komputera, dzięki czemu działa ona bardzo szybko. Wszystkie wyrazy zawarte w bibliotece, przynależą do jednej z poniższych klas:

- rzeczownik (53867);
- czasownik (20067);
- przymiotnik lub imiesłów przymiotnikowy (38066);
- liczebnik (168);
- zaimek (182);
- przysłówek (5068);
- wyraz nieodmienny (504).

Biblioteka CLP - API

Biblioteki CLP w wersji 2.0 udostępniają użytkownikowi API składające się z następujących funkcji języka C [7]:

- *void clp_init();*
  - inicjalizacja biblioteki,
- *char *clp_ver();*
  - zwraca numer wersji,
- *int clp_pos(int id);*
  - zwraca numer oznaczający część mowy danego wyrazu,
- *void clp_label(int id, unsigned char *out);*
  - zwraca etykietę fleksyjną wyrazu,
- *void clp_bform(int id, unsigned char *out);*
  - zwraca formę podstawową wyrazu,
- *void clp_forms(int id, unsigned char *out);*
  - zwraca zbiór wszystkich form wyrazu,
- *void clp_formv(int id, unsigned char *out);*
  - zwraca wektor wszystkich form wyrazu,

- *void clp_vec(int id, const char *inp, int *out, int *num);*
  - funkcja zwraca tablicę z numerami pozycji w wektorze odmiany wyrazu,
- *void clp_rec(const char *inp, int *out, int *num);*
  - funkcja zwraca tablicę z numerami id dopasowanych wyrazów,
- *int clp_tag1(const char *inp, char *out);*
  - tag wyrazu jednosegmentowego.

Należy zauważyć, że cały czas prowadzone są prace nad słownikiem i wraz z kolejnymi wersjami, dodane są nowe funkcjonalności.

3.2. Metody ekstrakcji informacji

Internet, czyli globalna sieć informacyjna, to bez wątpienia jedno z największych osiągnięć XX wieku, które dało człowiekowi przede wszystkim nieograniczony dostęp do publikowanych w sieci informacji. Każdy użytkownik Internetu może, bez większych ograniczeń, publikować własne strony oraz posiada nieograniczone możliwości przeglądania zamieszczonych źródeł.

Podejście zapewniające ogólną dostępność Internetu powoduje, że jego zasoby rosną w bardzo szybkim tempie. Codziennie przybywają nowe strony internetowe, zawierające informacje zapisane w postaci bloków tekstu w języku naturalnym. Ilość dostępnych materiałów jest przytłaczająca. Ciężko jest przeglądnąć je wszystkie w poszukiwaniu szczegółowej informacji. To spowodowało, że zaczęto zastanawiać się nad możliwością zbudowania algorytmów oraz systemów, potrafiących samodzielnie wyszukiwać i interpretować informacje.

Wraz z rozwojem badań nad językiem naturalnym, powstały dziedziny podejmujące temat wyszukiwaniem informacji w zasobach internetowych. Wykształciły sie tutaj dwa popularne podejścia:

- **Information Retrieval**
- **Information Extraction**

Zostaną one szczegółowo omówione w dalszej części rozdziału.

3.2.1. Analiza tekstu

Aby komputer w pełni potrafił zrozumieć tekst, należałoby przeprowadzić trzy niżej opisane analizy, o rosnącym stopniu trudności. Są one niezbędne, aby w pełni zrozumieć wypowiedź [1][5].

Analiza leksykalna

Polega ona na wyszukaniu, w dostępnych słownikach, słów wchodzących w skład zdania w celu ustalenia ich znaczenia oraz formy gramatycznej. Dlatego potrzebujemy słowników komputerowych, dostarczających informacji o słowach i ich formach leksykalnych. Nie jest to zadanie łatwe, ponieważ bogata fleksja języka polskiego dodatkowo je utrudnia. Występują liczne wieloznaczności językowe, niemożliwe do wykrycia na tym etapie analizy.

Pocieszające może być to, że w wielu wypadkach, nie potrzebujemy rozpoznać wszystkich znaczeń. Dla dobrze zdefiniowanego scenariusza wystarczy jedynie rozpoznanie tzw. słów kluczowych, aby rozpoznać temat i kontekst wypowiedzi.

Od strony technicznej, zadaniem analizatora leksykalnego jest przetworzenie tekstu na ciąg tokenów, które będą używane w dalszej analizie.

Analiza syntaktyczna

Gramatyczny rozbiór zdania. Jej głównym celem jest wydobycie powiązań pomiędzy elementami zdania i zrozumienie jaką funkcję w tym zdaniu pełnią. Analiza składniowa dla języków naturalnych jest zagadnieniem trudnym, ze względu na ich charakter. Reguły gramatyczne często są skomplikowane i zawierają liczne wyjątki, a znaczenie poszczególnych wyrazów w wielu wypadkach zależy od kontekstu. Wybór gramatyki, która będzie używana dla danego języka, zależy od specyfiki konkretnego języka i rozpatrywanego zagadnienia, liczą się też względy obliczeniowe. Niektóre systemy analizy używają gramatyk funkcjonalnych, ale generalnie parsowanie z ich wykorzystaniem jest problemem NP-zupełnym. Etap ten jest łatwiejszy dla języków o zdeterminowanym szyku zdania, jak dla języka angielskiego. Niestety język polski nie posiada tej cechy.

Innym ważnym problemem może być to, że ludzie często budują zdania niepoprawne gramatycznie, co dla analizującego tekst jest kolejnym dużym wyzwaniem. Wymaga od niego dopasowania prawidłowego wzorca gramatycznego.

Większość istniejących parserów dla języków naturalnych jest, przynajmniej częściowo, oparta na statystyce. W pierwszym etapie, analizie statystycznej poddawany jest specjalnie zbudowany korpus języka, co pozwala systemowi zgromadzić informacje o częstości występowania poszczególnych wyrazów i fraz, w różnych kontekstach. Wykorzystuje się przy tym metody takie jak: PCFG, badanie entropii oraz sieci neuronowe. Większość wiodących systemów używa statystyk leksykalnych (porównują podobieństwo części mowy i poszczególnych słów). Systemy te wymagają jednak pewnych korekt.

Parsery dla języków naturalnych, niestety nie mają do dyspozycji prostych gramatyk, takich jak te dla języków programowania. Używane gramatyki formalne są bardzo skomplikowane obliczeniowo, zwłaszcza w parsingu. Często są stosowane jakieś bezkontekstowe przybliżenia, aby wykonać pierwsze dopasowanie.

Algorytmy, które używają gramatyk bezkontekstowych, często odrzucają mało prawdopodobne dopasowania i rezygnują z dokładności na rzecz szybkości.

Analiza semantyczna

Ten etap ma na celu odgadnięcie znaczenia wypowiedzi. Jest to zadanie, które daje niejednoznaczne wyniki, bo analizowane zdania mogą być zależne od całego kontekstu lub co gorsza, mogą być one niedokładne, urwane lub mogą zawierać metafory. Wtedy posługując się algorytmem, jaki stosuje człowiek w takiej sytuacji, należy zgadnąć znaczenie na podstawie przesłanek dostarczonych przez cały kontekst. Jest to zadanie wyjątkowo skomplikowane, bo wymaga zastosowania uniwersalnej wiedzy, co dla ludzi nie jest zbyt trudne, to dla komputerów stanowi poważny problem.

Pomimo, iż człowiek opanował analizę tekstu do perfekcji, nie udało się stworzyć algorytmów, które wzbogacą o takie umiejętności komputery. Zaczęto zatem opracowywać metody, pozwalające na osiągnięcie zbliżonych efektów bez przeprowadzania tak skomplikowanych rozważań.

3.2.2. Pozyskiwanie informacji - Information Retrieval (IR)

Jest to metodologia pozyskiwania informacji, która obecnie jest bardzo intensywnie wykorzystywana przez współczesne wyszukiwarki, zwłaszcza internetowe. Efektem pracy metod z rodziny IR jest zbiór adresów, które zostały zwrócone przez wyszukiwarkę jako te, które najprawdopodobniej zawierają dane o interesującej nas kwerendzie. Użytkownik musi analizować dokumenty, w celu wyekstrahowania istotnych dla niego informacji z tekstu.

Systemy IR wykorzystują dwa podejścia:

- **dokładne (exact match)** - w większości wypadków związane z zastosowaniem jakiegoś języka wyszukiwania (czy też raczej języka specyfikacji zapytań – query language),
- **rozmyte (fuzzy)** – wykorzystuje metody statystyczne do oceny odpowiedniości dokumentu do zapytania - główny problem: konstruowanie zapytania.

Największe trudności metod opartych o metodologię IR:

- zapewnienie wysokiej relewancji odpowiedzi,
- zapewnienie wysokiej kompletności odpowiedzi,
- przedstawienie wyniku w zrozumiały i efektywny sposób.

Powyższe problemy są głównym powodem rozwoju metod opartych o ekstrakcję informacji.

3.2.3. Ekstrakcja informacji - Information Extraction (IE)

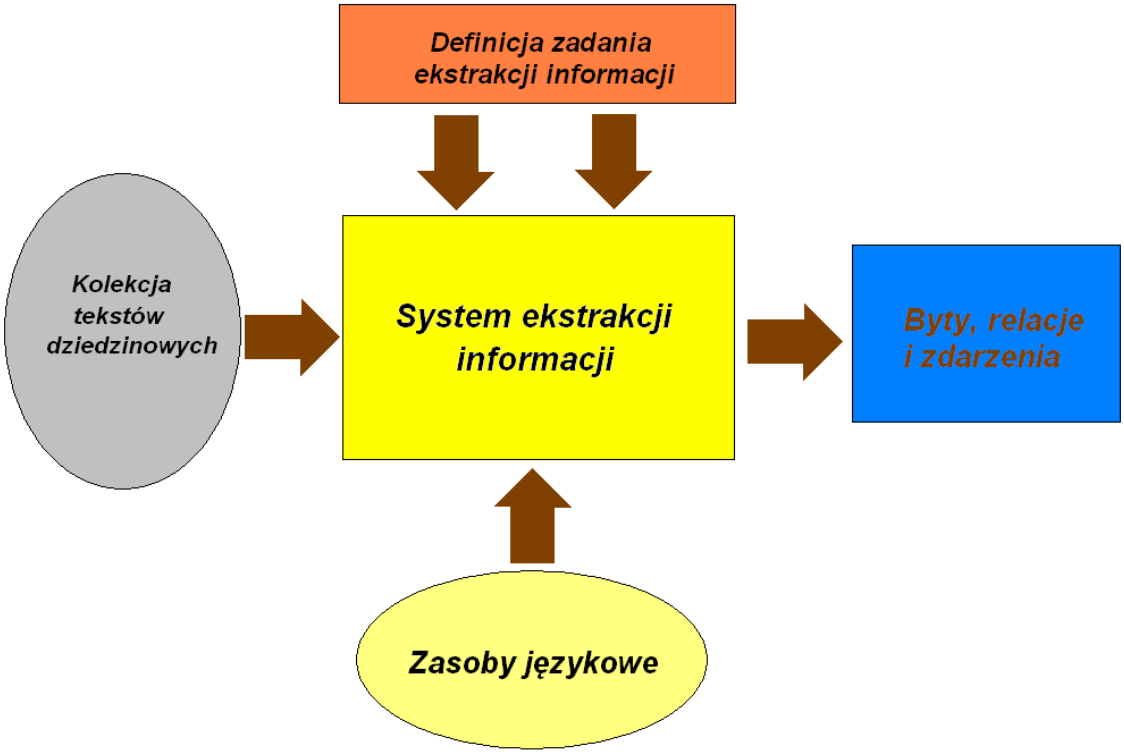
Ekstrakcja informacji jest to automatyczna identyfikacja wybranych typów bytów, relacji i zdarzeń w tekście. Identyfikacja bytów jest tutaj rozumiana jako rozpoznanie fragmentów tekstu i przyporządkowanie im znanych typów bytów, czyli polega na znalezieniu w tekście np. imion, nazwisk, nazw miast, ulic, lokalizacji, miejsc. Identyfikacja relacji jest poszukiwaniem znanych typów połączeń występujących pomiędzy bytami, natomiast identyfikacja zdarzeń jest wyszukiwaniem zmian, które zaszły w relacjach pomiędzy bytami. Podrozdział powstał w oparciu o źródła [8][10][12][27].

Dlaczego zadanie Information Extraction jest trudne?

- Ta sama informacja może być zapisana w różny sposób w języku naturalnym.
- Informacja może być zapisana w skomplikowany sposób, może być ukryta za metaforami.
- Komputer musi przeanalizować cały tekst, dokonując wstępnej analizy syntaktycznej i dokonać dopasowania do wzorca.
- Ludzie bardzo często stosują metafory, niedopowiedzenia lub urywają zdania.
- Mnogość form fleksyjnych dla języka polskiego oraz bardzo skomplikowana gramatyka powodują, że zadanie to staje się dodatkowo trudniejsze i nie można dla niego stosować tych samych reguł, co dla innych języków.

Zasada działania systemów IE

Systemy Information Extraction przetwarzają i analizują bloki tekstu, w celu wydobycia z nich pewnych struktur i relacji, które są nośnikiem istotnych, z punktu widzenia zadania, informacji. Większość tego typu systemów zapamiętuje połączania pomiędzy tekstem bazowym, aby dodatkowo umożliwić użytkownikowi analizę kontekstu. Informacje, jakie są w ten sposób pozyskiwane, mogą być bardzo zróżnicowane.



Rysunek 3.1. Zasada działania systemów IE.

Wyróżnić można trzy fazy działania algorytmu IE:

- lokalna analiza tekstu, która polega na wytworzeniu wzorców i zapełnieniu ich lingwistycznymi realizacjami relacji, przy wykorzystaniu technik NLP,
- integracje wyszukanych faktów,
- przekształcenie do wymaganego formatu.

Informacja, jaka będzie pozyskiwana z dokumentu, jest definiowana przez użytkownika, ponieważ musi on wcześniej stworzyć wzorzec, w oparciu o który ekstrakcja będzie dokonywana. Na wzorzec taki składają się pewne fragmenty tekstu, zawierające luki, które następnie są wypełniane dopasowanymi wyrazami, bądź wybranymi, wcześniej przedefiniowanymi wartościami lub odwołaniami. Niezależnie jednak od tego, schemat działania jest taki sam. Do systemu IE wprowadzane są bloki tekstu napisane w języku naturalnym, zaś pozyskiwane są dokumenty strukturalizowane, będące odwzorowaniem ich treści.

Najważniejszymi modułami systemu Information Extraction są:

- Procesor tekstów

- Generator wzorców

Zadaniem procesora tekstów jest analiza leksykalna, w celu pozyskania jak najszerszego obrazu językowego tekstu (struktury lingwistycznej). Stosować można tutaj dwa podejścia:

- DTP(deep text processing) - oznaczające analizowanie wszelkich możliwych interpretacji i gramatycznych relacji w dokumencie. Ze względu na dużą złożoność, związaną z przetwarzaniem języka naturalnego, jest rzadko stosowana.
- STP(shallow text processing) - jest mniej czasochłonną metodą i dlatego jest częściej stosowana - nie analizuje wszystkich możliwych interpretacji i relacji gramatycznych w dokumencie. Jest kompromisem pomiędzy wyszukiwaniem wzorców, a pełną analizą leksykalną. Występujące w tekście językowe nieregularności, które zwykle stwarzają problemy, nie są przetwarzane, a zamiast analizowania całości tekstu, badane są jedynie wcześniej ustalone struktury.

Generator wzorców, następnie wykorzystuje wiedzę dziedzinową oraz efekty pracy procesora tekstu, w celu zbudowania specjalistycznych relacji. W praktyce, rola tych modułów może być sprowadzona do pojedynczego algorytmu, a podział na rozłączne elementy może ulec zatarciu.

Zadaniem procesora tekstu, jest miedzy innymi podział tekstu na zdania, z których każde jest sekwencją wyrazów wraz z wyspecyfikowanymi atrybutami leksykalnymi, interpretacja skrótów oraz analiza podstaw słowotwórczych. Wynikiem działania tego modułu ma być zestaw prawidłowości oraz reguł, istotnych z punktu widzenia określonej dziedziny oraz zadania.

Wzorce dla systemu IE

Wyróżnić można dwa podejścia stosowane w kontekście tworzenia wzorców:

- Metody inżynierii wiedzy - reguły definiowane są przez tzw. eksperta, który korzysta z własnej wiedzy, dokonując operacji na dokumencie,
- Podejście uczące - system sam uczy się reguł ekstrakcji, w oparciu o ręcznie opisywane dokumenty i interakcję z użytkownikiem.

Przeważnie lepsze wyniki otrzymuje się w przypadku systemów tworzonych we współpracy z ekspertem, jednakże w pewnych wypadkach, automatycznie wygenerowane wzorce mogą okazać się doskonalsze w pozyskiwaniu informacji z dokumentów. Ciekawe może okazać się połączenie tych dwóch technik, z punktu widzenia możliwości opisu

analizowanego dokumentu.

3.2.4. Text mining

Systemy z rodziny Text mining wykorzystują podejście bazujące na statystycznych metodach analizowania tekstu. Służą one do wydobywania informacji statystycznych, wykorzystywanych do dalszej obróbki. Najczęściej zakładamy tutaj, że skoro zdania mają określoną strukturę i zawierają pewne informacje, to duża liczba zdań na dany temat pozwoli na wydobycie najistotniejszych powiązań. Działanie takich metod, przeważnie polega na zliczeniu częstości pojawiania się słów lub ich form fleksyjnych, w kontekście danego problemu. Pozyskana w ten sposób informacja ilościowa, pozwala na zdecydowanie, czy dane słowo jest ważne i czy może być ono częścią informacji, istotnej dla wyszukiwanej kwerendy.

Do typowych zadań text mining należy:

- znajdowanie dokumentów najbardziej podobnych do zapytania użytkownika,
- tworzenie rankingów dokumentów,
- grupowanie dokumentów (analiza skupień),
- klasyfikowanie dokumentów (kategoryzacja),
- analiza powiązań między jednostkami tekstu,
- dokonywanie automatycznych streszczeń dokumentów (ang. summarizing),
- identyfikacja słów kluczowych,
- analiza treści dokumentów oraz automatyczne wykonanie streszczenia dokumentu,
- znajdowanie i klasyfikacja dokumentów najbardziej podobnych do zapytania użytkownika,
- wyszukiwanie dokumentów pasujących do zadanych wzorców.

Lista skojarzeniowa

Jedną z najczęściej używanych metod ilościowych są listy skojarzeniowe. Rozumiane są one jako zbiór wyrazów, uporządkowanych częstotliwościowo oraz semantycznie powiązanych z wyrazem wejściowym. np. dla wyrazu FLOTA wyrazami semantycznie związanymi są: OKRĘT, MARYNARZ, ADMIRAŁ, PORT, PŁYNĄĆ, itd.

Działanie takiego algorytmu jest bardzo proste. Wczytuje on cały korpus tekstu, zliczając ilość wystąpień każdego słowa sprowadzonego do formy podstawowej. Następnie budowana jest lista rankingowa, posortowana według tzw. miary skojarzenia, gdzie:

lw - częstość bezwzględna, ilość wystąpień w korpusie tekstu,

cw - częstość względna, ilość wystąpień w zdaniach, wspólnie z wyrazem definiowanym,

sk - miara skojarzenia będąca ilorazem **cw/lw**.

Podstawą konstrukcji mechanizmu wnioskowania statystycznego jest założenie:

- jeżeli np. wyraz SZALUPA występuje w korpusie tekstów 12 razy, a w zdaniach z wyrazem FLOTA 10 razy, to wyraz SZALUPA jest silnie skojarzony z wyrazem FLOTA,
- jeżeli np. wyraz NIEBO występuje w korpusie tekstów 150 razy, a w zdaniach z wyrazem FLOTA 15 razy, to wyraz NIEBO jest słabo skojarzony z wyrazem FLOTA.

Przykład został zaczerpnięty z pozycji [1].

Teoria zależności pojęciowych - ramki wiedzy

Kolejną metodą służącą do wyszukiwania informacji w tekście jest technika bazująca na teorii tzw. ramek wiedzy. Ramka jest szablonem, który narzuca pewien charakter wynikom wyszukiwania, bo zawiera definicje pewnych zdarzeń, posiadających określoną liczbę cech, należących do ustalonego zbioru wartości. Algorytm ekstrakcji na podstawie tekstu tematycznego wypełnia odpowiednie pola w formularzu bazując na regułach, jakie narzuca szablon.

Rozwinięte tematyczne algorytmy, mające pozyskiwać szczegółową wiedzę, powinny mieć zdefiniowane ramki dla wszystkich sytuacji, jakie mogą być opisywane przez analizowany dokument. Niestety w ogromnej większości przypadków nie możemy przewidzieć różnych możliwości. Wyjątkami mogą być systemy posiadające bardzo wyraźnie zdefiniowany temat, gdzie struktura analizowanego tekstu jest wcześniej ustalona lub łatwa do odgadnięcia: np. “System oceny CV”, Czy “System budujący bazę przepisów kulinarnych”. Wtedy można pokusić się o budowę skryptów, potrafiących ekstrahować dokładne informacje. Większość systemów jednak tego nie używa, pracując w oparciu o tzw. szkice skryptów, które zawierają tylko zarysy konceptualne.

Kolejną wadą tego podejścia, jest przymus ręcznego konstruowania szablonów, co powoduje konieczność zawężenia dziedziny tematycznej dla takiej analizy. Ta wada spowodowała, że zaczęto konstruować algorytmy, które same potrafią budować tego typu wzorce, wykorzystując wiedzę pozyskaną w wyniku ekstrakcji informacji ze specjalnie zdefiniowanych dokumentów.

Skuteczność tej metody, zależy od trudności samego zadania. Jeżeli mamy do czynienia z ekstrakcją tekstów o ściśle określonej strukturze, mogą być osiągnięte

bardzo wysokie dokładności analizy, np. w przypadku ogłoszeń o kupnie/sprzedaży nieruchomości, jednak przy tekstach o nieustalonej strukturze, skuteczność jest wyraźnie niższa.

Istnieje także podejście, zakładające tworzenie ramek wiedzy, związanych raczej z leksykalnymi i syntaktycznymi właściwościami tekstu, a nie z jakąś konkretną dziedziną tematyczną. Ramka tego typu skupia się na przedstawieniu cech charakterystycznych danego zdarzenia, zamiast dokładnie odwzorowywać zagadnienia danego tematu. Zaletą powyższego podejścia jest to, że nie potrzeba dostosowywać takiej ramki do specyficznej dziedziny wiedzy, jednak ma to negatywny wpływ na jakość i precyzję takiej ekstrakcji informacji.

Metody statystyczne bazujące na teorii spójności leksykalnej

Metody statystyczne zakładają odzwierciedlenie spójności semantycznej tekstu za pomocą spójności leksykalnej. Ludzie postrzegają tekst jako spójną całość, charakteryzującą się pewną ciągłością semantyczną. Można zatem przyjąć, że jeżeli semantycznie powiązane słowa tworzą kolekcje leksykalne, które powtarzają się ze znaczącą częstotliwością w określonym fragmencie tekstu i jeżeli powtarzają się one ze szczególną częstotliwością, to pojawienie się tych słów obok siebie jest nieprzypadkowe.

Metody bazujące na powyższym założeniu są bardzo popularne :

- techniki segmentacji tematów,
- techniki tworzenia klastrów zdań,
- konstruowanie łańcuchów leksykalnych,
- w podsumowywaniu tekstów łańcuchy leksykalne mogą być używane dla znajdowania istotnych zdań w tekście.

Budowa takiego łańcucha rozpoczyna się od selekcji słów, które będą włączone do analizy. W najprostszej strategii, brane pod uwagę są jedynie rzeczowniki wykryte w tekście. Jadnak, aby uzyskać najlepsze wyniki, wszystkie części zdania powinny być wzięte pod uwagę. Co więcej, powinny zostać uwzględnione powiązania między synonimami, w czym mogą pomóc tezaurusy. Bardziej zaawansowane algorytmy, również radzą sobie z wieloznacznością.

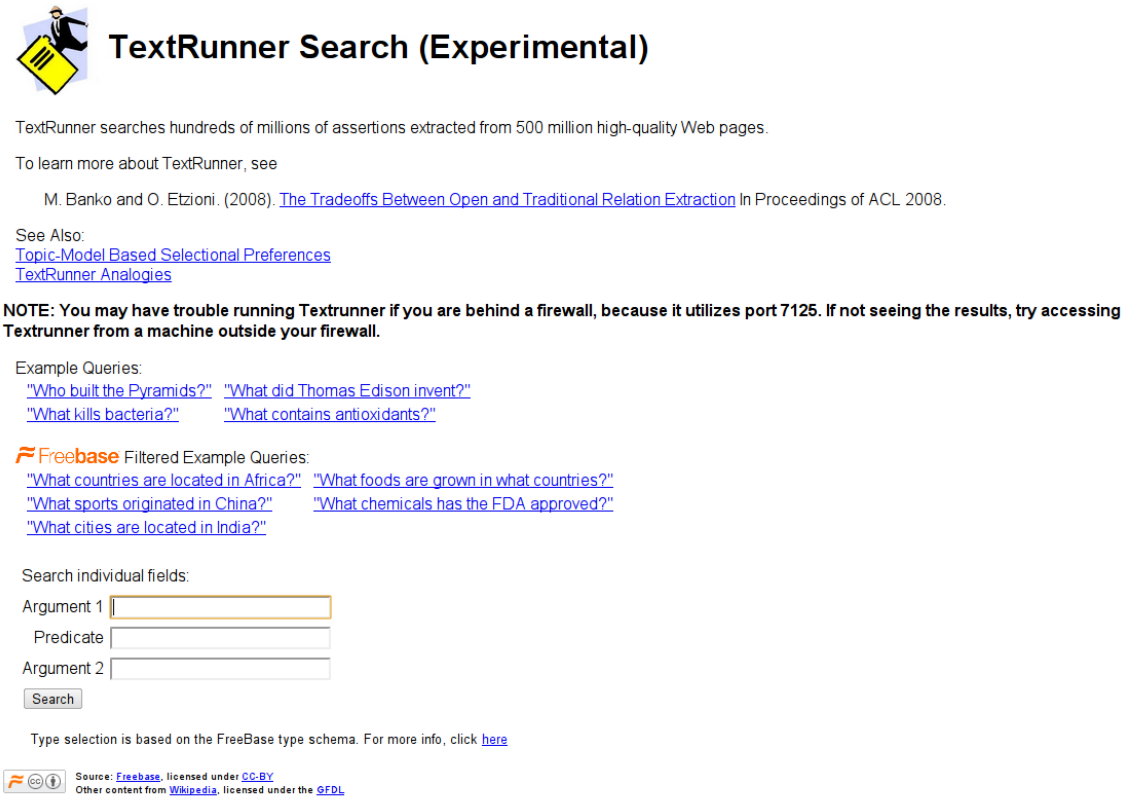
Metod, działających na tej zasadzie, często używa się do automatycznego generowania abstraktów. Przykładem może być tutaj projekt, opracowany przez przez Barzilay'a i Elhadada. Działanie algorytmu polega na tym, iż bierze on pierwsze słowa z najmocniejszych łańcuchów, a następnie ekstrahuje z dokumentu pierwsze zdania zawierające dany

wyraz. Siła łańcucha jest obliczana jako funkcja jego długości i częstości jego elementów.

Znaczna większość algorytmów, budowanych w celu wyszukiwania najważniejszych informacji z tekstu, bada częstość występowania słów kluczowych, natomiast wnikliwa analiza lingwistyczna pokazuje, iż najwartościowsze dla dokumentu są występujące w nim kolokacje.

3.3. Istniejące systemy ekstrakcji informacji

3.3.1. TextRunner



Rysunek 3.2. Główne okno aplikacji TextRunner.

TextRunner jest narzędziem działającym dla języka angielskiego i został stworzony na Uniwersytecie w Waszyngtonie. Tamtejsi naukowcy opracowali silnik wyszukiwania, który jak piszą, gromadzi skojarzenia i fakty, z ponad 500 milionów pojedynczych stron internetowych. Co więcej, narzędzie podobno ekstrahuje przy tym informacje z miliardów linii tekstu, analizując podstawowe związki językowe między wyrazami. Koncern internetowy Google przekazał projektowi ogromną bazę pojedynczych stron WWW, które TextRunner analizuje, co na pewno znacząco wpłynęło na wydajność

systemu, bo ominięta jest cała faza pobierania stron z Internetu [18].

TextRunner, na podstawie wprowadzonych danych, dokonuje przeszukania bazy stron internetowych, w poszukiwaniu fraz zawierających podane dane wejściowe. Nie wyszukuje on całych zdań, tylko ich fragmenty. Główną rolę odgrywają tutaj czasowniki, rzeczowniki i przymiotniki. Należy zwrócić uwagę na to, że system wyszukuje jedynie te części mowy, które znajdują się w bezpośrednim sąsiedztwie wprowadzonych informacji hasłowych. Jeżeli pewne informacje, w zadanym sąsiedztwie, występują dostatecznie często, tworzona jest z nich nowa grupa i dla każdej z grupy, przeprowadzane jest kolejne wyszukiwanie.

Fragment rezultatu zwróconego przez aplikację TextRunner znajduje się na rysunku 5.2. Po lewej stronie obecna jest lista grup, do których aplikacja przydzieliła odpowiednie frazy. Dla każdej z nich wyświetlany jest rezultat, zwrócony przez aplikację.

Zadanie to jest zdecydowanie prostsze dla języka angielskiego, niż dla języka polskiego, który posiada bardzo bogatą fleksję i znacznie bardziej skomplikowaną gramatykę. Powoduje to przede wszystkim to, że wyszukiwanie dokładnych dopasowań w tekście nie ma sensu. Wiązałoby się to z utratą ważnych informacji i prowadziło do uzyskania bardzo małej ilości wyników. Dla języka polskiego konieczne jest uwzględnienie pozostałych form fleksyjnych dla pojedynczych słów.

Rysunek 3.3. Wynik zwrócony przez aplikację TextRunner dla zapytania: machine vision.

3.3.2. Inne systemy ekstrakcji informacji

ATRANS – ekstrakcja informacji z prostych wiadomości teleksowych o przelewach bankowych. Oparty o proste przetwarzanie tekstu i ramkach skryptowych.

JASPER – ekstrakcja informacji o zarobkach z krótkich zdań z użyciem technik NLP.

SCISOR – ekstrakcja informacji z tekstów w Internecie. System stworzony dla ekstrakcji faktów dotyczących korporacji oraz informacji finansowych.

eAQUA - zaawansowany system ekstrakcji informacji z tekstów historycznych.

SProUT - ekstrakcja informacji z tekstów mammograficznych. Posiada wsparcie dla 11 języków, w tym także dla języka polskiego. Służy do automatycznego wydobycia informacji, zawartych w raportach mammograficznych.

FESTUS - system ekstrakcji najważniejszych informacji z tekstów w języku angielskim lub japońskim. *http://www.ai.sri.com/ appelt/fastus.html*

PROTEUS - Projekt rozwijany na Uniwersytecie w Nowym Yorku. Jego celem jest skonstuowanie systemu ekstrakcji informacji, który będzie potrafił znaleźć w tekstach informacje, wyekstrahować najważniejsze elementy i zaprezentować je w wybranym przez uzytkownika języku. *http://nlp.cs.nyu.edu/index.shtml*

4. Opis rozwiązania

4.1. Wstęp

Rozdział ten zawiera opis systemu, który powstał w ramach tej oto pracy dyplomowej. Zostały tutaj opisane główne moduły aplikacji, zgodnie z kolejnością ich wykonywania oraz wszystkie problemy z jakimi zetknąłem się podczas prowadzonych badań.

Na samym początku rozdziału warto podkreślić, że automatyczna analiza tekstu jest zagadnieniem dość skomplikowanym. Gramatyka języków naturalnych jest bardzo złożona i dostarcza ogromne bogactwo konstrukcji gramatycznych, trudnych do opisania i sklasyfikowania. Zagadnienie to jest dodatkowo trudniejsze dla języka polskiego z powodu bogatej fleksji, szczególnie trudnej gramatyki oraz dużej liczby wyjątków językowych. To powoduje, że nie udało się jeszcze stworzyć uniwersalnego mechanizmu, analizującego tekst.

Cechy charakterystyczne gramatyki języka polskiego sprawiają, że do pewnych zastosowań lepsze wydaje się podejście bazujące głównie na statystyce. Analiza tekstu służy wtedy do pozyskania danych ilościowych, opisujących częstotliwość pojawiania się pewnych połączeń pomiędzy słowami w tekstach pochodzących z różnych źródeł.

Podejście statystyczne ma podłoże kognitywne, bo jest próbą naśladowania metod nauki, jakie stosuje człowiek. Aby zapamiętać nowe informacje w ustalonej formie, człowiek często stosuje regułę powtórzeń. Polega ona na tym, że powtarzamy sobie pewne zdania kilka razy, aby dokładnie zapamiętać ich treść.

W pracy zostało wykorzystane podejście bazujące na statystyce, zaproponowane przez promotora pracy dra Adriana Horzyka. Jest to kontynuacja prac związanych z rozwojem metod bazujących na diagramach lingwistycznych neuroasocjacji, w postaci diagramów przyzwyczajeń lingwistycznych. Głównym celem pracy jest stworzenie algorytmu, który będzie samodzielnie wyszukiwał w zasobach internetowych zdania, opisujące definiowane znaczenie, a następnie zbuduje z nich maksymalnie duży graf LHG. Kolejno, na jego podstawie, algorytm ma wygenerować prostą definicję znaczenia.

Pracę można wstępnie podzielić na odrębne moduły, których realizacja zostanie opisana w niniejszym rozdziale:

- Specjalistyczny robot internetowy, którego głównym zadaniem ma być pozyskiwanie tekstu w postaci stron internetowych, zawierających zdania opisujące definiowane znaczenie.
- Moduł ekstrahujący zdania ze stron internetowych, składające się ze słów będących w relacjach leksykalnych oraz semantycznych z definiowanym znaczeniem.
- Moduł przetwarzający zdania, generujący graf lingwistycznych neuroasocjacji, w postaci grafu przyzwyczajeń lingwistycznych LHG.
- Moduł przetwarzający informacje ze zbudowanego grafu i generujący na jego podstawie uproszczoną definicję znaczenia.
- Graficzny interfejs użytkownika.

4.2. Robot internetowy

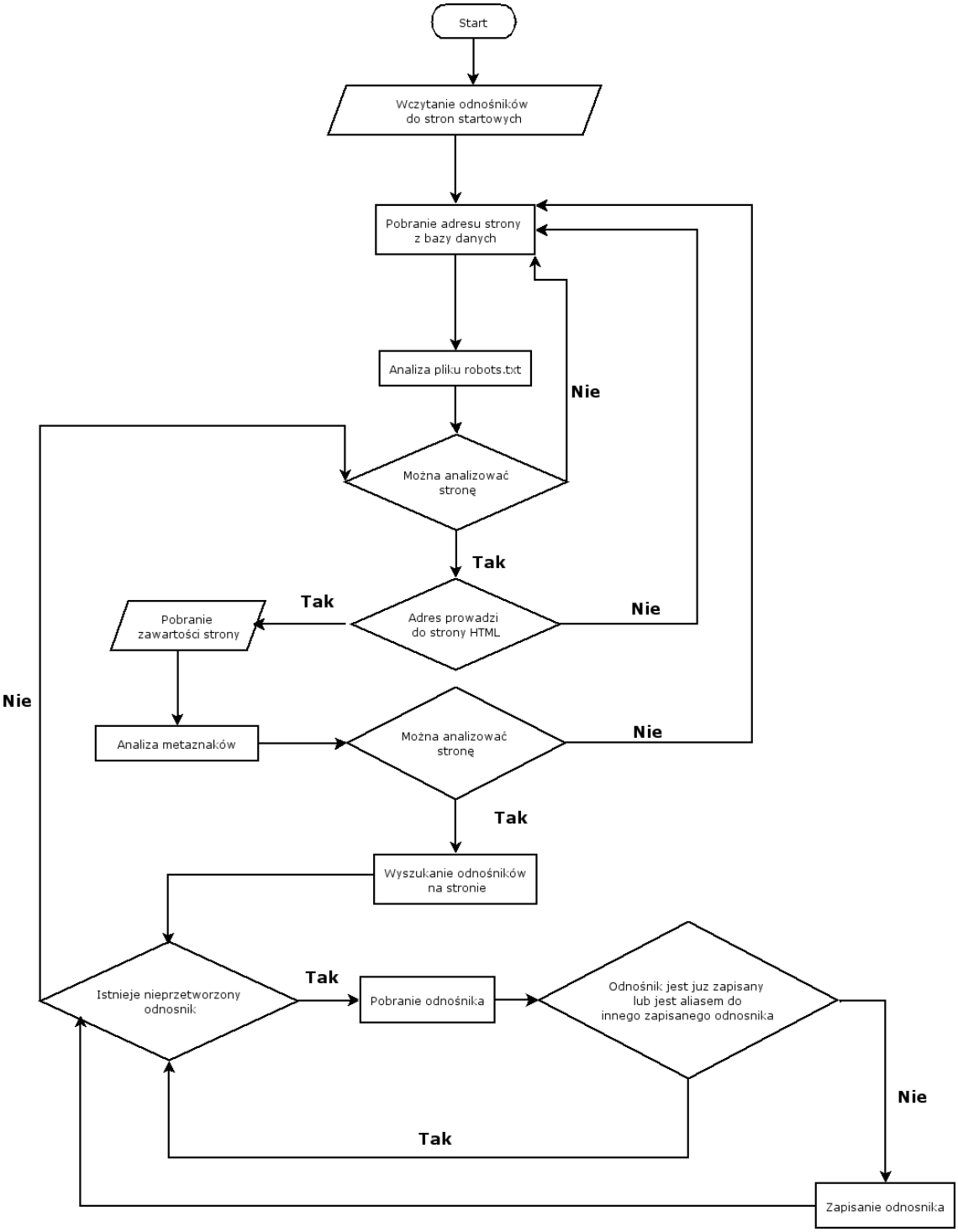
Roboty internetowe, zwane także pająkami internetowymi, stanowią podstawowy element działania wszystkich współczesnych wyszukiwarek. Algorytm ich działania jest bardzo prosty [2][3].

Przechodzą one po stronach internetowych zbierając informacje. Kiedy napotkają nową stronę, której jeszcze nie zindeksowały, najczęściej pobierają jej treść bez obrazków oraz zapisują znalezione odnośniki i przechodzą do analizy kolejnych stron.

Ciekawą informacją wydaje się być to, że wszystkie pająki internetowe, podobnie jak zwykli użytkownicy, pozostawiają po sobie ślad na serwerze i analizując pliki dziennika możemy znaleźć informację kiedy dana strona została odwiedzona. Pająki internetowe nie potrafią znaleźć stron, które nie są polinkowane z innymi i leżą na serwerze bez żadnych powiązań. Jako twórcy witryny internetowej możemy także sami zdecydować, czy chcemy aby odwiedzały one naszą stronę, czy też nie. Wykorzystuje się w tym celu specjalny plik, umieszczony na serwerze o nazwie: *robots.txt* lub definiując specjalne meta, znaki na stronie HTML.

Istnieje kilka znanych kwestii, o których warto wiedzieć, analizując działanie pająków internetowych:

- **Duplikaty stron** - pająki internetowe najczęściej posiadają bardzo skomplikowane mechanizmy rozpoznawania, czy pewna strona nie jest duplikatem innej, albo czy adres nie jest aliasem do już zindeksowanej strony, aby przechowywać ją tylko raz.



Rysunek 4.1. Algorytm działania typowego robota internetowego.

- **Zapętlanie się** - stanowi ono jeden z popularniejszych problemów. Ma ono miejsce w momencie, gdy robot zaczyna przechodzić po linkach z jednej strony na drugą i z powrotem.
- **JavaScript** - roboty mają duże problemy z analizą linków generowanych za pomocą skryptu.
- **Dodatkowy narzut** - roboty internetowe generują sztuczny ruch w sieci.

4.2.1. Własne rozwiązanie

Budowa i zasada działania pająków internetowych jest bardzo ciekawym zagadnieniem. Powstało już wiele wersji robotów, które są stosowane do różnych celów. Na potrzeby pracy został zaprojektowany i zaimplementowany algorytm takiego robota, wzbogacony w kilka dodatkowych modułów:

- Zaprojektowana została metoda inicjująca wyszukiwanie tekstu, oparta na najpopularniejszej wyszukiwarce internetowej google.com.
- Wykorzystane zostały bazy danych PostgreSql do przechowywania kodu stron oraz informacji o odwiedzonych stronach.
- Powstał mechanizm sprawdzający oraz zapewniający pobieranie stron wyłącznie w języku polskim.
- Robot został wzbogacony o moduł odpowiedzialny za usuwanie znaczników HTML z kodu strony.
- Dodano moduł przetwarzający zdania, generujący graf lingwistycznych neuroasocjacji w postaci grafu przyzwyczajeń lingwistycznych LHG.

Specyfika pracy wymaga bardziej skomplikowanego mechanizmu wyszukiwania nowych odnośników, co zostało dokładniej opisane w dalszej części tego rozdziału. Ponadto, dzięki własnej implementacji, mamy całkowitą kontrolę nad wszystkimi aspektami jego działania, co pozwala na wyciąganie bardziej szczegółowych wniosków.

Tworząc własne rozwiązanie, świadomie zostały pominięte kwestie, które nie mają większego wpływu na działanie robota i realizowanie celów wymienionych w temacie pracy:

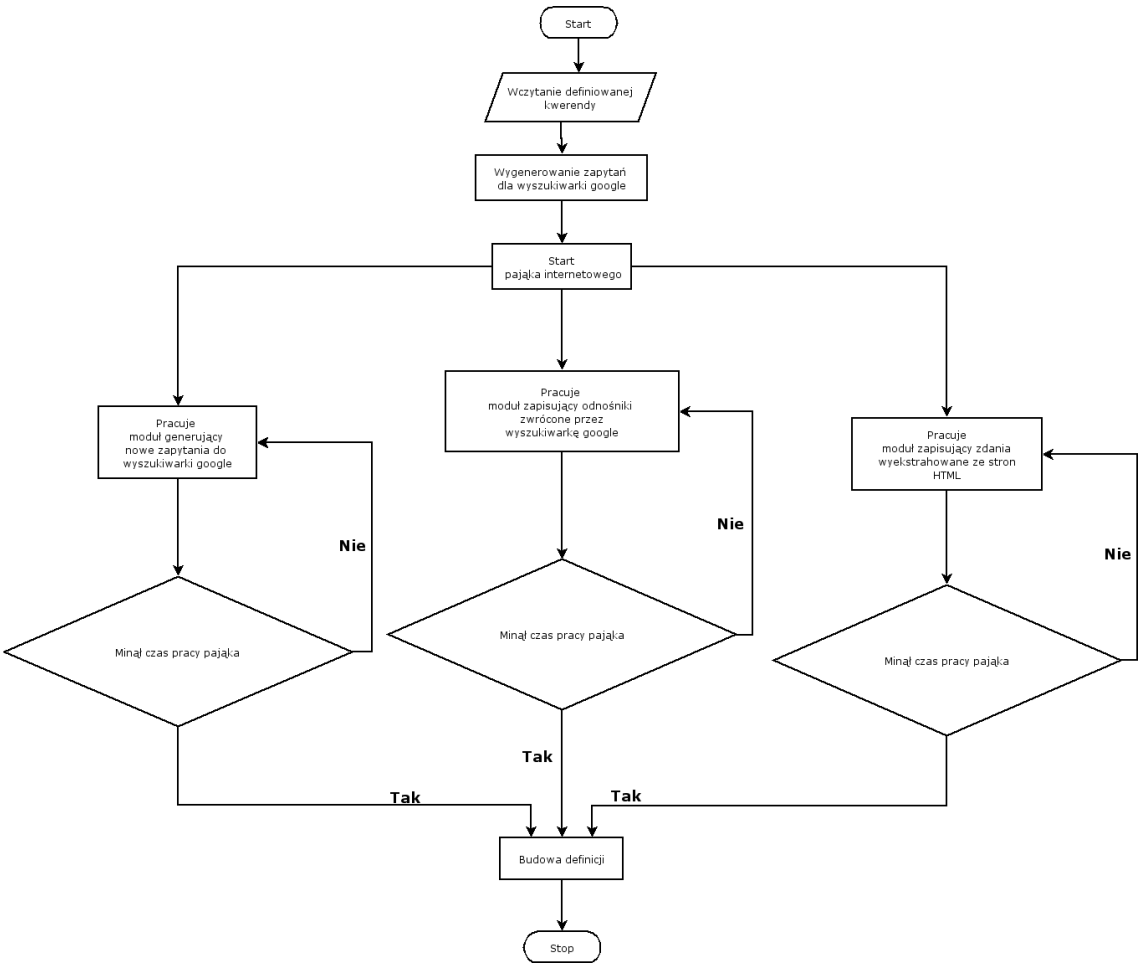
- robot nie interpretuje pliku *robots.txt* ani nie wyszukuje metaznaczników na stronie, co skutkuje tym, że zawsze wchodzi on tam, gdzie prowadzi odnośnik i wykonuje swoje zadania;
- robot na każdą stronę, pod danym adresem, wchodzi tylko raz. Jeżeli już wcześniej odwiedzał daną lokalizację, już więcej tam nie wejdzie, nawet gdy zmieni się kod tej strony;

- robot nie analizuje zmian treści na wcześniej odwiedzonych stronach.

Powyższe ograniczenia nie wpływają na działanie pająka internetowego, natomiast są to zagadnienia skomplikowane i dlatego zostały pominięte.

4.2.2. Algorytm działania aplikacji

Dzięki własnemu rozwiązaniu, znamy dokładny algorytm poruszania się robota po stronach internetowych i dzięki temu, w zależności od potrzeby, możemy go modyfikować.



Rysunek 4.2. Algorytm działania robota internetowego zaimplementowanego w ramach pracy.

Rozpoczęcie wyszukiwania

Robot internetowy, w pierwszej kolejności, uzupełnia sobie bazę adresów URL i rozpoczyna analizę treści umieszczonych w niej stron startowych. Jest to bardzo ważny etap, ponieważ mamy tutaj pełną kontrolę nad tym, od jakich stron zostanie rozpoczęte przeszukiwanie. Pojawia się tutaj automatycznie problem ustawiania takich

stron, bo przecież tylko niektóre, z dostępnych w sieci, będą zawierały istotne informacje.

Przeprowadzone w ramach pracy doświadczenia wykazały, że wybór stron startowych ma ogromny wpływ na kierunek poszukiwań, w jakim będzie zmierzał pająk internetowy. Wnioski są w tym wypadku bardzo logiczne. Jeżeli źle dobierzemy punkt startowy, robot zacznie przetwarzać strony, których treść jest całkowicie nieprzydatna w kontekście aktualnych poszukiwań, a mimo to algorytm będzie przetwarzał ogromne ilości danych tekstowych. Jest to efekt tego, że nie mamy możliwości sprawdzenia, czy dany adres prowadzi do strony zawierającej interesujące informacje.

Zaproponowane rozwiązanie wykorzystuje w tym celu współcześnie najlepszą wyszukiwarkę stron internetowych: google.com. Na początku wykorzystywane były też takie jak bing.pl, netsprint.com, szukacz.pl i inne, jednak jakość wyszukań była bardzo słaba.

Z podanego na początku działania algorytmu frazy oraz reguł służących do wyszukiwania informacji, które zostaną opisane w kolejnych sekcjach rozdziału, zostaje skonstruowane odpowiednie zapytanie, wysłane metodą GET do wyszukiwarki.

Podczas testów natknąłem się na problem blokowania automatycznych zapytań do wyszukiwarki google.com. Jest to najprawdopodobniej zabezpieczenie stosowane w celu ochrony zasobów wyszukiwarki. Udało się go pokonać wykorzystując serwis onet.pl, który używa wyszukiwarki firmy Google, natomiast nie posiada już powyższych blokad.

Przykładowe zapytanie dla słowa "kot"do wyszukiwarki ma postać:

```
http://szukaj.onet.pl/query.html?lr=Polska&qt=kot&p=1
```

Po otrzymaniu wyniku z wyszukiwarek, kod strony jest parsowany, analizowany, a następnie zapisywane są pozyskane adresy URL. Będą one służyły do wyszukiwania informacji.

Na zwróconej stronie znajduje się standardowo dziesięć odnośników. Takich stron może być dużo - nawet kilkadziesiąt. Aby zapisać wszystkie, robot buduje odpowiednie zapytanie dla każdej strony z wynikami. W pracy to użytkownik ma możliwość zdefiniowania ile stron ma przeglądać robot. Dodatkowo w zapytaniu dodany jest parametr *lr=Polska*, który oznacza, że chcemy uzyskać odnośniki prowadzące wyłącznie do stron napisanych w języku polskim.

Rozpoznawanie odpowiednich połączeń fleksyjnych

Biblioteka lingwistyczna CLP posiada możliwości zwracania wszystkich form fleksyjnych dla wybranego słowa. Nie mamy natomiast możliwości sprawdzenia, czy dana forma bezpośrednio łączy się z podanym słowem podstawowym. Jest to ważne w momencie inicjalizacji wyszukiwarki. Chcemy, aby generowane zapytania, z których będziemy pobierać linki do kolejnych stron, były poprawne, bo to wpływa na wydajność algorytmu.

Problem udało się rozwiązać w stopniu wystarczającym przy wykorzystaniu wyszukiwarki google.com. Konstruowane są zapytania GET i pobierana jest częstotliwość pojawienia się danej frazy, skonstruowanej z połączenia różnych form fleksyjnych kolejnych słów definiowanego hasła. Jeżeli jest ona większa od pewnej zadanej wartości, połączenie uznajemy za prawidłowe.

Analiza zawartości stron z pobranych adresów URL

Równolegle rozpoczyna się analiza zawartości stron, znajdujących się pod zapisanymi przez inicjalizator, adresami. Moduł ten jest wielowątkowy, aby poprawić wydajność algorytmu, ponieważ testy pokazały, że najdłużej trwa pobieranie zawartości strony po wysłaniu zapytania. To powoduje, że przez większą część pracy procesor jest nieużywany, a czas działania algorytmu znacznie się wydłuża.

Pracują tutaj dwa moduły. Pierwszy, zajmuje się analizą wyników, zwracanych przez wyszukiwarkę google.com, ekstrahuje odnośniki do stron internetowych i zapisuj je do bazy danych. Wcześniej zapisane linki nie są powielane - mamy zatem gwarancję, że nie przetwarzamy wielokrotnie tych samych danych. Drugi moduł zajmuje się analizą stron, znajdujących się pod tymi linkami. Na tym etapie następuje odrzucenie znaczników HTML z kodu strony. Ważną cechą, w odniesieniu do klasycznych pająków internetowych, jest to, że odnośniki znajdujące się na stronie, nie są zapisywane. Testy pokazały, że algorytm staje się mniej wydajny, gdy zaczniemy analizować kolejno zagnieżdżone strony. Spowodowane jest to tym, że bardzo często nie zawierają one interesującej nas treści, w kontekście aktualnego wyszukiwania, a jedynie wydłużają czas pracy aplikacji.

Budowanie zapytań do wyszukiwarki

Dostarczaniem nowych zapytań do wyszukiwarki zajmuje się specjalny moduł. Działa on niezależnie od innych i jest uruchamiany co pewien czas. Wyszukuje on najbardziej popularne połączenia słowne z szukaną frazą i używa ich do wygenerowania nowego zapytania. Takie podejście pozwala na wyszukiwanie bardziej szczegółowych informacji dotyczących specjalistycznych asocjacji słownych, gwarantując przy tym, że odnalezione

odnośniki będą zawierały szukane dane. Przykład takiego zapytania dla asocjacji *kot perski*, gdzie słowo *kot* stanowi tutaj definiowaną kwerendę:

```
http://szukaj.onet.pl/query.html?lr=Polska&qt=kot+perski&p=1
```

Sprawdzenie języka strony

Na początku analizy pobranej strony, należy sprawdzić język, w jakiej została ona napisana. W książce [1] został zaproponowany algorytm, bazujący na histogramie, stworzonym z analizy statystycznej słów wchodzących w skład treści strony. Metoda ta daje dużą pewność, jednak jest czasochłonna i w budowanym systemie może powodować nadmierne opóźnienia.

Inną metodę zaproponował mgr Marcin Gadamer w swojej pracy magisterskiej. Jego rozwiązanie bazuje na analizie częstotliwości wystąpień, charakterystycznych dla języka polskiego, znaków alfabetu. Polega ona na wyszukiwaniu w tekście, znaków: ’ą’,’ę’,’ć’,’ś’,’ł’,’ń’. I jeżeli stwierdzimy, że stanowią one określoną część zawartości strony, uznajemy test za zdany prawidłowo.

Innym skutecznym mechanizmem mogłaby być również analiza słów za pomocą słownika fleksyjnego języka polskiego.

Usuwanie znaczników HTML(HTML Stripping)

Kolejnym krokiem jest usunięcie wszystkich znaczników HTML użytych na stronie, w celu pozyskania samego tekstu. Jedną z propozycji wykonania takiego filtrowania mogłoby być użycie wyrażeń regularnych. Jednak ma to swoje wady, bo jest to zagadnienie kłopotliwe. Idealnym rozwiązaniem byłoby pozyskanie tekstu, gdyby znajdował się on w pewnych specjalnych znacznikach. Niestety większość stron, które występują w sieci, jest niezgodna ze standardami. Znaczniki, które miały służyć do innych celów, są używane do wprowadzania tekstu. To powoduje, że wyekstrahowanie pełnych bloków ze stron jest bardzo trudne.

Proste usunięcie tagów HTML powoduje powstanie bardzo niekorzystnego zjawiska zaszumienia kontekstu wypowiedzi i pomieszania różnych zdań ze sobą. Weźmy na przykład prostą stronę, która posiada menu zakładkowe. Gdy usuniemy znaczniki HTML, z treści zakładek powstaje ciąg słów. Następnie jeżeli taka strona okaże się dużym serwisem informacyjnym, z którego pobierzemy dużą liczbę adresów i wszystkie przeanalizujemy, okaże się, że powstały przypadkowo ciąg słów, stanie się statystycznie bardzo popularny. Zjawisko to może bardzo wypaczyć wyniki działania algorytmu, bo głównym elementem, na którym bazujemy, są dane ilościowe.

Zaproponowane rozwiązanie wykorzystuje parser kodu HTML. Jego działanie jest następujące:

1. Parser analizuje cały kod strony.
2. Wyszukujemy znaczniki, które spełniają jedno z poniższych założeń. Jeżeli jakiś rodzic został wybrany, automatycznie jego dzieci już nie mogą być wybierane, bo to spowodowałoby zwielokrotnienie tych samych danych.
   - posiadają dzieci służące do wprowadzania tekstu jak: $< center >, < a >, < b >$,
   - nie posiadają dzieci w postaci obiektów HTML,
   - posiadają dzieci, ale nie zawierają one żadnego tekstu,
3. dla tak wybranych znaczników, usuwany jest kod HTML i powstały w ten sposób tekst, jest następnie przetwarzany w celu pozyskania zdań.

Ekstrakcja zdań

Proces bardzo często nazywany tokenizacją. Polega na podziale bloku tekstu na zdania oraz następnie na słowa. Temat, dla człowieka banalnie prosty, okazuje się być dość trudny dla komputera.

Najprostsze podejście bazuje na twierdzeniu, że zdanie to ciąg słów, zakończony znakami '.','!' lub '?'. Założenie to ma oczywiste wady:

- kropka występuje także w skrótach, adresach internetowych oraz adresach mailowych,
- często zdania maja budowę hierarchiczną, co oznacza, że jedynie połączone ze sobą, niosą prawdziwą informację,
- zdania złożone zawierają także znaki '-', ';', ':'.

Trudno jest pominąć powyższe cechy języka, bo są one bardzo popularne i występują powszechnie w języku polskim. Takie zaniedbanie, może być przyczyną uzyskania błędnych wyników ekstrakcji informacji.

Kolejne podejście zakłada dodanie pewnych udoskonaleń do powyższego założenia. Po dokonaniu wstępnego podziału zdań względem znaków kończących, algorytm wykonuje dodatkowe czynności. Na początku uwzględniane są cytowania. Granica końca i początku zdania jest odpowiednio przesuwana. Następnie kasujemy podział na zdania w następujących przypadkach:

- jeśli jest on poprzedzony znanym skrótem, po którym występuje zwykle znak spacji, a po nim nazwa własna – np. Prof. lub vs.

- jeśli jest on poprzedzony znanym skrótem, po którym nie występuje słowo rozpoczęte wielką literą,
- jeśli podział zdania wynikał z wystąpienia '!' lub '?' oraz następuje po nim mała litera.

Inne metody są już bardzo skomplikowane, ale przynoszą dzięki temu lepsze rezultaty. Najpopularniejsze z nich to:

- drzewa decyzyjne (Riley, 1989),
- sieci neuronowe (Hearst, 1997).

Ekstrakcja zdań - własne rozwiązanie

Temat pracy zakłada budowę algorytmu, który sam wyszukuje sobie teksty, zawierające opis definiowanego znaczenia, a więc nie posiadamy wcześniej zdefiniowanego tekstu. Co więcej, informacje mają pochodzić ze stron HTML, a to dodatkowo utrudnia proces ekstrakcji pełnych zdań. Jak wykazały wykonane w ramach niniejszej pracy testy, strony internetowe zawierają bardzo często zdania niedokończone, źle skonstruowane, nieposiadające zakończeń w postaci znaków interpunkcyjnych. Często szukana fraza występuje w nagłówkach sklepów internetowych, serwisów aukcyjnych, reklamach, czy też na forach. Informacje te są wartościowe, jednak bardzo trudno jest wyodrębnić z takich stron osobne zdania.

Te cechy spowodowały, że w ramach pracy został zbudowany algorytm ekstrakcji jedynie fraz ze stron HTML, które nie zawsze są zdaniami, a mamy pewność, że zawierają informacje o definiowanej kwerendzie.

W pracy wykorzystano podejście pierwsze, opisane we wcześniejszym podrozdziale, z tym, że dodatkowo uznajemy, że znaki ',',':',';' rozdzielają kolejne frazy. Dzięki temu wyszukujemy krótkich fraz, ale mamy większą pewność, że są one rzeczywiście wartościowe.

4.2.3. Selekcja zdań

Po wstępnym przetworzeniu tekstu, następuje selekcja tych zdań, które zawierają wyszukiwaną przez system definicyjny kwerendę lub którąś z jej form fleksyjnych. Przyjmujemy tutaj założenie, że takie zdania zawierają pewne ważne informacje o wyszukiwanej kwerendzie. Kolejno są one przekazywane do modułu odpowiedzialnego za budowę diagramów LHG.

4.2.4. Budowa grafu LHG

Następnym etapem pracy było zbudowanie maksymalnie rozbudowanego grafu przyzwyczajeń lingwistycznych LHG.

Grafy LHG

Grafem G nazywamy strukturę składającą się z uporządkowanej pary $G = (V, E)$ gdzie:

V - jest to niepusty zbiór elementów, które nazywamy **wierzchołkami** grafu,
E - jest to zbiór dwuelementowych podzbiorów zbioru wierzchołków V, które nazywamy **krawędziami** grafu, czyli $E \subseteq u, v : u, v \in V, u \neq v$

Dodatkowo jeżeli dwuelementowe podzbiory zbioru V zawierają elementy uporządkowane, to takie grafy nazywa się **skierowanymi**. Przyjmuje się, że krawędź $e = (x, y)$ jest skierowana z x do y, czyli wychodzi z wierzchołka x, a wchodzi do wierzchołka y. Jeżeli kolejność występowania wierzchołków w parach nie jest ważna, to grafy takie nazywamy **nieskierowanymi**.

Grafy LHG są uproszczoną wersją Grafów Lingwistycznych Neuroasocjacji Semantycznych (GLAS), które zostały zaprojektowane przez promotora niniejszej pracy dra Adriana Horzyka. Ich najważniejszą cechą jest utrzymywanie kontekstu całej wypowiedzi. Mają one także odwzorowywać konstrukcję poprawnych zdań w wymiarze fleksyjno - częstotliwościowym oraz umożliwić algorytmom lingwistycznym budowanie poprawnych wypowiedzi, dostarczając wzorce konstrukcyjne oraz niezbędną do tego wiedzę [2].

Budowę grafu LHG rozpoczyna się od grafu pustego. Następnie dokonuje się podziału zdań na słowa, które stanowią wierzchołki. Żadne słowo nie może się powtórzyć, ale mogą one być jakąś formą fleksyjną słów znajdujących się na grafie. Dodatkowo każdemu wierzchołkowi dodaje się etykietę, zawierającą części mowy, do jakich może należeć słowo.

Bardzo ważną część grafu stanowią krawędzie, które powstają w wyniku parsowania zdań. Stanowią one odzwierciedlenie połączeń, jakie mogą występować pomiędzy słowami i są one nazywane asocjacjami. Wyróżnić możemy następujące rodzaje asocjacji:

- *Asocjacje fleksyjne* - łączą słowa w tych samych formach fleksyjnych,
- *Asocjacje definicyjne, kategorii i zawierania* - łączą słowa opisujące sie nawzajem,
- *Asocjacje działania* - opisują czynności i działania,

- *Asocjacje liczebności* - zawierają opisy mnogości, liczebności zakresów,
- *Asocjacje łączące czynności z działaniem* - łączą czynności z przedmiotem ich działania,
- *Asocjacje miejsca, czasu przyczyny i warunku działania* - opisują miejsca, czas, przyczyny i warunki działania.

Własne rozwiązanie

Grafy przyzwyczajeń lingwistycznych zawierają przede wszystkim informacje dotyczące słów oraz relacji, w jakie mogą one wchodzić z innymi słowami. Jeżeli przeanalizujemy odpowiednio dużą liczbę zdań, to jesteśmy w stanie ocenić z dużą skutecznością, bazując na statystyce, które słowa oraz połączenia są językowo poprawne. Dodatkowo możemy spróbować wyróżnić pewne asocjacje i określić funkcje, jakie pełnią w zadanym kontekście.

Aby wykorzystać tak zbudowane grafy do zapamiętywania i odtwarzania informacji, posiadają one zdolność zapamiętywania ścieżek, będących sekwencją połączeń miedzy słowami, które odwzorowują rzeczywiste zdania. Dla każdej asocjacji definiuje się zbiór par, gdzie pierwsza wartość opisuje zdanie do którego należy, natomiast druga pozycję w tym zdaniu.

Bardzo rzadko informacje możemy zapisać za pomocą pojedynczych asocjacji słownych, a wręcz przeciwnie, z reguły składają się one z sekwencji słów, które dopiero odpowiednio zestawione ze sobą mają sens. Weźmy następujący przykład zdań statystycznie prawdopodobnych:

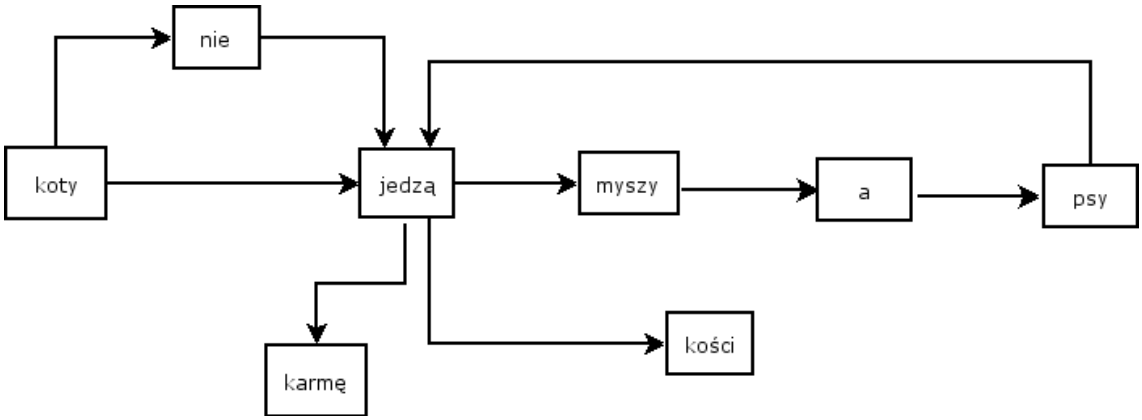
Koty jedzą myszy, a psy jedzą kości.
Koty nie jedzą kości.
Koty jedzą karmę.

Zdania te zapisane na diagramie LHG zostały zaprezentowane na rysunku 4.3.

Analizując diagram, powstały w wyniku rozkładu powyższych zdań, możemy dojść do błędnych wniosków:

Koty nie jedzą myszy.
Koty jedzą kości.

Posiadając jedynie informacje dotyczące połączeń między dwoma słowami, nie jesteśmy w stanie odtworzyć prawdziwej informacji, zapisanej na diagramie LHG.



Rysunek 4.3. Przykład prostego grafu LHG.

Niezbędne jest tutaj posiadanie wiedzy o pełnej sekwencji słów.

Zaletą grafów LHG jest możliwość zbadania, które elementy zapisanych ciągów najczęściej występują w zbiorach tekstów oraz odrzucenie tych mało popularnych, a więc najprawdopodobniej błędnych.

Grafy LHG posiadają także pewne wady. Najbardziej uciążliwa jest konieczność przetworzenia bardzo dużej ilości danych, aby uzyskać wystarczającą ilość informacji statystycznych. Dodatkowo, każda zmiana algorytmu wyszukującego i sprawdzającego nowe zdania wymaga przebudowania grafu. Niestety, jest to bardzo czasochłonne i czasami potrzeba kilku dni, aby uzbierać wystarczającą ilość danych.

4.3. Algorytm ekstrakcji informacji

Po skonstruowaniu diagramów przyzwyczajeń lingwistycznych, kolejnym etapem pracy była budowa modułu, odpowiedzialnego za ekstrakcję informacji. Najbardziej pożądaną cechą, byłaby z pewnością uniwersalność zaproponowanej metody, aby działała ona prawidłowo dla różnych danych.

Większość współczesnych algorytmów służy do wyszukiwania pewnych specyficznych informacji, charakterystycznych dla zadania, jakie mają do wykonania. Najczęściej struktura przeglądanych bloków tekstu jest wcześniej znana lub posiadamy wiedzę na temat tego, jakie informacje chcemy wyekstrahować z tekstu. Bardzo często posiadają one bazę wzorców do których prosty algorytm dopasowuje słowa znalezione w tekście. Z racji tego, że temat pracy zakłada budowę uniwersalnego algorytmu, to podejście niestety jest zbyt proste, ponieważ nie możemy zdefiniować wzorców dla wszystkich tematów.

Możemy także wyróżnić algorytmy służące do generowania abstraktów, których efektem działania ma być usystematyzowany skrót najważniejszych informacji z tekstu. Zadanie to wydaje się być bardzo podobne do tego, co chcemy osiągnąć w tej oto pracy. Niestety algorytmy takie bardzo często działają na kompletnym tekście, z którego wybierają te informacje, które wydają się być najistotniejsze, bo statystycznie najczęściej występują w tekście.

Jak już było wcześniej wspomniane, praca zakłada budowę uniwersalnego narzędzia, które będzie w stanie samodzielnie wyszukać najważniejsze informacje i stworzyć pewien opis, zawierający odpowiedzi na najważniejsze pytania, dotyczące opisywanego hasła.

4.3.1. Wyszukiwanie części mowy

Analizując grafy LHG okazuje się, że możemy bez większych problemów wyekstrahować zbiór słów, które najczęściej występują w zdaniach, obok definiowanej kwerendy. Największe znaczenie mają na pewno przymiotniki, czasowniki oraz rzeczowniki i na nich skupiać się będzie algorytm w tej pracy.

Wyszukiwanie takich słów można przeprowadzić bardzo prosto. Poszukujemy wszystkich form, będących w relacji na diagramie LHG, z definiowaną kwerendą. W niektórych wypadkach warto także odróżnić te, które znajdują się przed i po, bo mogą znaczyć co innego.

Np. dla zdań:

Janek lubi koty.

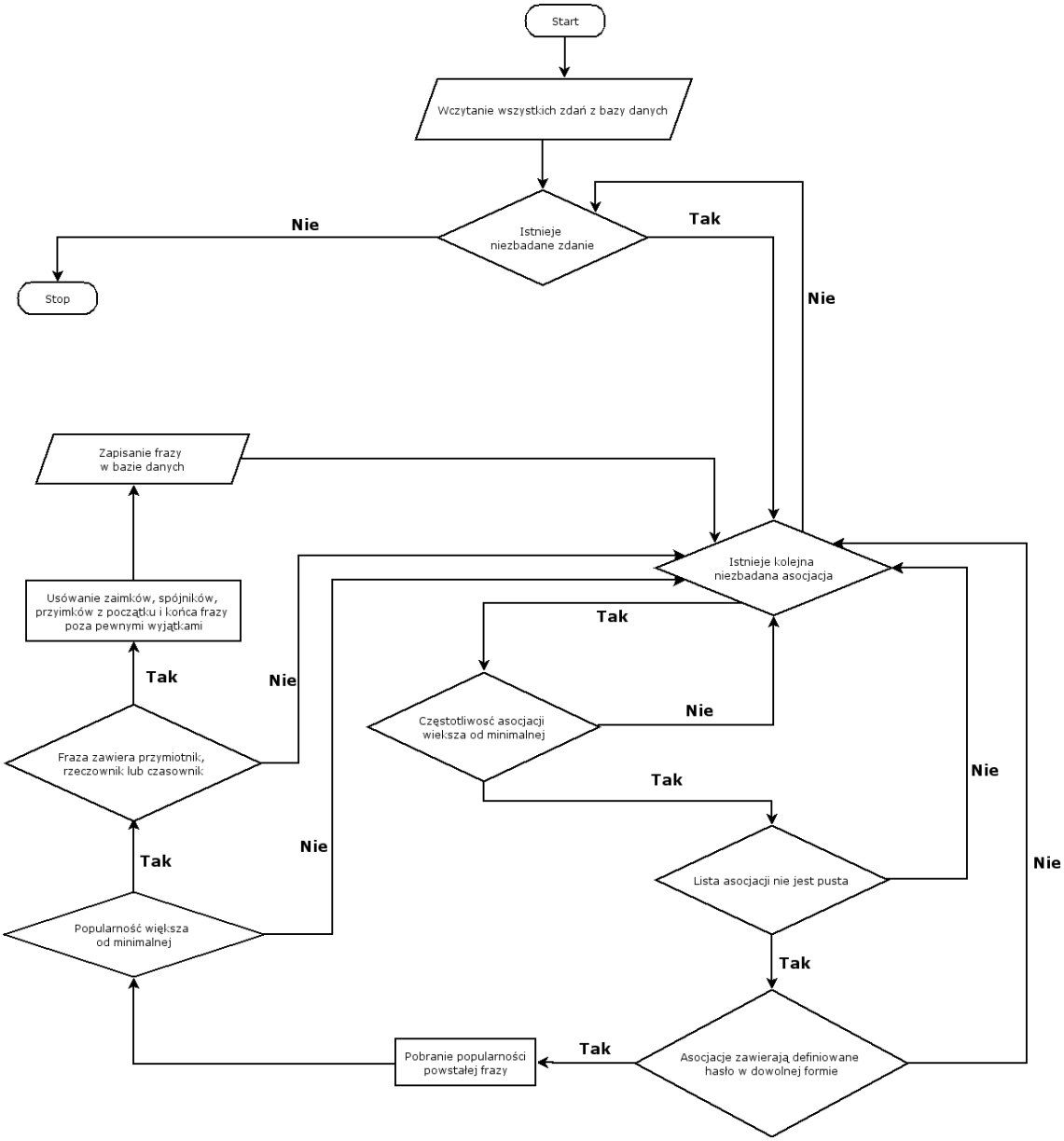
Koty lubią mleko.

Czasownik *lubi* oznacza zupełnie co innego. Koty mogą być *"lubiane"* oraz koty potrafią *"lubić"*.

4.3.2. Algorytm ekstrakcji informacji - schemat blokowy

Niżej znajduje się schemat blokowy oraz opis działania algorytmu ekstrakcji informacji, który został wypracowany w ramach niniejszej pracy.

1. Pobierane są wszystkie frazy wyekstrahowane z tekstów przez robota internetowego.
2. Dla każdej frazy, następuje analiza częstości występowania konkretnych połączeń pomiędzy słowami, zapisanymi w bazie danych.



Rysunek 4.4. Algorytm ekstrakcji informacji.

3. Jeżeli jest ona większa od pewnej minimalnej wartości, dane połączenie uznawane jest za ważne. W przeciwnym wypadku połączenie uznajemy za nieważnie.
4. Tworzone są listy zawierające ciągi połączeń ważnych.
5. Wybierane są te listy, które zawierają opisywaną kwerendę.
6. Następnie dla każdej z listy wybranych połączeń sprawdzamy, czy zawierają one przynajmniej czasownik, rzeczownik, lub przymiotnik. Pozostałe są odrzucane.
7. Kolejno dla każdej uzyskanej listy dokonujemy usunięcia słów występujących na początku i na końcu frazy, które są: spójnikami, zaimkami oraz przysłówkami, poza słowami ważnymi takimi jak: “nie”.
8. Tak uformowane frazy zostają zapytaniem GET przesyłane do wyszukiwarki google.com, w celu pobrania ilości otrzymanych wyników.
9. Frazy, które posiadają popularność większą od pewnej wartości, uznajemy za prawidłowe. Pozostałe odrzucamy.
10. Następnie usuwane są frazy, które są fragmentami innych fraz, a więc są albo niepełne, albo dublują inne dłuższe.

4.3.3. Formowanie definicji.

Budowa definicji polega na wyszukiwaniu w pobranych frazach przymiotników, czasowników oraz rzeczowników. Wyszukujemy takie części mowy, które stoją w sąsiedztwie definiowanej frazy i nie są oddzielone od niej rzeczownikami, bo mogą one stanowić nowy podmiot w danym zdaniu. Mogą one z kolei być rozdzielone z definiowana frazą innymi częściami mowy.

4.3.4. Problemy i trudności, do rozwiązania.

Ekstrakcja informacji jest zagadnieniem bardzo skomplikowanym. Wiele zagadnień wymaga szczegółowego dopracowania, na co potrzeba jeszcze dużo czasu i wielu testów. Najważniejsze problemy, dotykające tematykę niniejszej pracy, zostały wymienione w tym rozdziale.

Problem zdań pytających?

Wyszukiwanie informacji na zadany temat w Internecie, kiedy dodatkowo nie mamy żadnych informacji o strukturze tekstu jest bardzo trudne. Nadrzędnym problemem jest tutaj pozyskiwanie zdań. Często jednak ciężko jest oszacować, w którym miejscu znajduje się jego początek oraz koniec, bo posiadamy jedynie urywki tekstu. To często powoduje, że pewne fragmenty początku lub końca zdania są pomijane.

Pojawia się tutaj także problem oceny, czy dane zdanie zawiera informacje prawdziwe. Dotyczy to głównie zdań pytających, które przeważnie zawierają informacje niepotwierdzone, do których nie możemy mieć pewności, że są prawdziwe.

Przykładem takiego zdania może być:

Czy koty jedzą myszy?

Zdanie to nie niesie informacji dającej nam pewność, że *koty jedzą myszy*, natomiast algorytm ekstrakcji zinterpretuje je tak, jak zdanie twierdzące.

Zdania z przecinkami.

Innym, podobnym problemem jest kwestia informacji niesionych przez kilka zdań, występujących kolejno po sobie oraz zdań zawierających przecinki. O ile jeszcze te pierwsze można pominąć, to te drugie mogą być źródłem wielu problemów. Weźmy pod uwagę zdanie *"To nie pies, to człowiek."* i usuńmy z niego przecinek. Algorytm ekstrakcji zaklasyfikuje je jako zdanie twierdzące niosące informacje, że *"... pies to człowiek ... "*.

Niestety zdania tego typu, są na tyle często spotykane, że mogą generować znaczące problemy. Powstały algorytm dzieli takie zdania na mniejsze frazy, jednak to powoduje utratę często ważnych informacji.

Rozpoznawanie części mowy w zależności od kontekstu.

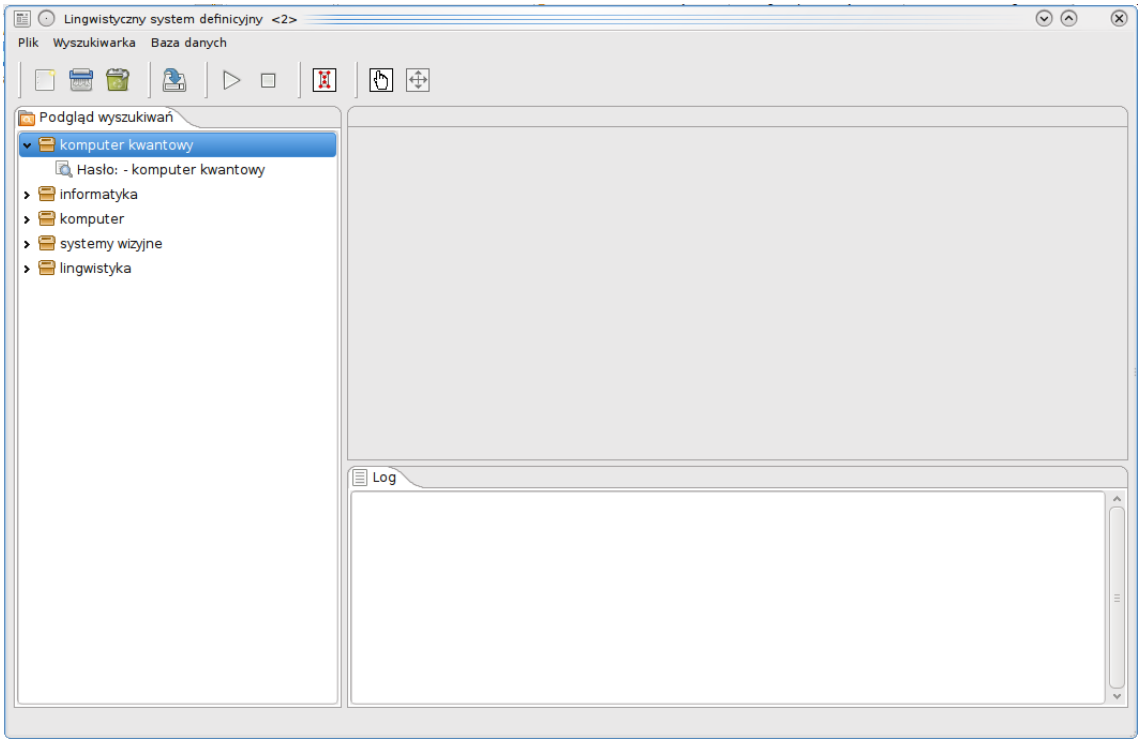
Niektóre słowa mogą być innymi częściami mowy w zależności od kontekstu, czasami bardzo szerokiego. To powoduje, że ciężko jest je rozróżnić. Słowa takie jak np. "jak", "młody", "stary" są wieloznaczne i trudno zbadać ich prawidłową formę, nie posiadając żadnej wiedzy o kontekście, w jakim one występują.

4.4. Prezentacja interfejsu użytkownika.

W ramach pracy powstała prosta aplikacja, napisana w języku programowania JAVA, która posiada zaimplementowany algorytm ekstrakcji informacji, skonstruowany zgodnie z opisem, zawartym w poprzedniej części rozdziału. W tym podrozdziale zawarta jest prezentacja interfejsu graficznego wraz z opisem jego funkcji.

4.4.1. Główne okno aplikacji.

Po uruchomieniu programu pojawia się główne okno aplikacji zaprezentowane na rysunku 4.5.



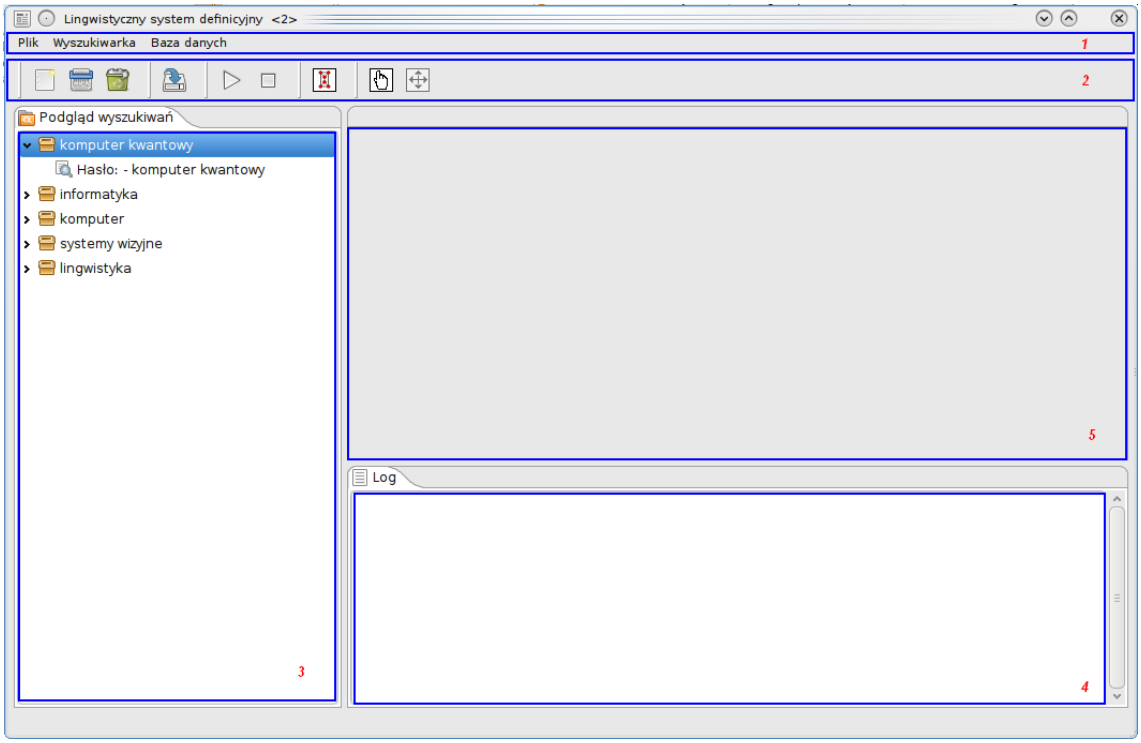
Rysunek 4.5. Główne okno aplikacji.

Główne okno aplikacji składa się z następujących elementów, które zostały zaznaczone na rysunku 4.6:

Menu główne (oznaczone numerem 1) - jest ono podzielone na następujące elementy menu: ***Plik***, ***Wyszukiwarka***, ***Baza danych***.

Menu: ***Plik*** posiada następujące elementy:

- *Nowe wyszukiwanie* - dodaje nowe hasło do wyszukiwarki,
- *Usuń wyszukiwanie* - całkowicie usuwa definiowane hasło z bazy danych,
- *Zapisz definicję* - zapisuje definicję w pliku HTML w podanej lokalizacji,



Rysunek 4.6. Elementy głównego okna aplikacji.

- *Zakończ* - kończy działanie programu.

Menu: ***Wyszukiwarka*** posiada następujące elementy:

- *Start* - rozpoczyna działanie algorytmu dla wybranego hasła,
- *Stop* - zatrzymuje działanie algorytmu dla wybranego hasła,
- *Parametry* - powoduje pojawienie się okna z parametrami algorytmu.

Menu: ***Baza danych*** posiada opcję:

- *Połączenie* - ustawia połączenie z bazą danych. Automatycznie wczytuje wszystkie hasła znajdujące się w bazie danych.

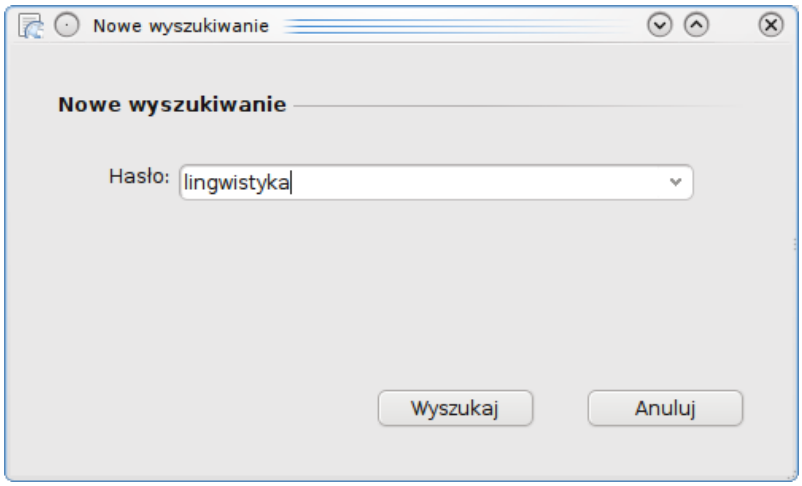
Paski narzędzi (oznaczone numerem 2) - znajdują się tutaj przyciski służące do obsługi aplikacji:

 - dodaje nowe hasło do wyszukiwarki,

 - całkowicie usuwa definiowane hasło z bazy danych,

 - usuwa aktualną definicję dla wybranego hasła,

 - zapisuje definicję dla wybranego hasła do pliku HTML,



Rysunek 4.7. Okno dodawania nowego hasła dla systemu.

- rozpoczyna działanie algorytmu dla wybranego hasła,
- zatrzymuje działanie algorytmu dla wybranego hasła,
- tworzy graf LHG dla wybranego hasła,
- włącza tryb pozwalający na przemieszczanie elementów grafu LHG,
- włącza tryb pozwalający na przeglądanie grafu LHG.

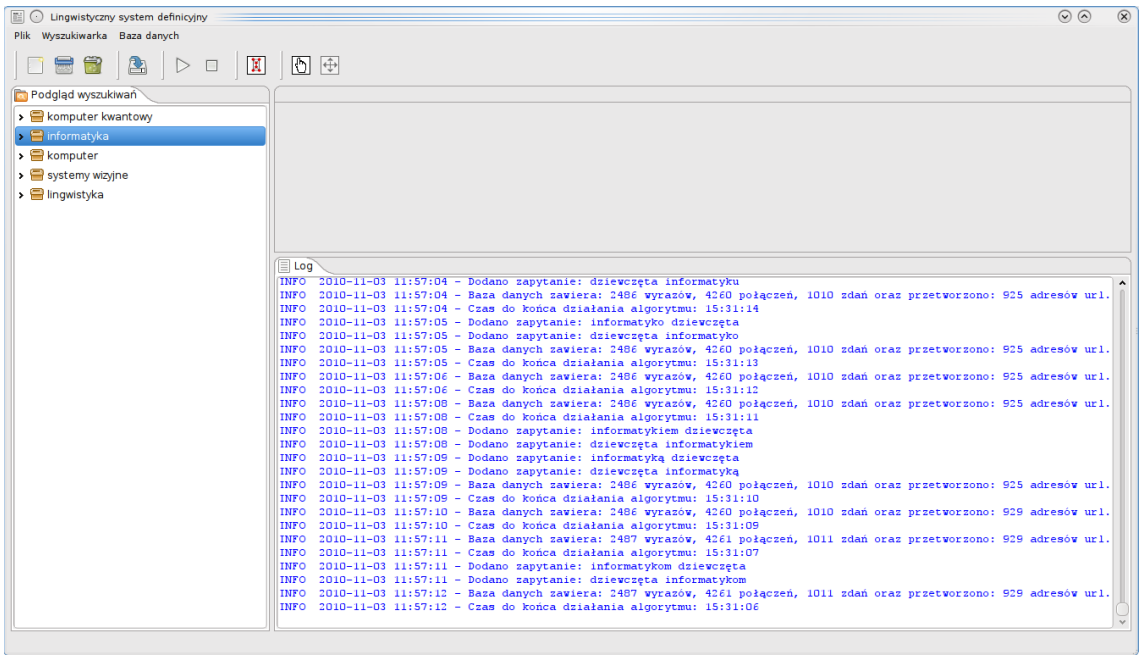
Okno: Podgląd wyszukiwań (oznaczone numerem 3) - zawiera drzewko haseł znajdujących się aktualnie w bazie danych. Gdy definicja zostanie wygenerowana, można ją otworzyć dwukrotnie klikając myszką na drzewku, na ikonie prezentującej kartkę z lupą. Z kolei, gdy wygenerujemy graf, zostanie on dodany do drzewka i można go otworzyć, klikając dwukrotnie na drzewku, na ikonie grafu.

Okno: Log (oznaczone numerem 4) - zawiera informacje o przebiegu pracy algorytmu.

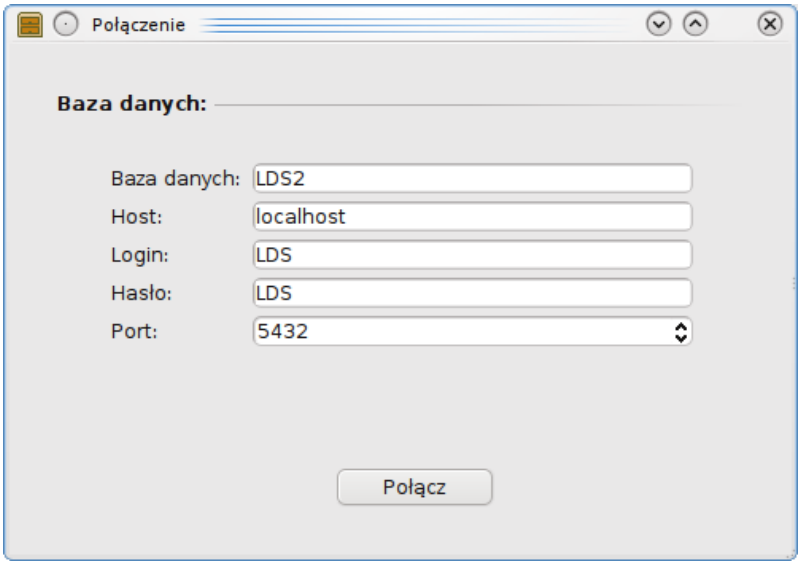
Zakładki definicji(oznaczone numerem 5) - otwierane są tutaj zakładki definicji oraz zakładki grafów LHG.

4.4.2. Połączenie z bazą danych.

Po wybraniu z menu *Baza danych* opcji *Połączenie* pojawia się okno zaprezentowane na rysunku 4.9. Możemy tutaj ustawić podstawowe parametry bazy danych. Po wciśnięciu przycisku zatwierdź, aplikacja automatycznie wczyta wszystkie definiowane hasła. Jeżeli w bazie nie znajdują się potrzebne tabele, aplikacja sama je sobie wygeneruje.



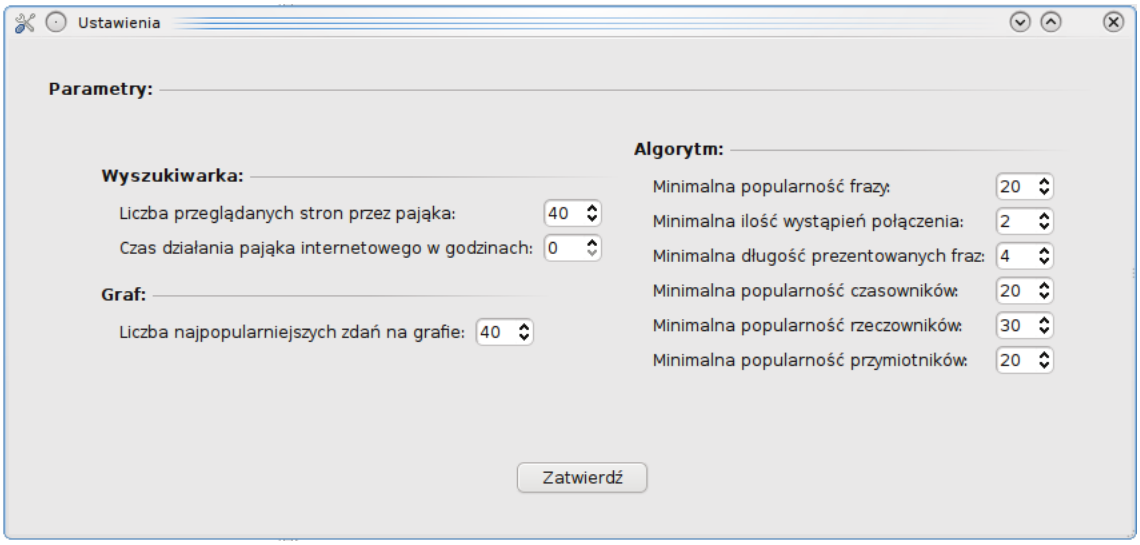
Rysunek 4.8. Log aplikacji w trakcie działania pająka internetowego.



Rysunek 4.9. Okno połączenia z bazą danych.

4.4.3. Definiowanie parametrów wykorzystywanych przez algorytm.

Menu *Wyszukiwarka* zawiera opcję *Parametry*. Po jej wybraniu pojawia się okno, zawierające wszystkie parametry, które wykorzystuje algorytm (rysunek 4.10), a które należy zdefiniować w celu wygenerowania definicji.



Rysunek 4.10. Okno ustawień parametrów działania algorytmu.

Parametry algorytmu:

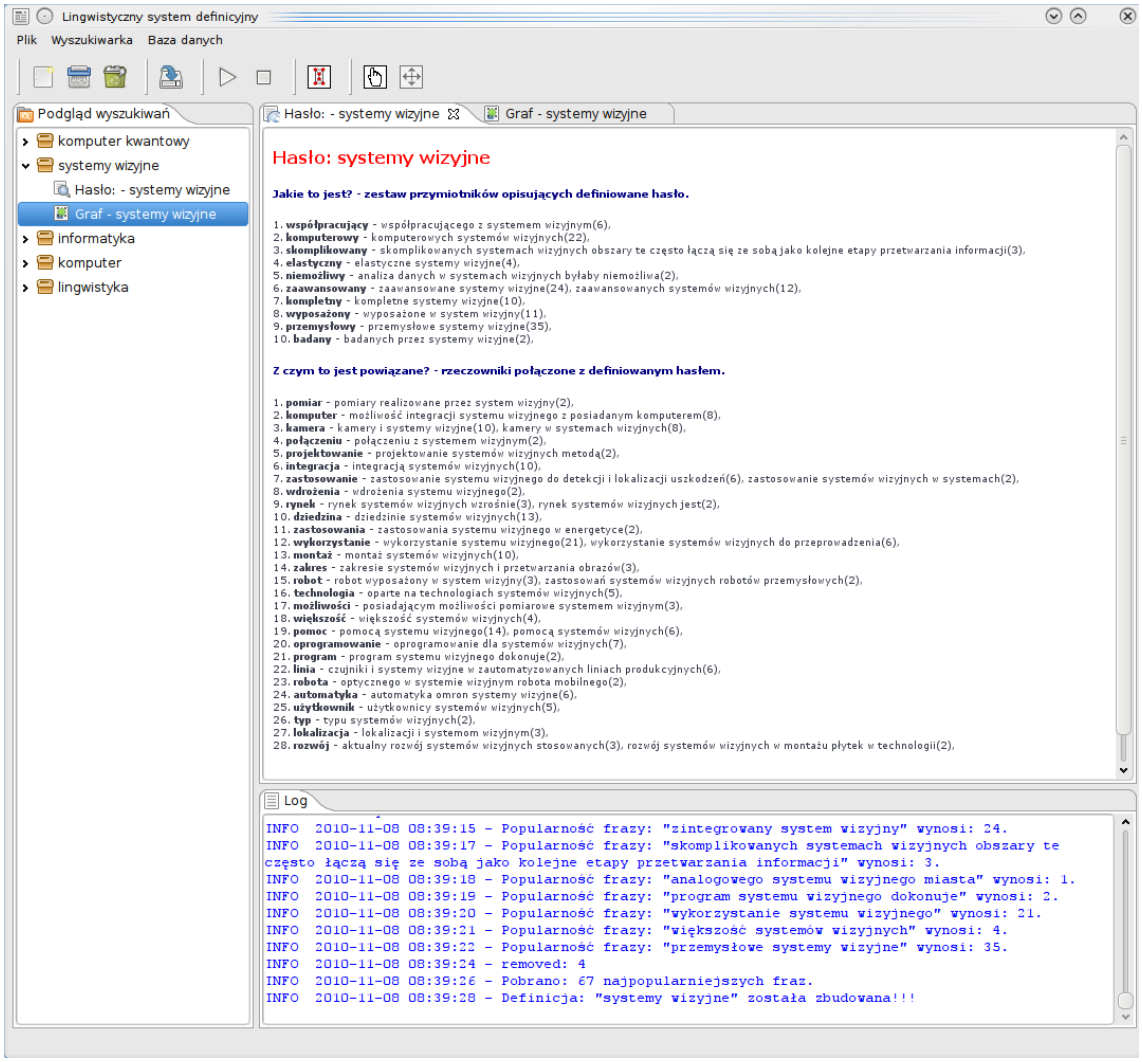
- **Liczba przeglądanych stron przez pająka** - parametr określa, ile stron z wynikami zwróconymi przez przeglądarkę analizuje pająk internetowy, w celu pobrania adresów stron.
- **Czas działania pająka internetowego w godzinach** - ilość godzin, od chwili uruchomienia algorytmu do momentu wygenerowania definicji.
- **Liczba najpopularniejszych zdań na grafie** - ilość najpopularniejszych zdań zaprezentowanych na grafie.
- **Minimalna popularność frazy** - popularność wyekstrahowanego fragmentu zdania, wyrażona jako ilość wyników zwróconych przez wyszukiwarkę.
- **Minimalna ilość wystąpień połączenia** - określa, ile razy dane połączenie ma się pojawić na diagramie LHG, aby uznać je za poprawne.
- **Minimalna długość prezentowanych zdań** - określa, ile jak długie frazy maja być zamieszczane w definicji. Parametr wykorzystywany do filtracji, w celu wyszukania dłuższych fraz dla danej definicji.
- **Minimalna popularność czasowników** - określa minimalną liczbę powtórzeń czasowników w sąsiedztwie definiowanego hasła na grafie LHG. Parametr wykorzystywany do filtracji, w celu wyszukania częściej występujących czasowników.

- **Minimalna popularność rzeczowników** - określa minimalną liczbę powtórzeń rzeczowników, w sąsiedztwie definiowanego hasła na grafie LHG. Parametr wykorzystywany do filtracji, w celu wyszukania częściej występujących rzeczowników.
- **Minimalna popularność przymiotników** - określa minimalną liczbę powtórzeń przymiotników, w sąsiedztwie definiowanego hasła na grafie LHG. Parametr wykorzystywany do filtracji, w celu wyszukania częściej występujących przymiotników.

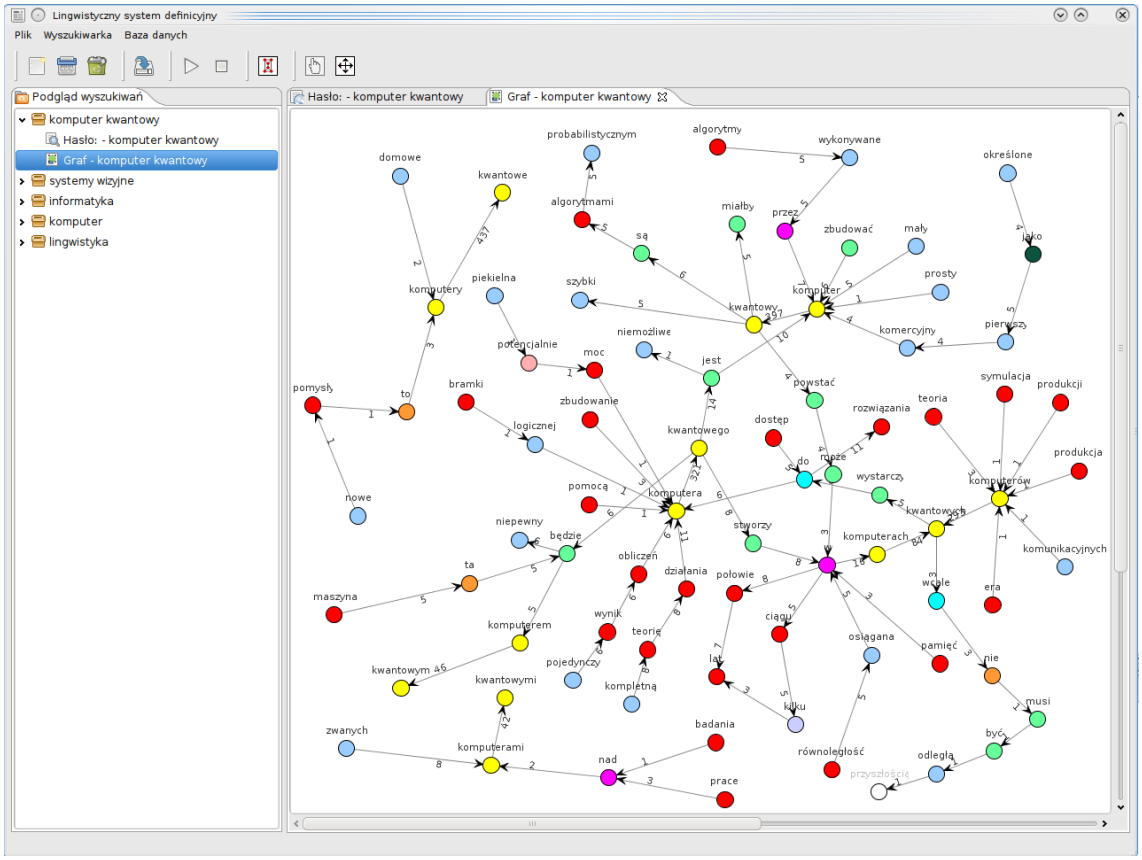
4.4.4. Budowanie definicji.

Scenariusz budowy definicji pewnego pojęcia za pomocą niniejszej aplikacji może wyglądać następująco:

1. Na samym początku należy utworzyć połączenie z bazą danych. W tym celu należy wybrać z menu *Baza danych* opcję *Ustawienia*, podać wymagane do połączenia parametry bazy danych i zatwierdzić je przyciskiem *Zatwierdź*.
2. Kolejno należy kliknąć na przycik *Nowe wyszukiwanie* i podać definiowane hasło. W tym momencie aplikacja zainicjalizuje początkowe zapytania dla wyszukiwarki. Może to potrwać kilka minut.
3. Dalej należy zdefiniować parametry działania algorytmu opisane w poprzednim paragrafie (menu *Wyszukiwarka* opcja *Parametry*).
4. Ostatnim krokiem jest zaznaczenie dodanego hasła i kliknięcie na przycisk oznaczający start algorytmu. W tym momencie następuje pozyskiwanie danych ze stron internetowych i zapisywanie ich do bazy danych w postaci diagramu LHG. Możemy przerwać ten etap wciskając przycisk *Stop* i ponownie go uruchomić przyciskiem *Start*.
5. Po minięciu zdefiniowanego czasu, algorytm przystępuje do budowy definicji. Może to trwać dość długo w zależności od ilości zgromadzonych danych.
6. Po wygenerowaniu definicji w oknie *Log* pojawi się napis mówiący o tym, że definicja została wygenerowana. Możemy wtedy rozwinąć drzewko pod definiowanym hasłem w oknie *Podgląd wyszukiwań* i klikając dwukrotnie otworzyć zbudowaną definicję.
7. Ostatnim opcjonalnym krokiem może być także wygenerowanie wizualnej prezentacji diagramu LHG za pomocą przycisku *Budowa grafu* i otwarcie go za pomocą podwójnego kliknięcia na ikonce grafu, dodanej do pod-drzewka definiowanego hasła.



Rysunek 4.11. Przykładowa definicja dla zapytania systemy wizyjne.



Rysunek 4.12. Graf LHG najpopularniejszych fraz dla hasła: komputer kwantowy.

5. Testy systemu

5.1. Wstęp

Rozdział ten ma na celu przedstawienie efektów działania zbudowanego algorytmu i próbę porównania uzyskanych wyników z innymi dostępnymi rozwiązaniami. Niestety, już na samym początku pojawia się problem, gdyż nie udało mi się znaleźć innych systemów, działających dla języka polskiego, które służyłyby do automatycznego definiowania pojęć.

Istnieją natomiast systemy ekstrahujące specyficzne informacje pewnego typu, najczęściej z dokumentów o określonej budowie. Popularne są także narzędzia do automatycznego generowania abstraktów i skrótów, jednak ich założenia różnią się od założeń tworzonego systemu. Przede wszystkim już na wejściu otrzymują one zdefiniowane wcześniej dokumenty, na których pracują, a zatem operują na pewnym skończonym zbiorze danych. Dzięki temu mogą wyszukiwać najważniejsze fragmenty w zadanym tekście, wyliczając pewne parametry ilościowe, względem całego dokumentu.

Jedyną aplikacją tego typu jest wspomniany w rozdziale 3.3.1 **TextRunner**. Działa on jednak wyłącznie dla języka angielskiego, dlatego możemy jedynie porównać formę prezentacji wyniku pomiędzy obydwiema aplikacjami.

Efekty działania stworzonego algorytmu zostaną zatem w niniejszym rozdziale przedstawione i skonfrontowane z wymaganiami pracy.

5.2. Testy działania aplikacji

Niniejszy rozdział zawiera opis pięciu definicji, wygenerowanych przez zbudowaną aplikację. Zostaną tutaj kolejno przedstawione parametry algorytmu, wykorzystane do budowy każdego kolejnego opisu pojęcia. Dalej przedstawiona zostanie właściwa definicja oraz krótkie wnioski. Z racji, że powstałe objaśnienia są bardzo długie, zostały one odpowiednio sformatowane i przeniesione do niniejszego dokumentu, aby ułatwić analizę. Dodane zostały tutaj także fragmenty grafów LHG, prezentujące jedynie najpopularniejsze frazy, z racji, że całe grafy są bardzo duże i niemożliwe do prezentacji na papierze.

Budowane definicje składają się z kilku bloków. Zawierają one kolejno wyekstrahowane z tekstu przymiotniki, czasowniki oraz rzeczowniki powiązane z aktualnie definiowanym znaczeniem. Zostały sprowadzone do formy podstawowej za pomocą biblioteki CLP, gdy biblioteka ta nie zwracała wielu form bazowych dla podanego słowa. W przeciwnym wypadku pozostawały bez zmian. Dalej za pozyskanymi słowami znajdują się frazy, z których słowa te zostały wyekstrahowane. W nawiasach okrągłych podano liczebność wystąpień fraz w wynikach, zwróconych przez wyszukiwarkę google.com.

Prezentacja wyniku jest podobna do tej z aplikacji **TextRunner**. Znaczącą różnicą jest to, iż nie ma tu wyszukanych ważniejszych grup pojęć, dla których przeprowadzane są kolejno następne ekstrakcje. To kolejne z zagadnień do poruszenia w ramach rozwijania niniejszego systemu. Jest to jednak zagadnienie trudniejsze dla języka polskiego. Przyczyną tego problemu jest konieczność uprzedniego stwierdzenia, czy różne frazy pochodzące od różnych form fleksyjnych należą do badanej grupy czy też nie, a to w wielu przypadkach może sprawiać wiele trudności.

5.2.1. Definicja formy hasłowej: kot

Parametry algorytmu

Liczba przeglądanych stron przez pająka:	40
Czas działania pająka internetowego w godzinach:	36
Liczba najpopularniejszych fraz na grafie:	50
Minimalna popularność frazy:	10
Minimalna ilość wystąpień połączenia:	3
Minimalna długość prezentowanych fraz:	2
Minimalna popularność czasowników:	5
Minimalna popularność rzeczowników:	30
Minimalna popularność przymiotników:	5

Przetworzone dane

Liczba przetworzonych adresów:	29750
Liczba zapisanych wyrazów w bazie danych:	31653
Liczba zapisanych połączeń w bazie danych:	29750
Liczba zapisanych zdań w bazie danych:	21797

Prezentacja definicji

Hasło: kot

Jakie to jest? - zestaw przymiotników opisujących definiowane hasło.

1. inny - dogaduje się z innymi kotami(105), lubi inne koty(71), toleruje inne koty(64), reaguje na inne koty(14), mieszkać z innymi kotami(14),
2. neurotyczny - żyć z neurotycznym kotem(84),
3. dziki - żyją dzikie koty(31), dzikie koty ze wsi(11),
4. duży - koty duże i małe(40), toaleta dla dużych kotów(19), dając swemu kotu duże ilości psiego jedzenia pozbawiasz go wielu niezbędnych dla organizmu składników(10),
5. ważny - jest dla kota ważną(33),
6. norweski - kot norweski leśny(259), koty norweskie leśne koty(141), kota norweskiego leśnego(101), kotach norweskich leśnych(31), kotów norweskich leśnych leśny(23),
7. europejski - kot europejski krótkowłosy(51), koty europejskie sprzedam(12),
8. syjamski - koty syjamskie i orientalne(73), kotów syjamskich i orientalnych(65), kotów syjamskich i perskich(11),
9. piękny - piękne koty o wspaniałym pochodzeniu(14),
10. turecki - kot turecki angora(38), kot turecki van jest(21), koty tureckie angora(11),
11. pospolity - oddam kota koty pospolite(15),
12. wydany - kotów wydanej przez oficynę wydawniczą(32),
13. rasowy - wystawie kotów rasowych(335), hodowlę kotów rasowych(286), międzynarodowa wystawa kotów rasowych(282), koty rasowe koty(120), kotów koty rasowe(56),
14. perski - hodowlą kotów perskich(246), koty perskie i egzotyczne(154), dostał w prezencie perskiego kota(35), mam kota perskiego(27), kot perski reproduktor(22),
15. suchy - kota suchą karmą(27),
16. starszy - starszy kot prawdopodobnie jest już bowiem zadomowiony i przywykł uważać mieszkanie oraz właścicieli za swoją własność dlatego drugiego zwierzaka początkowo będzie uznawać za intruza stąd najlepiej jest wybrać młodego kociaka przeciwnej płci tzn do kocura dobrać małą kotkę lub na odwrót(10),
17. szczęśliwy - sekret szczęśliwych kotów(23),
18. śpiący - nie zbudzić śpiącego kota(15),
19. śmieszny - kot śmieszne filmiki(11), koty śmieszne teksty(11),
20. wszystek - nie wszystkie koty są(50), wszystkie koty lubią(31), wszystkie koty za pan brat(19), wszystkie koty są zwierzętami mięsożernymi(17),
21. niektóry - niektóre koty mogą(105), niektóre koty mają(64), niektóre koty lubią(49), niektóre koty potrafią(27), niektóre koty źle(12),
22. amerykański - kot amerykański krótkowłosy(49), kot amerykański szorstkowłosy(18),
23. wykonany - kota wykonane z materiału(11),
24. lewy - otocz kota lewym ramieniem(63),
25. bezdomny - koty bezdomne smutki i radości(13),
26. wychodzący - jest kotem wychodzącym(43),
27. orientalny - kot orientalny krótkowłosy(47), koty orientalne krótkowłose(11),
28. długowłosy - szampon dla kotów długowłosych(24),
29. czarny - czarny kot przebiegł(251), czarny kot przebiegający(132), kot czarny kot(120), czarny kot jest(112), czarne koty są(97),
30. biały - kot biały kot(295),

31. domowy – największe koty domowe(64), kotów domowych jest(48), zwykły kot domowy(32),
największą rasą kotów domowych(14), duże koty domowe(10),
32. potrzebny – kotu potrzebne jest towarzystwo(15),
33. wielki – wielki kot w małym mieście(153),
34. leśny – norweski kot leśny(133), norweskich kotów leśnych(57), inne koty leśne(21),
35. łagodny – łagodny kot szybko stał się zabawką dzieci(11),
36. śliczny – koty śliczne cudeńka(15),
37. egipski – klątwa egipskich kotów(23),
38. pierwszy – jest to mój pierwszy kot(12),
39. kastrowany – kotów kastrowanych i sterylizowanych(12),
40. specjalny – koty specjalnej troski(17), kota specjalne świąteczne potrawy(12),
41. prawdziwy – prawdziwe koty nigdy nie jedzą z miseczek(28), prawdziwe koty jedzą(18),
42. drugi – wziąć drugiego kota(11),
43. egzotyczny – koty egzotyczne i perskie(38), kotów egzotycznych i perskich(17),

Co można z tym robić? Jaki można mieć na to wpływ?

1. być – są to koty(321), to jest kot(298), to był kot(139),
jest ten kot(137), jest mój kot(125),
2. zależeć – to zależy od kota(97),
3. oddam – oddam kota w dobre ręce(93),
4. mają – nie mają kota(27), mają kota na punkcie kota(16),
5. sprawdź – sprawdź czy kot(12),
6. pozbyć się – pozbyć się kota(102),
7. sprzedam – sprzedam koty koty(16),
8. powinien – powinno się kota(10),
9. reagować – nie reaguje na koty(19),
10. licząc – nie licząc kota(205),
11. mówić – wszyscy mówią do kotów(65), mówią że koty(50), mówi twój kot(18),
12. należą – należą do kotów(39),
13. patrząc – patrząc na kota(27),
14. pochodzić – prawdopodobnie pochodzi od kota(103),
15. lubię – nie lubię kotów(279), bardzo lubię koty(222),
16. chodzić – chodzi o koty(181),
17. utrzymać – utrzymać kota w zdrowiu(13),
18. mieć – chcesz mieć kota(67), może mieć kota(36), chciałbym mieć kota(22),
nie chce mieć kota(15),
19. wystarczyć – wystarczy że kot(16),
20. bierz – nie bierz kota(33),
21. da – da się kota(20),
22. zamieszkać – zamieszka z nami kot(14),
23. miałam – nie miałam kota(76), miałam kiedyś kota(49),
24. bawi – bawi się z kotem(79),
25. podawać – nie należy podawać kotu(13), należy podawać kotom(12),
26. zdarzają – zdarzają się koty(51),
27. zrobić – zrobić z kotem(99),
28. robi – nie robi się kotu(193),
29. dowiedzieć się – dowiedzieć się więcej o kotach(11),
30. lubią – nie lubią kotów(187), bardzo lubią koty(18),
31. zdenerwować – każde poruszenie mogłoby zdenerwować kota(13),
32. kupować – nie kupuj kota w worku kup(21),
33. kupić – chciałbym kupić kota(17),
34. nauczyć – można nauczyć kota(24),
35. należy – należy do kotów(85), należy do kota(34),
36. pojawiać się – pojawiają się koty(18),
37. mieszkać – mieszkać z kotem(30),

38. wychodzić – wychodzić z kotem(13),
39. dbać – dbać o kota w każdym(17), dba o kota(14),
40. posiadać – nie posiada kota(10),

Co to robi? Co się z tym dzieje?

1. obchodzi – kotów nie obchodzi(11),
2. obwąchać – kot obwącha dokładnie drzwi i zdecyduje się czy ma wyjść
czy też może jednak wejść(20),
3. mają – koty te mają(113), kotów które mają(94), koty tak mają(69),
koty mają tendencję(55), koty mają po pięć palców u każdej z przednich łap
ale u tylnych tylko po cztery(18),
4. oddam – koty oddam w dobre ręce(25),
5. dostawać – kot powinien dostawać(39),
6. czują – kota nie czują(73),
7. używać – koty nie używają tak jak inne zwierzęta jakichkolwiek
zbiorników wodnych do mycia się myją(24),
8. tolerować – kot nie toleruje(15),
9. występować – koty te występują w wielu odmianach barwnych(17),
10. towarzyszyć – koty towarzyszą człowiekowi(16),
11. spać – koty lubią spać(27), kot nie śpi(22), kot może spać na łóżku(14),
12. musieć – kot musi być(159), kota musi być(76), kot musiał być(18),
kot nie musi być(12), kota i muszę(11),
13. zasłużyć – przyjaźń tego kota trzeba sobie zasłużyć(14),
14. umieć – koty nie umieją(23), kot nie umie(19),
15. robi – kot też tak robi(27),
16. lubią – koty bardzo lubią(92), to co koty lubią najbardziej(40),
koty lubią perfumy(17), koty nie lubią zapachu(17), koty lubią i muszą pić mleko(12),
17. zrobił – kot nie zrobił(18),
18. wrócił – kot nie wrócił(10),
19. zagra – kot zagra w och(12),
20. pomóc – kot pomoże swojemu nowemu panu zdobyć serce pięknej
królewny której największą życiową pasją jest taniec nie obejdzie się przy tym
bez zasadzek które na dzielną parę zastawiają różne typy spod ciemnej gwiazdy(82),
21. przepadać – kotami nie przepadam(47),
22. nauczyć – koty można nauczyć(46),
23. przypomina – kota bardziej przypomina ludzki niż psi(13),
24. pobawi – widział kota kto pobawi się z myszką(26),
25. wymagać – koty nie wymagają(11),
26. robić – koty zawsze będą robić(95), kota co robić(12),
27. akceptować – kot nie akceptuje(18),
28. robią – koty tak robią(36), koty nie robią(22),
29. zaprasza – kotów syberyjskich zaprasza(21),
30. produkują – koty produkują sardynki tekst(10),
31. wygrać – kota i wygraj(10),
32. wiedzieć – kot wie co dobre(55), kot nie wie(46), kot nie wiem(43),
kota i nie wiem(19), kot to wie(13),
33. porusza – kot porusza się bezszelestnie(52),
34. myją – koty się myją(16),
35. ucieka – kot nie ucieka(18),
36. znosi – kot nie znosi(61), kotów nie znosi(13),
37. widzą – koty widzą świat i ludzi(41),
38. drapie – kot drapie meble(21),
39. chcieć – kota nie chcesz(26), kot chce być(20),
40. jeść – koty nie jedzą(78), kot nie chce jeść(68), kot nie chcę jeść(67),
kot może jeść(52), kot powinien jeść(39),

41. **widziałem** – kota nie widziałem(15),
42. **uwielbiać** – każdy kot uwielbia(17), kot też uwielbia(11),
43. **chodzić** – koty chodzą własnymi drogami(56), kot chodzi własnymi drogami(40), koty chodzą swoimi ścieżkami(18),
44. **rozróżniać** – koty nie rozróżniają(10),
45. **znoszą** – koty nie znoszą(48),
46. **pasuje** – kot nie pasuje(10),
47. **zrobić** – kota co zrobić(11),
48. **stworzyć** – kotów mógł stworzyć tak zabawne(12),
49. **sikać** – kot sika poza kuwetą(15),
50. **wydaje** – kot wydaje się być(11),
51. **pić** – kot nie pije(21),
52. **należy** – kota nie należy(50),
53. **wyglądać** – kot nie wygląda(20), kot będzie wyglądał(20), koty zawsze tak wyglądają(10),
54. **udomowić** – koty i kto je udomowił(20),
55. **bawić** – koty uwielbiają się bawić(22), koty lubią się bawić(16),
56. **być** – koty nie są(299), kot nie będzie(269), kotów nie jest(226), kot jest chory(162), kot jest zwierzęciem(145),
57. **daje** – kot nie daje(55),
58. **powinien** – koty powinny być(111), kota powinna być(49), kota powinno być(37), każdy kot powinien(36), kotom nie powinno(17),
59. **wybiera** – to kot wybiera(13),
60. **zaszkodzić** – kotu nie zaszkodzi(10),
61. **mówić** – kot mówi mama(12),
62. **dostaje** – kot dostaje szału(14),
63. **istnieć** – kot nie istnieje(14),
64. **śmiać się** – kot się śmieje(22),
65. **mieć** – kot powinien mieć(194), kot może mieć(129), koty muszą mieć(91), kota powinien mieć(34), kota może mieć(32),
66. **zostać** – kot może zostać(66), koty mogą zostać(15),
67. **da** – kota nie da(86), kot nie da(51), kot da łapką(17),
68. **myje** – kot sie myje(74), kot się myje(72),
69. **śmierdzieć** – koty nie śmierdzą(25),
70. **złapać** – kot próbuje złapać(12),
71. **nudzi** – kot się nudzi(27),
72. **pływać** – koty nie lubią pływać(47),
73. **drapał** – kot nie drapał(10),
74. **nadaje** – kot jarosława nadaje(22), kot nadaje domowi duszę(11),
75. **proponować** – kot na wionę proponuje zabawę kolorami(11),
76. **grozić** – zabicie kota grozi(12),
77. **traktuje** – przedstawicielami swojego gatunku niektórzy zoolodzy uważają że kot traktuje opiekuna jak zastępczą matkę pozostając w stanie przedłużonego dzieciństwa(25),
78. **trzymaj** – kota i trzymaj(31),
79. **przestawać** – sprawia że kot nigdy nie przestaje być(10),
80. **przynosić** – koty przynoszą nieszczęście(20),
81. **załatwia** – kot załatwia się poza kuwetą(45),
82. **znaleźć** – kotów można znaleźć(20),
83. **powodować** – kotów może powodować(10),
84. **wychodzić** – koty nie wychodzą(32),
85. **chadzać** – koty chadzają własnymi drogami(15),
86. **wykazuje** – kot nie wykazuje(20),
87. **miauczeć** – kot nie miauczy(36),
88. **cierpieć** – kot nie cierpi(14),
89. **reagować** – kot nie reaguje(35),
90. **potrafić** – koty nie potrafią(80), kot nie potrafi(79), koty potrafią wymruczeć chorobę(26), kot to potrafi(10),
91. **szczepić** – koty należy szczepić(20),
92. **widzi** – kot nie widzi(56),
93. **wpadać** – niegrzeczne koty wpadają w kłopoty(22),

94. stać – kot stoi przy tym na ugiętych łapkach podniecenie(21),
95. zobacz – kot zobacz wpis(55),
96. łapią – koty nie łapią(10),
97. lubię – kotów nie lubię(107),
98. otrzymywać – koty mogą otrzymywać następujące oceny(14),
99. bawi – kot się bawi(27),
100. miałam – kota nie miałam(28),
101. móc – kot może być(271), koty mogą być(203), kota może być(150),
kotów może być(125), kot nie mógł(95),
102. zwiększa – kot zwiększa swą prędkość ze stałym przyspieszeniem
aż będzie gotów by się zatrzymać(41),
103. potrzebować – koty nie potrzebują(49), kot nie potrzebuje(33),
104. składa – kot składa się z materii(54),
105. zwariować – kotami i nie zwariować(22), kotem i nie zwariować(12),
106. czuł – kot będzie czuł(34),
107. wymyślić – to koty wymyśliły(15),
108. przyjść – kot nie przyjdzie(33),
109. dbać – koty same dbają o czystość(11),
110. przetrwać – kotom przetrwać zimę(26),

Z czym to jest powiązane? – rzeczowniki połączone z definiowanym hasłem.

1. dom – kota do domu(261), kota z domu(133), kot szuka domu(128),
mam w domu kota(119), kotem w domu(113),
2. miska – miski dla kotów(37),
3. imię – imię dla kota(221), kot ma na imię(137),
4. tygrys – koty domowe tygrysy(27),
5. historia – historia pewnego kota(102), kot historia i legendy(62),
historia kota domowego(21), prawdziwa historia kota w butach zwiastun(18),
krótka historia kota(12),
6. informacja – informacje o kotach(161),
7. koniec – kot w końcu(79), koty w końcu(31), kot na końcu(25),
8. łapa – kot zawsze spada na cztery łapy(246), koty zawsze spadają na cztery łapy(120),
kot spada na cztery łapy(77), wszystkie koty spadają na cztery łapy(36),
ubezpieczenie psów i kotów cztery łapy(34),
9. wiadomości – wiadomości o kotach(28),
10. zoo – koty w zoo(10),
11. pies – pies lub kot(336), kota czy psa(328), psa czy kota(328), psa lub kota(314),
psami i kotami(305),
12. butach – kotem w butach(232), koty w butach(62), prawdziwą historię kota w butach(37),
prawdziwej historii kota w butach(30), kotu w butach(29),
13. humor – humor śmieszne koty(13),
14. korzystanie – kota korzystania z kuwety(29), zachęcić kota do korzystania(10),
15. płot – pierwsze koty za płoty(294),
16. rola – kotem w roli głównej(138), kotami w roli głównej(66),
17. kategoria – kot w kategorii(20), kota w kategorii(13),
18. noc – koty w nocy(80), kot w nocy(53), kota w nocy(40), koty na noc(10),
19. miłość – miłości do kotów(183),
20. zagrożenie – koty wychodzące zagrożenia(10),
21. ręka – kota na ręce(194),
22. rodowód – kota z rodowodem(88), kot z rodowodem(71), koty z rodowodem(49),
kup kota rasowego z rodowodem(10),
23. ludzi – kotów i ludzi(115),
24. sterylizacja – kotów po sterylizacji(84), koty do sterylizacji(15), kotem po sterylizacji(12),
kot po sterylizacji(12), kota po sterylizacji(10),
25. pomoc – pomoc dla kotów(89), kota przy pomocy(25),

26. przykład – kota na przykład(31), kot na przykład(28),
27. ogłoszenia – koty rasowe ogłoszenia(15),
28. poznaniu – kotów rasowych w poznaniu(58), kotów w poznaniu(53),
29. przyjaciel – kot to najlepszy przyjaciel(11), kot najlepszy przyjaciel(10),
30. worek – kupować kota w worku(340), kupowanie kota w worku(284), kot w worku(277),
kupuje kota w worku(268), kupić kota w worku(260),
31. rok – kota co roku(13), kota w tym roku(11),
32. pokój – kota w pokoju(33), czarnego kota w ciemnym pokoju(30), kotów w pokoju(12),
33. miłośnik – wszystkich miłośników kotów(284), klub hodowców i miłośników kotów
norweskich leśnych(54), miłośników kotów i osoby poszkodowane przez los(41),
serwis dla miłośników kotów(16), kotów i ich miłośników(10),
34. filtr – kuweta dla kota z filtrem(22),
35. sprzedaż – koty na sprzedaż(73), kot na sprzedaż(15),
36. szczęście – kot na szczęście(77), czarny kot przynosi szczęście(31), kotów na szczęście(22),
37. ścieżka – kot chodzi własnymi ścieżkami(49), koty chodzą własnymi ścieżkami(43),
koty chadzają własnymi ścieżkami(30),
38. miejsce – kot zawsze wyląduje w najbardziej miękkim miejscu(31), kot przywiązuje
się do miejsca(25), kot do miejsca(23), zapewnia idealne miejsce dla kota w zimne dni(17),
koty przywiązują się do miejsca(17),
39. kotka – kot czy kotka(125), kota i kotkę(63), kot lub kotka(16),
40. śmierć – kota na śmierć(120),
41. kocię – koty i kocięta(296), kotów i kociąt(222), kociąt i kotów(87),
zwierzaki koty kocięta(41), koty kocięta galeria(32),
42. kotki – koty i kotki(259), koty kotki kocięta(155), koty kotki koteczki(152),
kot koty kotki(46), koty kotki i kocięta(42),
43. życie – życie z kotem(82), kota w życiu(40), koty w życiu(15), koty przez całe życie(13),
44. zachowanie – mówi nam zachowanie kota(37), kota można odczytać poprzez jego zachowanie(11),
45. maj – dwa koty maja(18),
46. książka – książki o kotach(93),
47. tekst – kotem tekst lek(12),
48. alergia – alergia na kota(119),
49. cel – kotów w celu(133), kota w celu(36),
50. adopcja – kotów do adopcji(109), psy i koty do adopcji(77), kot do adopcji(57),
dorosłe koty do adopcji(12),
51. pielęgnacja – pielęgnacja kota perskiego(18), kuwety i pielęgnacja kota(17),
52. mleko – mleko dla kotów(139), koty piją mleko(23), mleka dla kotów(19),
53. kuweta – kuwety dla kota(139), kuwety dla kotów(84), kuweta dla kotów(64),
kuweta kryta dla kota yarro z motywem kota(12),
54. witaminy – witaminy dla kotów(49),
55. dzień – kota w dzień(12),
56. dziecko – dzieci i koty(107), kotów i dzieci(28), kot dla dzieci(15), kot dla dziecka(11),
57. okolica – wszystkie koty w okolicy(25), kota w okolicy(17),
58. drzwi – drzwi dla kota(112), kota za drzwi(17),
59. smycz – kota na smyczy(63), kot na smyczy(61), kotem na smyczy(54), koty na smyczy(18),
nauczyć kota chodzić na smyczy(14),
60. zdrowie – książeczka zdrowia kota(33), kota na zdrowie(12),
61. ciekawostka – ciekawostki o kotach(82),
62. przypadek – kota w przypadku(56), przynajmniej w przypadku tego kota(26),
przypadku kotów jest(20), kota w tym przypadku(10),
63. drapak – drapak dla kota(257), drapaki dla kotów(176),
64. orbita – paulina w orbicie kotów(197),
65. jedzenie – jedzenie dla kota(171), kota do jedzenia(35), kotu do jedzenia(30),
66. warszawa – kotów w warszawie(63), wystawa kotów rasowych w warszawie(57), kota w warszawie(28),
67. głowa – kotem na głowie(39), kot śpi na twojej głowie(12),
68. pewność – koty z pewnością(47),
69. robak – kot ma robaki(22),
70. rasa – hodowlą kotów rasy(328), domowa hodowla kotów rasy(183), kota tej rasy(98),
wiele ras kota domowego(64), nabyć kota rasy(47),
71. potrzeba – koty w potrzebie(91), kotów w potrzebie(24), kota w potrzebie(14),potrzeby

energetyczne takich kotów utrzymują się na umiarkowanym poziomie a naturalne mechanizmy(12),
72. kontakt – kontakt z kotem(146), kontakt z innymi kotami(46),
73. podróż – kot w podróży(48), kota do podróży(13),
74. tabletka – podać kotu tabletkę(148), zaaplikować kotu tabletkę(109), tabletki dla kotów(55), kotu tabletkę do pyszczka(31),
75. dżungla – kot z dżungli(22),
76. ogon – odwraca kota ogonem(320), odwracać kota ogonem(315), odwrócić kota ogonem(310), odwracanie kota ogonem(283), odwracasz kota ogonem(259),
77. tapeta – koty tapety na pulpit(54),
78. ogród – kot w ogrodzie(61), koty w ogrodzie(32), psa i kota w ogrodzie(18),
79. zwierzę – koty i inne zwierzęta(289), kotów i innych zwierząt domowych(155), karma dla kotów i innych zwierząt(128), kotów i zwierząt(123), koty zwierzęta w domu(39),
80. kotek – kotów i kotek(26),
81. wystawa – kota do wystawy(34), koty rasowe wystawy(12),
82. akcesorium – akcesoria dla kota(106),
83. schronisko – kota ze schroniska(185), koty ze schroniska(86), kot ze schroniska(43), kot do schroniska(11),
84. możliwość – kot ma możliwość(26), kot musi mieć możliwość(11),
85. pech – czarny kot przynosi pecha(110), koty przynoszą pecha(33),
86. powód – kota z powodu(33),
87. szelki – szelki samochodowe dla kota(42), szelki dla kota odblaskowe(19),
88. champion – koty world champion(10),
89. wąsy – kotu potrzebne są wąsy(18),
90. ryba – koty lubią ryby(27),
91. kaloryfer – legowisko dla kota na kaloryfer(36),
92. zabawka – zabawki dla kotów(190), zabawki dla kota(141), idealna zabawka dla kota(10),
93. porada – porady o kotach(19),
94. droga – kot przebiegnie ci drogę(148), czarny kot przebiegnie drogę(77),
95. hodowla – hodowla kotów rasowych(316), hodowla kotów norweskich leśnych(299), hodowle kotów rasowych(299), hodowli kotów rasowych(263), hodowla kotów perskich i egzotycznych(214),
96. alba – hodowli koty rasowe brytyjskie albo(26), masz kota albo(19),
97. ogół – koty w ogóle(97), kot w ogóle(73), kotów w ogóle(34),
98. strona – strona główna koty(75), jednej strony jest kotem(12), kotach strona główna(11),
99. tama – tam gdzie kot(56), koty się tam(12),
100. oko – oczy kota są(28), kotu w oczy(23),
101. hodowca – hodowców kotów rasowych(184), kota od hodowcy(10),
102. mysz – kot i mysz(352), koty i myszy(128), kota i myszy(113), kot i myszy(88), kotów i myszy(41),
103. weterynarz – kota do weterynarza(87), kotem do weterynarza(66), kota u weterynarza(14),
104. forum – kotów na forum(20), forum poświęcone kotom(16),
105. raz – kot od razu(76), kota w razie(27), koty na razie(21),
106. sztuka – koty w sztuce(43), kot w sztuce(36),
107. żwirek – żwirek dla kotów(253), żwirek silikonowy dla kotów(50), żwirek dla kota z drewna pinii(12),
108. woda – koty nie lubią wody(66), kota do wody(25),
109. właściciel – właściciel kota zajmuje się głównie dostarczaniem mu jedzenia pieszczot i udostępnianiem łóżka pozostałe rozrywki darcie zasłon czy fotela zostawianie kłaków w każdym miejscu wywracanie kuwety leżenie na świeżo upranej bieliźnie polowanie kot zapewnia sobie sam pewne jest również to że od przebywania z kotem simona można dostać kota czego i państwu życzymy(21), właściciel kota jest(17),
110. prawo – prawie wszystkie koty(19),
111. europ – koty w europie(11),
112. ciąża – ciąża u kotów(28), kot a ciąża(27), ciąża u kota(22), ciąża i kot(21),
113. serwis – kotów w serwisie(14), serwisie który nazwaliśmy koty rasowe rasowe koty znajdziecie bierzące informacje ze świata kociego i stowarzyszenia koty(11), koty w serwisie(10),
114. poradnik – kot poradnik opiekuna(32),
115. zdjęcie – zdjęcia kotów rasowych(12), galerie zdjęć kotów(10), kota ze zdjęcia(10),
116. karma – karmy dla kota(233), karmę dla kota(214), karma dla kotów dorosłych(160),

sucha karma dla kota(106), karma sucha dla kota(81),
117. człowiek – kot też człowiek(60), kot i człowiek(52), skrzyżować człowieka z kotem(36), kota na człowieka(30), kotami człowiek żyje(18),
118. piwnica – kota w piwnicy(21),
119. stocznia – ratujmy koty ze stoczni(21), kotów ze stoczni(16), kotów w stoczni(16),
120. świat – koty tłumaczą sobie świat(87), kota na świecie(80), koty na świecie(68), kotów na całym świecie(39), najdłuższy kot na świecie(39),
121. oddania – koty do oddania(59),
122. dwór – kota na dwór(92), kot nie wychodzi na dwór(26),
123. sierść – uczulony na sierść kota(29), sierść kota i dokładnie(13), kot gubi sierść(13),
124. zabawa – zabawa z kotem(167), zabawy z kotem(164), zabawa w koty(54), kotu do zabawy(27),
125. rzecz – rzecz kotów bezdomnych(14), rzecz o kotach(12),
126. mieszkanie – kot w mieszkaniu(60), kota w mieszkaniu(49), począć kot w pustym mieszkaniu(44), koty w mieszkaniu(29), kota do mieszkania(13),
127. prawda – koty co prawda(26),
128. problem – problem z kotem(107), kot ma problemy(55), kot ma problem(19),
129. czas – kot w tym czasie(44), kota na czas(25),
130. większość – większość kotów uwielbia(24), większość kotów nie lubi(18),
131. ludzie – koty i ich sławni ludzie(113), koty i ludzie(109),
132. kociak – koty i kociaki(137), koty kociaki kotki(26),
133. ciało – masy ciała kota(22), budowy ciała kotów(17),
134. wieku – kotów w różnym wieku(51), kotów w podeszłym wieku(28), koty w różnym wieku(21),

Wnioski

Jak można zauważyć, niektóre frazy są niedokończone, ucięte lub niepoprawne gramatycznie jak np: *nie wszystkie koty są*, *niektóre koty mogą*, *niektóre koty mają* itd. Wiąże się to z tym, że połączenie pomiędzy kolejnymi słowami było już za mało popularne, aby algorytm dodał kolejne słowo do budowanej frazy.

Jest to spowodowane zbyt małymi możliwościami obliczeniowymi komputerów. Gdyby były one większe, pozwoliłoby to na uzyskanie większej ilości dokładniejszych danych w krótszym czasie. Ponadto dane w postaci stron internetowych pobierane są z sieci, co powoduje, że czas działania algorytmu jest znacznie dłuższy w porównaniu z aplikacją **TextRuner**, która posiada wbudowaną bazę stron, przekazaną przez firmę Google na rzecz projektu w celach badawczych.

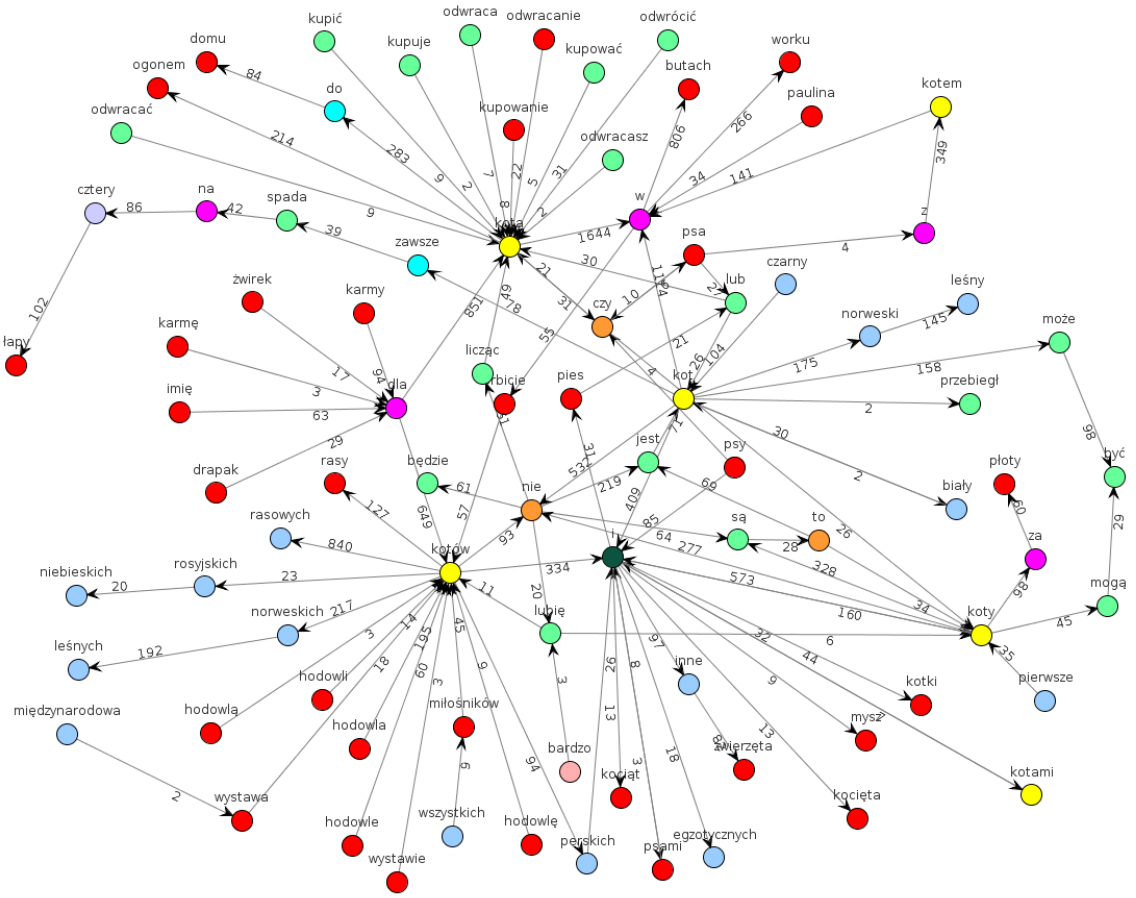
Warto wspomnieć także o tym, że dużo zdań pochodzących z Internetu jest niepoprawna gramatycznie. Wystarczy, że takie zdanie zostanie kilka razy zacytowane, aby algorytm uzyskał wystarczającą ilość powtórzeń, by uznać je za poprawne. W tym wypadku, przy tak zdefiniowanych parametrach, wystarczą jedynie trzy razy. Decyduje o tym parametr: *Minimalna częstotliwość połączenia*.

Analizując zaprezentowaną definicję, można jednak wyszukać bardzo wiele pozytywnych aspektów działania przedstawionego algorytmu. Potrafił on, bez dodatkowej pomocy, stworzyć bardzo obszerny zbiór fraz, które rzeczywiście zawierają jakieś informacje, charakteryzujące definiowane pojęcie. Można zauważyć tutaj, że dzięki

częstotliwościowej analizie wystąpień, zostały one bardzo trafnie wyselekcjonowane i dodane do definicji.

Zaprezentowany w tym przykładzie zbiór słów z pewnością pozwala na uzyskanie prostych informacji, opisujących definiowane pojęcie, a biorąc pod uwagę fakt, że został on wygenerowany automatycznie, można stwierdzić, że podejście wykorzystane w pracy jest skuteczne i daje nadzieje na uzyskiwanie jeszcze lepszych i dokładniejszych wyników w przyszłości.

Graf LHG



Rysunek 5.1. Fragment grafu LHG dla hasła: kot.

5.2.2. Definicja formy hasłowej: drzwi

Parametry algorytmu

Liczba przeglądanych stron przez pająka:	40
Czas działania pająka internetowego w godzinach:	30
Liczba najpopularniejszych fraz na grafie:	50
Minimalna popularność frazy:	10
Minimalna ilość wystąpień połączenia:	3
Minimalna długość prezentowanych fraz:	2
Minimalna popularność czasowników:	5
Minimalna popularność rzeczowników:	30
Minimalna popularność przymiotników:	5

Przetworzone dane

Liczba przetworzonych adresów:	26084
Liczba zapisanych wyrazów w bazie danych:	19831
Liczba zapisanych połączeń w bazie danych:	65436
Liczba zapisanych zdań w bazie danych:	15372

Prezentacja definicji

Hasło: drzwi

Jakie to jest? - zestaw przymiotników opisujących definiowane hasło.

1. przeznaczony - drzwi przeznaczone są do stosowania wewnątrz budynku(17),
2. montowany - drzwi montowane są(42),
3. tarasowy - drzwi tarasowe przesuwne(68), duże drzwi tarasowe(50),
4. płytowy - drewniane drzwi płytowe(20),
5. automatyczny - drzwi automatyczne przesuwne(64), drzwi automatyczne aluminiowe(43),
przesuwne drzwi automatyczne(40), montujemy drzwi automatyczne(14),
drzwi automatyczne przesuwane(11),
6. nowy - czekając na nowe drzwi(123), drzwi nowy sącz(40),
7. garażowy - drzwi garażowe segmentowe(80), automatyczne drzwi garażowe(15),
zewnętrzne drzwi garażowe(14),
8. chiński - drzwi chińskie groemix lakierowane(10),
9. przesuwny - szklane drzwi przesuwne(229), systemem drzwi przesuwnych(106),
drzwi przesuwne szklane(87), są drzwi przesuwne(87), automatyczne drzwi przesuwne(81),
10. drewniany - producent okien i drzwi drewnianych(225), drzwi drewniane drzwi wewnętrzne(177),
drzwi drewniane sosnowe(132), drzwi drewniane drzwi zewnętrzne(126),
drzwi drewniane zewnętrzne i wewnętrzne(108),
11. przeciwpożarowy - drzwi przeciwpożarowe drzwi(158), oferuje drzwi przeciwpożarowe(43),
drzwi przeciwpożarowe profilowe(13), drzwi przeciwpożarowe płaszczowe(13),
12. metalowy - drzwi metalowe zewnętrzne(38),
13. mieszkaniowy - drzwi mieszkaniowe wewnętrzne(12),
14. każdy - nie pukaj do każdych drzwi(238),
15. produkowany - drzwi produkowane są(144),
16. wewnętrzny - drzwi wewnętrzne drzwi zewnętrzne drzwi(138), drzwi wewnętrzne pełne(130),
drzwi wewnętrzne pokojowe(114), drzwi wewnętrzne przesuwne(111),
drewniane drzwi wewnętrzne i zewnętrzne(109),

17. suwany – drzwi suwanych szaf wnękowych i szuflad(10),
18. wykonany – drzwi wykonane są(291), drzwi wykonane z tłoczonych płyt drzwiowych(11),
19. balkonowy – oknach i drzwiach balkonowych(172), montaż okien i drzwi balkonowych(80),
są drzwi balkonowe(62), produkcja okien i drzwi balkonowych(48),
drzwi balkonowe i drzwi zewnętrzne(26),
20. obrotowy – to drzwi obrotowe(15),
21. dwuskrzydłowy – drzwi dwuskrzydłowe z witrażem(17),
22. uchylny – posiada ona drzwi uchylne(15),
23. stalowy – drzwi stalowe przeciwpożarowe(163), drzwi stalowe zewnętrzne(155),
drzwi stalowe wzmocnione(140), drzwi stalowe profilowe(102), drzwi stalowe płaszczowe(78),
24. pełny – drzwi pełne lub przeszklone(19),
25. zewnętrzny – drzwi zewnętrzne stalowe(247), drzwi zewnętrzne aluminiowe(151),
aluminiowe drzwi zewnętrzne(139), drzwi zewnętrzne oraz wewnętrzne(120), to drzwi zewnętrzne(114),
26. zamknięty – odbywa się za zamkniętymi drzwiami(83), zamknięte drzwi są(16),
zamkniętych drzwi na klucz(13), zamkniętych drzwi jest(11),
posiedzeniu przy drzwiach zamkniętych(11),
27. prawy – drzwi prawe i lewe(84), drzwi prawe czy lewe(36),
tylne drzwi prawe(24), przednie drzwi prawe(24),
28. objęty – drzwi objęte są(15),
29. tylny – drzwi tylne lewe(135), drzwi tylne prawe(131), głośników w tylnych drzwiach(40),
30. szklany – drzwi szklane hartowane(61), drzwi szklane wewnętrzne(59),
drzwi szklane wahadłowe(45), drzwi szklane drzwi wewnętrzne(34),
drzwi szklane kabiny prysznicowe lustra(31),
31. zabudowany – decydujemy się na drzwi zabudowane(11),
32. solidny – solidne drzwi pokojowe i bardzo dobre(12),
33. wahadłowy – drzwi wahadłowe można(13),
34. składany – drzwi składane łamane(18), oferuje drzwi składane(12),
drzwi składane i przesuwne(11),
35. boczny – drzwi boczne do garaży(46),
36. wykonywany – drzwi wykonywane są(169),
37. wzmocniony – atestowane drzwi wzmocnione(17),
38. przesuwany – zabudowy wnęk z drzwiami przesuwanymi(135), drzwi przesuwane drzwi(130),
centrum drzwi przesuwanych(123), szafa z przesuwanymi drzwiami(92), drzwi przesuwane składane(41),
39. zielony – drzwi zielona góra(55),
40. rycerski – drzwi rycerskie na straży twojego domu(10),
41. aluminiowy – drzwi aluminiowe zewnętrzne(45), drzwi aluminiowe drzwi stalowe(18),
42. techniczny – drzwi techniczne przeciwpożarowe(62), drzwi techniczne i metalowe(42),
drzwi techniczne akustyczne(16), drzwi techniczne i przeciwpożarowe(12),
43. malowany – drzwi malowane są(75),
44. ewakuacyjny – zamykanie drzwi ewakuacyjnych w sposób(81),
45. atestowany – drzwi atestowane drzwi(14),
46. prowadzące – otworzą się drzwi prowadzące(17),
47. zgodny – drzwi zgodne z prawem(11),
48. lewy – drzwi lewych i prawych(106), drzwi lewe lub prawe(61),
drzwi lewych lub prawych(42), drzwi lewe czy prawe(24),
49. otwarty – zaprasza na drzwi otwarte(109), drzwi otwarte w centrum(34),
50. frezowany – drzwi frezowane to szeroki wybór wzorów i kolorów dający duże
możliwości kreacji wnętrz na specjalne zamówienia realizujemy niestandardowe rozwiązania dla
osób poszukujących własnego indywidualnego stylu oferujemy wysoki standard wykończenia(12),
51. wejściowy – drzwi wejściowe wewnętrzne(235), drzwi wejściowe i wewnętrzne(234),
drzwi wejściowe aluminiowe(162), drzwi wejściowe wzmocnione(149),
drzwi wejściowe drzwi zewnętrzne(141),
52. otwierany – drzwi otwierane do tyłu(37), drzwiach otwieranych na zewnątrz zadaszenie
jest obowiązkowe(20), drzwi otwieranych do wewnątrz jak i na zewnątrz pomieszczenia uniemożliwia
ich demontaż celem zdjęcia skrzydła drzwi(19), drzwi otwierane na zawiasach(17),
dolnej części skrzydła w drzwiach otwieranych(13),
53. znany – oferujemy drzwi znanych(14),
54. sosnowy – drzwi sosnowe wewnętrzne(27), drzwi sosnowe dębowe(12),
drzwi sosnowe producent drzwi(10), są to drzwi sosnowe i drzwi dębowe(10),

55. prysznicowy - drzwi prysznicowe parawany nawannowe baterie pralki(84),
drzwi prysznicowe przesuwne(84), szklane drzwi prysznicowe(53), drzwi prysznicowe obrotowe(47),
drzwi prysznicowe drzwi(35),
56. skrzydłowy - drzwi skrzydłowe z segmentem stałym(14),
57. oklejony - drzwi oklejone płytą w kolorze(10),
58. wyposażony - drzwi wyposażone są(289),
59. wzmacniająca - wklejona rama drewniana wzmacniająca drzwi oraz
podtrzymująca szybę i płycinę(19),

Co można z tym robić? Jaki można mieć na to wpływ?

1. być - to są drzwi(206), jest za tymi drzwiami(110), to nie są drzwi(43),
są to również drzwi(22), dostępne są także drzwi(17),
2. otwórz - otwórz te drzwi(250),
3. móc - być może drzwi(17),
4. domykać - nie domykajmy drzwi(240),
5. sprawdzą - sprawdzą się drzwi(28),
6. kupić - można kupić drzwi(41),
7. oferować - oferujemy również drzwi(76), producent i importer oferuje drzwi(11),
8. otwiera - otwiera sobie drzwi(79),
9. zamontować - prawidłowo zamontować drzwi(13),
10. zamknąć - zamknąć drzwi na klucz(184),
11. znajduje - znajduje się za drzwiami(87),
12. patrząc - patrząc od drzwi(22),
13. wybierając - wybierając drzwi warto(10),
14. puka - puka do drzwi(329),
15. otwieraj - nie otwieraj drzwi(289),

Co to robi? Co się z tym dzieje?

1. być - drzwi są zamknięte(272), drzwi nie jest(260), drzwi to jest(235),
drzwi są bardzo(162), drzwi nie będą(161),
2. podzielić - drzwi możemy podzielić(10),
3. mają - drzwi nie mają(146), drzwi mają być(138), drzwi antywłamaniowe mają(42),
4. wiedzieć - drzwiach nie wiem(52),
5. zajmują - drzwi nie zajmują(18),
6. zamykają - drzwi się nie zamykają(179), drzwi nie zamykają(104),
drzwi zamykają się samoczynnie(34),
7. powinien - drzwi powinny być(287), drzwi nie powinny być(28), drzwi nie powinno być(22),
8. otwieramy - drzwi i otwieramy(16),
9. stanowić - drzwi mogą stanowić(42),
10. stać - drzwi mogą stać(13),
11. puka - drzwi puka wiatr(11),
12. działać - drzwi nie działają(21),
13. musieć - drzwi musi być(118),
14. masz - drzwi nie masz(34),
15. chodzić - drzwi i nie chodzi(11),
16. zostać - drzwi nie zostaną(66), dodatkowo drzwi mogą zostać(13),
17. mieć - drzwi powinny mieć(62), drzwi mogą mieć(50), drzwi muszą mieć(42),
18. podnoszono - drzwi podnoszono przesuwnych(45),
19. da - drzwi nie da(133),
20. otwierać - drzwi powinny otwierać(74), drzwi powinny się otwierać(24),
drzwi mogą się otwierać(22),
21. móc - drzwi może być(231), drzwi nie mogą(144), drzwi mogą być wykonane(110),

drzwi to może(85), drzwi mogą być wyposażone(66),
22. zrobić – drzwi można zrobić(17),
23. wyposażyć – drzwi można wyposażyć(35),
24. trzeba – drzwi to trzeba(30),
25. należy – otworzyć drzwi należy(31),
26. otwiera – drzwi otwiera miłość(67),
27. posiadać – drzwi mogą posiadać(25), drzwi gerda gtt posiadają(14),
28. otwierają – drzwi się otwierają(305),

Z czym to jest powiązane? – rzeczowniki połączone z definiowanym hasłem.

1. szkła – drzwi ze szkła hartowanego(123), drzwi wykonane ze szkła(32),
drzwi szklane ze szkła(20), drzwiach ze szkła(10), drzwi wewnętrzne ze szkła(10),
2. dom – drzwi w domu(279), drzwi wejściowe do domu(263), drzwi wejściowych do domu(131),
drzwi zewnętrzne do domu(94), drzwiach wejściowych do domu(92),
3. wymiana – wymiana drzwi wejściowych(276), drzwi wymiana drzwi(56),
4. klamka – klamki do drzwi wewnętrznych(170), drzwi i klamki(148),
klamki do drzwi zewnętrznych(141), drzwi klamki do drzwi(25), klamki do drzwi wejściowych(17),
5. natura – drzwi natura koncept(32), drzwi classen natura(27),
6. salon – drzwi do salonu(276), salon drzwi drzwi(17), drzwi wrocław salon(14),
7. opel – drzwi opel omega(30),
8. wnęka – drzwi do wnęki(174), wnęki z drzwiami przesuwanymi(129),
wnęki z drzwiami przesuwnymi(86), wnęki z drzwiami suwanymi(33), drzwi do wnęk(33),
9. aluminium – drzwi pcv i aluminium(208), okna i drzwi z aluminium(154),
producent okien i drzwi z pcv i aluminium(114), drzwi zewnętrzne z aluminium(33),
montaż okien i drzwi z pcv i aluminium(23),
10. szkło – drzwi szklane szkło(102),
11. montaż – drzwi z montażem(211), drzwi montaż okien(143), drzwi wraz z montażem(114),
montaż drzwi zewnętrznych i wewnętrznych(113), drzwi montaż drzwi(107),
12. granit – drzwi antywłamaniowe granit(39),
13. rozwiązaniem – idealnym rozwiązaniem są drzwi(12),
14. budynek – drzwi wejściowe do budynku(282), drzwi wejściowych do budynku(279),
drzwi do budynku(267), drzwi w budynkach(115), drzwi wejściowych w budynku(57),
15. szczecina – okna i drzwi szczecin(89), drzwi szklane szczecin(59),
drzwi zewnętrzne szczecin(27),
16. delta – drzwi stalowe delta(82), drzwi antywłamaniowe delta(57),
drzwi metalowe delta(23),
17. klasa – drzwi w klasie(100),
18. sieć – drzwi w sieci(40),
19. wersja – drzwi w wersji(214), wersji z drzwiami(77),
drzwi dostępne są w wersji płaskiej(52),
20. pomoc – drzwi za pomocą(304), drzwi przy pomocy(213), drzwiami za pomocą(22),
21. system – system drzwi przesuwnych(274), systemy drzwi przesuwanych(189),
systemy do drzwi przesuwnych(142), systemy drzwi suwanych(92), okien i drzwi w systemie(83),
22. rozwiązania – rozwiązania dla drzwi(17),
23. ościeżnica – drzwi i ościeżnice(284), drzwi z ościeżnicą(195),
drzwi wraz z ościeżnicą(53), drzwi i ościeżnicy(41), drzwi wewnętrzne i ościeżnice(36),
24. mebel – drzwi i mebli(192), drzwiami przesuwnymi meble(148), drzwi oraz mebli(49),
drzwi meble na zamówienie(39), drzwi przesuwne meble łazienkowe(18),
25. automatyka – automatyka do drzwi(106), bramy automatyka drzwi(32),
26. jakość – drzwi najwyższej jakości(93), dobrej jakości drzwi(54),
wysokiej jakości drzwi drewniane(32), wysokiej jakości drzwi zewnętrznych(26),
najwyższej jakości drzwi drewniane(18),
27. typ – drzwi tego typu(107), drzwi antywłamaniowe typu(56), każdego typu drzwi(35),
drzwi typu porta(29), drzwi różnego typu(14),
28. pokój – drzwiach do pokoju(207), drzwi prowadzące do pokoju(40),

drzwi do pokoju są(27), drzwi wewnętrzne do pokoju(19),
29. Internet – drzwi przez internet(27),
30. sprzedaż – okna drzwi sprzedaż montaż(60), sprzedaż drzwi zewnętrznych(56),
sprzedaż drzwi wewnętrznych(51), drzwi sprzedaż okien(37), drzwi antywłamaniowe sprzedaż(37),
31. miejsce – drzwi w innym miejscu(17),
32. stolarka – okna drzwi stolarka budowlana(62), okna drzwi stolarka okienna(24),
33. płyta – drzwi z płyty(181),
34. przeszklenie – drzwi z przeszkleniem(141), przeszklenia w drzwiach(75),
drzwi antywłamaniowe wejściowe zewnętrzne z przeszkleniem(14),
35. wnętrze – drzwi do wnętrza(183), drzwi do wnętrz(42), drzwi do nowoczesnego wnętrza(14),
36. włókna – drzwi z włókna szklanego(65), drzwi zewnętrzne z włókna szklanego(16),
37. garaż – drzwi do garażu(262), otworzyć drzwi garażu(10),
38. zastosowanie – drzwi te znajdują szerokie zastosowanie(36),
umożliwiając szerokie zastosowanie tych drzwi(22), zastosowanie tych drzwi zapewnia(11),
39. kolorach – drzwi w kolorach(46), drzwi dostępne są w kolorach(10),
40. warunek – drzwi odporne na warunki atmosferyczne(32),
41. kierowca – drzwi od kierowcy(134), drzwiach od kierowcy(91),
zamek w drzwiach kierowcy(78), szyby w drzwiach kierowcy(41), szyb w drzwiach kierowcy(26),
42. góra – drzwi otwierane do góry(162),
43. produkt – drzwi przesuwne produkty(11),
44. głośnik – głośniki w drzwi(118), głośniki w przednich drzwiach(97),
głośniki w tylnych drzwiach(67),
45. ścianka – drzwi i ścianki przeciwpożarowe(72), drzwi i ścianki profilowe(50),
ścianki przeciwpożarowe i drzwi przeciwpożarowe(38), drzwi i ścianki prysznicowe(25),
drzwi i ścianki stalowe profilowe(10),
46. porto – drzwi wewnetrzne porta(172), drzwi zewnętrzne porta(81),
drzwi wejściowe porta(75), drzwi antywłamaniowe porta(47), drzwi porta natura koncept(33),
47. podłoga – podłóg i drzwi(257), podłogi i drzwi(211), okna drzwi podłogi(210),
salon drzwi i podłóg(90), okna drzwi i podłogi(57),
48. pomieszczenia – drzwi do pomieszczenia(288), drzwi od wewnątrz pomieszczenia(11),
49. łazienka – drzwi do łazienki(280), drzwiach do łazienki(233),
drzwiami do łazienki(103), drzwi wejściowych do łazienki(11),
50. przypadek – drzwi w przypadku(276), przypadku drzwi zewnętrznych(71),
jest w przypadku drzwi(13),
51. ściana – drzwi i ściany(251), drzwi na ścianie(90), drzwi w ścianę(16),
montaż drzwi w ścianie nośnej i działowej(14),
52. warszawa – drzwi wewnętrzne warszawa(253), drzwi antywłamaniowe warszawa(206),
drzwi zewnętrzne warszawa(168), drzwi warszawa drzwi(116), montaż drzwi warszawa(115),
53. producent – producent drzwi wewnętrznych(191), drzwi producent drzwi(108),
producent drzwi stalowych(78), drzwi wewnetrzne producent(64), drzwi producent okien(64),
54. produkcja – drzwi naszej produkcji(58), produkcja drzwi wejściowych(38),
produkcja drzwi metalowych(23),
55. roleta – okna drzwi rolety bramy(207), drzwi i rolety(173),
rolety zewnętrzne oraz drzwi(44), okna i drzwi pcv rolety zewnętrzne(11),
56. stadion – drzwi otwarte na stadionie miejskim(36),
57. odporność – drzwi o zwiększonej odporności ogniowej(13),
58. parapet – okna drzwi parapety(279), drzwi i parapety(128), drzwi pcv parapety(120),
drzwi parapety rolety żaluzje(73),
59. zabudowa – zabudowy z drzwiami przesuwnymi(127), zabudowy z drzwiami przesuwanymi(99),
drzwi i zabudowy szklane(36), drzwi szklane i zabudowy(31), drzwi przesuwnych do zabudowy(13),
60. moderna – drzwi dre modern(27),
61. drewno – okien i drzwi z drewna(257), okna i drzwi z drewna(214),
drzwi z litego drewna(207), drzwi z drewna klejonego(102), drzwi wykonane z drewna(92),
62. okno – drzwi okna okucia zamki nie meblowe(324), okien i drzwi zewnętrznych(303),
drzwi oraz okna(280), okna i drzwi są(268), drewniane okna i drzwi(268),
63. okleina – drzwi w okleinie naturalnej(60), drzwi w naturalnej okleinie(39),
drzwi w naturalnych okleinach(22), drzwi wewnętrzne w naturalnych okleinach(16),
drzwi w okleinach naturalnych(15),
64. państwo – państwo drzwi wewnętrzne(45), drzwi oferujemy państwu(12),

65. wymiar – drzwi pod wymiar(52), drzwi przesuwane na wymiar(30),
drzwi przesuwne na wymiar(29), okna i drzwi na wymiar(28),
szafy z drzwiami przesuwnymi na wymiar(26),
66. kuchnia – drzwi do kuchni(309), drzwi od kuchni(193), drzwi w kuchni(107),
drzwiami w kuchni(27),
67. brama – drzwi bramy rolety(217), drzwi bramy ogrodzenia(185),
okna drzwi bramy garażowe(179), drzwi i bramy przeciwpożarowe(160),
drzwi i bram przeciwpożarowych(106),
68. powierzchnia – całej powierzchni drzwi(29),
69. ogród – drzwi ogrody zimowe(267),
70. wzgląd – drzwi ze względu(144),
71. szyba – drzwi z szybą(253), szybą w drzwiach(206), szyba w drzwiach jest(17),
72. zamówienie – drzwi na zamówienie(222), drzwi na indywidualne zamówienie(75),
drzwi drewniane na zamówienie(60), zamówienie na drzwi(23), drzwi na specjalne zamówienie(13),
73. rodzaj – różne rodzaje drzwi(42), drzwi różnego rodzaju(28), drzwi tego rodzaju(17),
drzwi rodzaje drzwi(15), rodzaju drzwi zewnętrzne(15),
74. akcesorium – drzwi i akcesoria(89), akcesoria do drzwi chłodniczych(21),
inne akcesoria do drzwi(20),
75. firma – drzwi zewnętrzne firmy(135), drzwi przesuwnych firmy(106),
drzwi firmy porta(81), drzwi antywłamaniowe firmy(81), drzwi firmy cal(23),
76. profil – drzwi z profili(192), drzwi balkonowe z profili(112),
drzwi wykonane z profili(28), drzwi zewnętrzne z profili(18),
77. kabina – drzwi do kabiny prysznicowej(74), drzwi szklane i kabiny prysznicowe(20),
78. komandor – drzwi przesuwnych komandor(56),
79. zamek – zamków w drzwiach(300), zamki w drzwiach(265), drzwi i zamków(200),
zamka w drzwiach(195), drzwi i zamki(191),
80. garderoba – drzwi do garderoby(164), garderoby z drzwiami przesuwnymi(97),
garderoby z drzwiami suwanymi(21),
81. środek – drzwi do środka(251), drzwi od środka(198), drzwiach otwieranych do środka(24),
82. omega – drzwi dwuskrzydłowe suwane omega(10),
83. skrzydło – skrzydła drzwi posiadają(52), wewnętrzne drzwi i skrzydła drzwiowe(50),
skrzydła drzwi wewnętrznych(48), skrzydło drzwi wykonane(44),
drzwi i skrzydła drzwiowe wewnętrzne(10),
84. wejście – drzwi na wejście(10),
85. strona – drugiej stronie drzwi(323), drzwi po prawej stronie(297),
drzwi z drugiej strony(182), drzwi na stronie(48), strona o drzwiach i producentach drzwi(18),
86. tama – tam są drzwi(282),
87. okucie – okucia do drzwi szklanych(102), drzwi i okucia(32),
okucia do drzwi przeciwpożarowych(15), drzwi oraz okucia(14), okucia drzwi szklane(11),
88. materiał – drzwi i materiały(22),
89. łódź – drzwi wewnętrzne łódź(56), drzwi szklane łódź(26),
90. włamanie – drzwi na włamanie(34),
91. otwieranie – awaryjne otwieranie drzwi i samochodów(62),
92. schody – drzwi i schody(286), schody i drzwi na terenie całego kraju(39),
93. wzór – wzory drzwi wejściowych(14),
94. klasy – drzwi antywłamaniowe klasy(196), wysokiej klasy drzwi(165),
drzwi stalowe antywłamaniowe klasy(33), drzwi najwyższej klasy(10),
95. kolor – drzwi w kolorze białym(43), drzwi w innym kolorze(21), drzwi w białym kolorze(12),
96. blok – drzwi w bloku(57), drzwi wejściowe w bloku(10),
97. budownictwo – drzwi zewnętrzne w budownictwie jednorodzinnym(37),
98. serwis – montaż i serwis drzwi(45), montaż serwis drzwi automatycznych(10),
99. klatka – drzwi wejściowych do klatki(75),
100. wygląd – estetyczny wygląd drzwi(60),
101. piwnica – drzwi do piwnicy(281), drzwi w piwnicy(73), drzwi stalowych w piwnicy(11),
102. konstrukcja – drzwi o konstrukcji(63), drzwi oraz konstrukcji(52),
drzwi i konstrukcje szklane(26), konstrukcja drzwi jest(23), konstrukcja drzwi zewnętrznych(17),
103. poznań – okna i drzwi poznań(32),
104. oferta – drzwi w ofercie(166), oferta drzwi zewnętrzne(82), drzwi w naszej ofercie(69),
ofercie drzwi zewnętrzne(63), ofercie drzwi wejściowe(45),

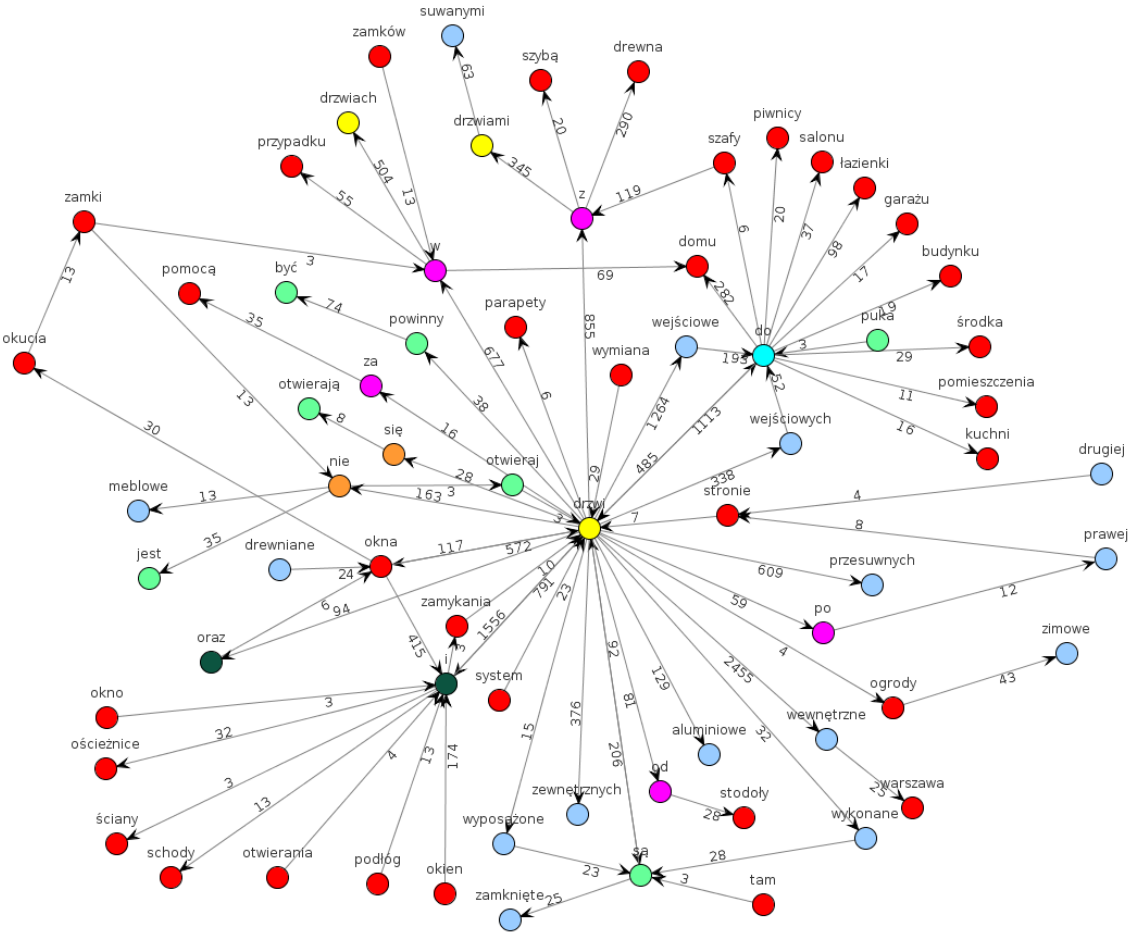
105. fasada - okna drzwi fasady(204), okna i drzwi fasady(37),
106. szafa - szafy z drzwiami suwanymi(267), drzwi do szafy(258), drzwi i szafy(80),
drzwi do szaf wnękowych(66), drzwi przesuwne szafy wnękowe(63),
107. cena - drzwi w dobrej cenie(23),
108. lat - drzwi przez wiele lat(36),
109. mieszkanie - drzwi w mieszkaniu(234), drzwi antywłamaniowe do mieszkania(88),
drzwiach do mieszkania(76), drzwiach wejściowych do mieszkania(65), drzwi wewnątrz mieszkania(61),
110. pomieszczeń - drzwiach do pomieszczeń(54), drzwi do pomieszczeń specjalnych(22),
drzwi garażowych oraz innych pomieszczeń gospodarczych(20),
drzwi do pomieszczeń reprezentacyjnych(11),
111. uwaga - uwagę na drzwi(95), drzwi należy zwrócić uwagę(27),
112. zabezpieczenia - drzwi i zabezpieczenia(32),
113. usługa - drzwi oraz usługi(25), okna i drzwi drewniane usługi(13),
114. styl - drzwi w stylu(131), bezpieczne drzwi wejściowe w stylu retro(18),
115. szyby - szyby do drzwi(140), drzwi i szyby(119), szklane drzwi szyby(25),
116. zabezpieczenie - zabezpieczenie drzwi przesuwnych(23),
117. zamykania - otwierania i zamykania drzwi(252),
118. problem - problem z drzwiami(193),
119. klient - drzwi do klienta(15), drzwi u klienta(10),
120. katalog - drzwi katalog stron(46), drzwi katalog firm(24),
121. stodoła - drzwi od stodoły(267), drzwi do stodoły(76), drzwiach od stodoły poleci(71),
122. zawias - zawiasy do drzwi szklanych(44),
123. wybór - drzwi do wyboru(83), wybór drzwi drzwi(40), drzwi ich wybór(34),
szeroki wybór drzwi zewnętrznych(26), wyborze drzwi wejściowych(26),
124. elementów - drzwi oraz innych elementów(21),
125. promocja - drzwi promocja na montaż drzwi(22),
126. elektryka - zabezpieczenie mebli drzwi i elektryki(19),
127. pomieszczeniach - drzwi w pomieszczeniach(154), drzwi balkonowe w pomieszczeniach(28),
128. otwarciu - drzwi po otwarciu(176), drzwi po ich otwarciu(40),
129. kolekcja - dostępne w drzwiach z kolekcji(10),

Wnioski

Dla kolejnego, dość popularnego zapytania, uzyskano również bardzo obszerny opis. Pająk internetowy pracował krócej niż dla definicji słowa “kot”, co przełożyło się na mniejszą liczbę pozyskanych zdań. Pojawiają się tutaj podobne problemy jak dla hasła opisanego wcześniej. Znów znaczną przeszkodą jest ograniczona moc obliczeniowa komputerów, która wpłynęła na liczbę wyszukanych zdań oraz na czas, jaki trzeba było przeznaczyć na to zadanie. Jednak bez wątpienia końcowy rezultat jest zadowalający, a liczba niepoprawnych fraz stosunkowo mała. Należy pamiętać, że jest to jedno z pierwszych podejść do tego zadania, dlatego wstępny rezultat można uznać za zadowalający.

Wygenerowana definicja zawiera obszerny zbiór przymiotników, czasowników i rzeczowników, które dają dość szczegółowe informacje opisujące definiowane znaczenie. Analizując uzyskany wynik, trzeba stwierdzić, że pomimo znacznych ograniczeń badawczych, uzyskany efekt jest bardzo dobry. Chodzi tutaj nie tylko o liczbę wyszukanych fraz, ale także bardzo poprawne dopasowania do konkretnych kategorii. Bez wątpienia, w tym wypadku system dostarcza obszerny zbiór informacji o badanym haśle, co należy uznać za bardzo dobry rezultat, potwierdzający skuteczność wykorzystanych metod.

Graf LHG



Rysunek 5.2. Fragment grafu LHG dla hasła: drzwi.

5.2.3. Definicja formy hasłowej: komputer

Parametry algorytmu

Liczba przeglądanych stron przez pająka:	40
Czas działania pająka internetowego w godzinach:	24
Liczba najpopularniejszych fraz na grafie:	50
Minimalna popularność frazy:	10
Minimalna ilość wystąpień połączenia:	3
Minimalna długość prezentowanych fraz:	2
Minimalna popularność czasowników:	5
Minimalna popularność rzeczowników:	20
Minimalna popularność przymiotników:	5

Przetworzone dane

Liczba przetworzonych adresów:	27124
Liczba zapisanych wyrazów w bazie danych:	18535
Liczba zapisanych połączeń w bazie danych:	56118
Liczba zapisanych zdań w bazie danych:	15490

Prezentacja definicji

Hasło: komputer

Jakie to jest? - zestaw przymiotników opisujących definiowane hasło.

1. wszystek - wszystkie komputery posiadają(52),
2. współczesny - współczesne komputery są(32),
3. nowy - komputery nowe i używane(125), wraz z nowym komputerem(55),
komputery nowy sącz(32), komputerów nowy sącz(28), komputery nowy dwór(21),
4. domowy - komputer domowej roboty(25),
5. dostępny - komputery dostępne są(42),
6. darmowy - komputera darmowe ruchome(23), komputer darmowe ruchome(23),
7. kompleksowy - komputery kompleksowe wyposażenie firm(14),
8. zielony - komputery zielona góra(85),
9. nurkowy - nowoczesny komputer nurkowy(14),
10. cały - cały komputer jest(62),
11. stacjonarny - komputerów stacjonarnych i przenośnych(281),
komputery stacjonarne i przenośne(159), to komputer stacjonarny(132),
komputer stacjonarny jest(102), komputery stacjonarne komputery(69),
12. przenośny - nowy komputer przenośny(122), większości komputerów przenośnych(89),
dostawa komputerów przenośnych(86), komputera przenośnego jest(83),
komputery przenośne są(80),
13. pokładowy - komputer pokładowy duży(23), mam problem z komputerem pokładowym(17),
aktywacja komputera pokładowego(14),
14. osobisty - pierwszy komputer osobisty(169), pierwsze komputery osobiste(96),
rozwój komputerów osobistych(76), twórca pierwszego komputera osobistego(35),
dostawa komputerów osobistych(18),
15. lokalny - komputerze komputer lokalny błąd(15),
16. pierwszy - pierwszy komputer jest(26),
17. starszy - starszych komputerach jest(11),

18. kwantowy – komputery kwantowe za 20 lat na każdym biurku(12),
to komputer kwantowy(11),
19. przemysłowy – wydajny komputer przemysłowy(12),
20. drugi – jednego na drugi komputer(14),

Co można z tym robić? Jaki można mieć na to wpływ?

1. być – jest komputer i(301), jest na komputerze(289), są to komputery(269),
był to komputer(174), to nie jest komputer(129),
2. włącza – włącza się komputer(100),
3. uruchamiać – uruchamia się komputer(75),
4. korzystać – korzysta z komputera(329), korzystając z komputera można(13),
5. znajdziemy – znajdziemy zarówno komputery(20),
6. podzielić – podzielić na dwa komputery(10),
7. zagrażać – nadal zagraża komputerom(12),
8. poleca – sklep komputerowy poleca komputery(106),
9. zawieszać – zawiesza sie komputer(115),
10. zabezpieczyć – skutecznie zabezpieczyć komputer przed atakami(12),
11. połączyć – połączyć z komputerem(270), połączyć się z komputerem(230),
połączyć dwa komputery(178),
12. siedzieć – siedzi przed komputerem(320),
13. zabezpiecz – zabezpiecz swój komputer(238),
14. przygotować – przygotować komputer do pracy i zainstalować system operacyjny(12),
15. psuje – psuje się komputer(19),
16. uruchomić – uruchom ponownie komputer(211), uruchom komputer ponownie(207),
ponownie uruchom komputer(188),
17. chodzić – chodzi o komputery(280), chodzi o komputer(267),
18. obsługiwać – obsługiwać komputer w stopniu(39),
19. pobierz – pobierz na komputer za darmo(12),
20. złożyć – chce złożyć komputer(43),
21. interesuje – interesuje się komputerami(299),
22. wyłącz – następnie wyłącz komputer(21),
23. znajdują – znajdują się komputery(294),
24. podłączyć – podłączyć dwa komputery(96),
25. składa – składa się komputer(115),
26. znajduje – znajduje się komputer(301),
27. dbać – dbać o komputer(71),

Co to robi? Co się z tym dzieje?

1. być – komputera nie jest(284), komputer jest włączony(275),
komputerze nie jest(266), komputer nie będzie(266), komputer jest wyłączony(254),
2. włącza – komputer się włącza(97), komputer się nie włącza(93),
komputer nie włącza(93),
3. uruchamiać – komputer nie uruchamia(169), komputer się nie uruchamia(127),
4. mają – komputery nie mają(201), komputery też mają wolne w weekend(75),
komputerach nie mają(28),
5. rozpoznaje – komputer nie rozpoznaje(160),
6. mylą – komputery się mylą(14),
7. wiedzieć – komputera nie wiem(203), komputerach i nie wiem(27),
8. pracować – komputer dalej pracuje(16),
9. zastąpić – komputery nigdy nie zastąpią(21),
10. obsługują – komputery nie obsługują(13),
11. psują – komputery się psują(37),

12. powinien – komputer powinien być(276), komputerem powinien być(24),
13. zawieszać – komputer się zawiesza(141), komputer sie zawiesza(118),
komputer się nie zawiesza(52), komputer często się zawiesza(17),
komputer zawiesza się zaraz po włączeniu(11),
14. potrafić – komputer nie potrafi(160),
15. obsługuje – jeden komputer obsługuje(12),
16. widzi – komputer nie widzi napędu(15),
17. widzą – komputery się widzą(81), komputery się nie widzą(58),
18. czytać – komputer nie czyta(139),
19. chcieć – komputer nie chce(150), komputerów i chcesz(15),
20. działać – komputer będzie działał(230),
21. musieć – komputer musi być(271), komputer nie musi być(146),
22. słyszałeś – polskich komputerów o których nie słyszałeś(13),
23. uruchomić – komputer się nie chce uruchomić(11),
24. chodzić – komputerze nie chodzi(32),
25. wykorzystać – komputer można wykorzystać(32),
26. zostać – komputer nie został(128), komputera może zostać(56),
komputera mogą zostać(25), oba komputery zostały(11),
27. mieć – komputer musi mieć(280), komputery muszą mieć(56),
komputer trzeba mieć(30),
28. zacina – komputer się zacina(110), komputer sie zacina(103),
29. wykrywa – komputer nie wykrywa(169),
30. pobierz – komputer pobierz za darmo(66),
31. móc – komputera może być(294), komputery mogą być(172),
komputera a może(155), komputerem może być(128), komputer nie może być(60),
32. trzeba – komputera nie trzeba(178), komputer i trzeba(43),
33. potrzebować – każdy komputer potrzebuje(20),
34. ustawić – komputer można ustawić(14),
35. pomóc – komputer pomoże walczyć z bólem pleców(16),
36. wyłącza – komputer się wyłącza(132), komputer sie wyłącza(125),
37. należy – komputera i należy(17),
38. znaleźć – komputer nie może znaleźć(42),
39. widziały – komputery się widziały(36),
40. zawiesić – komputer się nie zawiesi(24),
41. przegrzewa – komputer się przegrzewa(66),
komputer się nie przegrzewa(24),
42. spełnia – komputer spełnia wymagania gry(91),
43. wymagać – komputerów nie wymaga(24),

Z czym to jest powiązane? – rzeczowniki połączone z definiowanym hasłem.

1. awaria – przypadku awarii komputera(233),
2. pobrania – komputer do pobrania za darmo(60),
komputer za darmo do pobrania(48), komputera do pobrania za darmo(35),
3. dom – komputerem w domu(280), komputery w domu(260),
komputer do domu(217), komputera do domu(76), masz w domu komputer(67),
4. sprzęt – komputery sprzęt komputerowy(175),
komputera i sprzętu(116), komputerów i sprzętu komputerowego(106),
komputery i sprzęt komputerowy(61), komputery oraz sprzęt(55),
5. sklep – komputer ze sklepu(27), komputery w sklepach(16),
6. informacja – informacji w komputerze(199), informacji na
komputerze(144), podstawowe informacje o komputerach(23), informacje o komputerach
najszybciej są umieszczane(15),
7. salon – komputer w salonie(43),
8. urządzenie – urządzenie z komputerem(155), komputer to urządzenie(116),
9. montaż – komputer do montażu(62), samodzielny montaż komputera(43),

10. praca – komputer w pracy(330), praca na komputerze(326),
pracy na komputerze(322), pracy przy komputerze(316), praca z komputerem(299),
11. sieć – komputera podłączonego do sieci(357), komputerów podłączonych do sieci(324),
komputerze w sieci(311), komputera do sieci(290), komputerów w sieć(282),
12. wersja – komputer pełna wersja(74), komputer za darmo pelna wersja(14),
13. technologia – komputery i nowe technologie(42), komputerów i technologii informacyjnej(31),
14. naprawa – rozbudowa i naprawa komputerów(222), naprawa komputerów pogotowie komputerowe(204),
rozbudowa i naprawa komputera(148), naprawa komputerów w warszawie(54),
naprawa komputerów przenośnych(53),
15. instalacja – komputera po instalacji(99),
16. pomoc – komputerze za pomocą(290), komputerem przy pomocy(281),
komputerów za pomocą(270), komputer przy pomocy(188), łączy się z komputerem za pomocą(132),
17. system – komputer z systemem(261), komputerze z systemem(197),
komputera z systemem(183), komputerów z systemem operacyjnym(171), komputery i systemy(164),
18. obudowa – komputer w obudowie(137), obudowy do komputerów(88), k
omputer stacjonarny w obudowie(13),
19. podzespół – komputery i podzespoły(111), komputery podzespoły komputerowe(96),
sprzedaż komputerów i podzespołów(79), podzespoły do komputera(61), komputera i podzespołów(40),
20. połączenie – połączenie dwóch komputerów(214), połączenie do komputera(72),
21. typ – typu komputer komputer(52), komputerów typu media center(44),
komputer osobisty typu(34), komputery stacjonarne typu(16),
22. silnik – komputer silnika ford(20),
23. Internet – komputera podłączonego do internetu(356), komputerem i internetem(354),
komputer podłączony do internetu(343), komputerów podłączonych do internetu(336),
internetu i komputera(319),
24. sprzedaż – sprzedaż komputerów stacjonarnych(133), sprzedaż komputerów przenośnych(71),
komputer laptop części do komputera software na sprzedaż(39),
komputer laptop części do komputera na sprzedaż(10),
25. gracz – komputer dla gracza(187), komputer dla graczy(160),
komputery dla graczy(121), komputerów dla graczy(89), komputera dla graczy(39),
26. miejsce – komputer w miejscu(47),
27. zasób – eksploracja zasobów komputera(17),
28. obsługa – kurs obsługi komputera(315), komputer do obsługi(102),
komputera do obsługi(64), komputery obsługa informatyczna(46),
obsługa komputera z elementami(32),
29. życie – komputer w życiu(136),
30. osoba – grupy tp do obsługi komputera dla osób(14),
31. osa – komputery z google chrome os(26),
32. książka – książki o komputerach(53),
33. cel – komputera w celu(294), komputerze w celu(279), komputerem w celu(278),
komputera osobistego w celu(37),
34. wirusami – komputer przed wirusami(278), komputerów przed wirusami(97),
zabezpieczenie komputera przed wirusami(72), ochrona komputera przed wirusami(63),
35. informatyka – komputery i informatyka(206), informatyka i komputery(170),
informatyka i komputer(10),
36. dzień – komputera na co dzień(64), to mój komputer na dzień dobry(12),
37. dziecko – komputer dla dzieci(164), komputer dla dziecka(157),
komputera przez dzieci(88), komputera dla dzieci(67), komputera dla dziecka(43),
38. grafika – komputer dla grafika(93),
39. kabel – komputery przez kabel(23), dwóch komputerów kablem(20),
komputerów przez kabel(17), komputera wystarczy kabel(11),
40. nauka – komputer do nauki(114), komputera do nauki(92),
komputerów do nauki(30), komputery do nauki(27),
41. urządzenia – komputery i inne urządzenia(215), komputery i urządzenia peryferyjne(110),
komputery urządzenia peryferyjne(78), wszystkie komputery i urządzenia(47),
komputera i urządzenia(44),
42. zdrowie – komputer a zdrowie(148),
43. trybie – komputer w trybie(245), komputerem w trybie(182),
komputera w trybie awaryjnym(113), komputer jest w trybie(60),

komputer uruchamia się w trybie awaryjnym(14),
44. konsola – komputery i konsole(153), komputer i konsole(74),
45. seria – komputery z serii(201), komputerów z serii(127),
komputerów przenośnych z serii(32),
46. stan – komputer w stan(139), komputera w stan(124),
47. obszar – komputer w obszarze(486), komputer multimedia w obszarze(171),
48. przypadek – komputera w przypadku(314), przypadku komputerów przenośnych(115),
przypadku komputerów jest(15),
49. dysk – twardego dysku komputera(203), komputer nie widzi dysku(63),
komputery dyski i napędy(21), jest na dysku komputera(15),
50. temat – zobacz temat komputer(177),
51. grafiki – komputer do grafiki(61),
52. warszawa – serwis komputerów warszawa(146),
53. biurko – biurko pod komputer(240), komputer na biurku(102), komputer w biurku(17),
54. pamięć – pamięci do komputerów(121),
55. potrzeba – komputerów na potrzeby(64), komputera na potrzeby(55),
56. wiedza – wiedzy o komputerze(67),
57. drukarka – komputery i drukarki(203),
58. włączeniu – komputer po włączeniu(132), komputera przy włączeniu(14),
59. grupa – komputery w grupie roboczej(21), komputera w grupie roboczej(14),
60. otoczeniu – komputera w otoczeniu sieciowym(54), komputerów w otoczeniu sieciowym(31),
61. rynek – rynku komputerów osobistych(254), komputerów na rynku(106),
komputery na rynku(46), komputerami na rynku(11), komputerem na rynku(10),
62. student – komputer dla studenta(46),
63. tapeta – darmowe tapety na komputer(128), komputer tapety na pulpit(110),
64. akcesorium – komputery akcesoria w obszarze(499), komputery i akcesoria komputerowe(153),
komputery akcesoria komputerowe(150), akcesoria do komputerów przenośnych(38),
komputery przenośne i akcesoria(35),
65. zastosowań – komputery do zastosowań(50), komputera do zastosowań(23),
zastosowań w komputerach przenośnych(19),
66. firma – komputerów w firmie(286), komputera w firmie(233), komputery dla firm(124),
komputera przenośnego firmy(80), komputery pc dla firm(24),
67. zadań – komputer do zadań(46),
68. domena – komputera do domeny(112), komputer jest podłączony do domeny(13),
69. pulpit – komputer na pulpit(40), tapety na pulpit komputera fajne(31),
70. monitor – komputery i monitory(265), komputer z monitorem(249),
komputer i monitor(236), komputera z monitorem(156), komputer stacjonarny z monitorem(84),
71. użycie – komputerze przy użyciu(216), komputerem przy użyciu(138),
72. elektronika – komputery i elektronika(89),
73. możliwości – komputer i wykorzystaj jego możliwości(18), komputer i jego możliwości(16),
74. sterownik – komputer sterownik silnika(65), komputer sterownik moduł(19),
komputer sterownik ford(15),
75. maca – komputerami pc i mac(206), komputery pc i mac(104),
użytkownicy komputerów mac(90), komputery apple mac(89), komputerów apple mac(78),
77. komunikacja – komunikacja z komputerem(219),
76. strona – komputery strona główna(26),
78. konfiguracja – komputerów w dowolnej konfiguracji(10),
79. oko – komputer za ok(61),
80. program – komputery i programy(225), komputerów i programów(204),
komputer i programy(125), komputera i programów użytkowych(104), komputera przez program(41),
81. materiał – komputery materiały eksploatacyjne(79),
82. porównanie – komputery porównanie cen(25),
83. użytkownik – wielu użytkowników komputerów(203), polskich użytkowników komputerów(84),
użytkownicy komputerów przenośnych(84), użytkownicy komputerów mogą(79),
użytkownicy komputerów są(42),
84. zajęcia – zajęcia z komputerem(305),
85. dźwięk – komputery dźwięk głośniki(13),
86. działania – zasada działania komputera(106),
87. plik – komputery i pliki(26), komputer ze zbędnych plików(25),

88. senior – komputer dla seniorów(193), komputer pc dla seniorów(113), komputera dla seniorów(58),
89. uczeń – komputerów dla uczniów(48),
90. klasy – praktycznie każdy komputer klasy(50), sprzedaż komputerów klasy(43),
komputerem dla klasy(25), współczesnych komputerów klasy(20),
91. forum – komputera forum komputerowe(20), komputer forum dyskusyjne(15),
92. sterowanie – sterowanie komputerem za pomocą telefonu(15),
93. pośrednictwo – komputerem za pośrednictwem(322), komputera za pośrednictwem(300),
komputerów za pośrednictwem(144),
94. urządzeń – komputerów i urządzeń peryferyjnych(300), komputerów i innych urządzeń(293),
komputera i urządzeń peryferyjnych(123), komputera oraz innych urządzeń(75),
komputerów przenośnych oraz urządzeń(11),
95. starcie – komputera po starcie(16),
96. serwis – komputerów serwis komputerowy(236), naprawa i serwis komputerów(189),
profesjonalny serwis komputerów(160), komputery serwis komputerowy(156),
komputery serwis naprawa(111),
97. godzina – komputer o określonej godzinie(29), wyłączenie komputera o określonej godzinie(18),
wyłączanie komputera o określonej godzinie(12),
98. pisania – komputer do pisania(32),
99. zdjęcie – zdjęcia z komputera(205),
100. szkoła – komputer w szkole(321), komputera w szkole(208), komputerów w szkole(202),
komputery w szkole(165), komputery dla szkół(124),
101. stolik – stolik pod komputer(241),
102. karty – komputer nie widzi karty(66), komputery przemysłowe karty(20),
103. człowiek – komputer i człowiek(27), komputerami człowiek żyje(17),
104. świat – komputerów na całym świecie(301), komputer świat twój niezbędnik(158),
komputer świat gry(110), pierwszy na świecie komputer(105), komputer świat ekspert(86),
105. ochrona – ochrona komputera przed zagrożeniami(23),
106. biuro – komputer w biurze(174), komputer do biura(145), komputerem w biurze(44),
komputery do biura(43), dostęp do komputera w biurze(12),
107. poznań – naprawa komputerów poznań(187),
108. zasilacz – zasilacz do komputera(140), komputerów i zasilacze(20),
109. telewizor – komputera do telewizora(224), komputer do telewizora(163),
110. lat – komputery z dawnych lat(10),
111. cena – komputerów w cenie(40), komputery znajdź i porównaj ceny(10),
112. narzędzie – komputer jako narzędzie(310), komputer to tylko narzędzie(53),
komputer stał się narzędziem(39), komputer to nie tylko narzędzie(20),
113. zakres – komputerem w zakresie(125),
114. dostęp – dostęp do komputera z zainstalowanym(58),
115. gra – komputera do gier(237), gry na komputery(208), komputery i gry(158),
komputera w grach(123), komputery do gier(97),
116. granie – komputer do grania(162), komputera podczas grania(49),
komputerów do grania(25), komputer zawiesza sie podczas grania(23),
komputer zawiesza się podczas grania(22),
117. usługa – internet i komputery usługi(22),
118. wzrok – komputera na wzrok(15),
119. problem – problemy z komputerem(265), problemów z komputerami(200),
problemy z komputerami(165), problem z wyłaczającym się komputerem(34),
problemów z komputerem jest(10),
120. telefon – telefon z komputerem(262), komputera na telefon(237),
komputer i telefon komórkowy(156), komputera i na telefon(41), komputera przez telefon(29),
121. katalog – komputerów katalog firm(11), darmowe tapety na pulpit komputera katalog stron(10),
122. klient – naprawa komputerów u klienta(127), naprawa komputera u klienta(24),
123. uruchamianie – komputera podczas uruchamiania(45),
124. czas – komputer w czasie(290), komputera w czasie rzeczywistym(215),
komputerów w czasie(86), komputerach w czasie(11), komputer od pewnego czasu(11),
125. oprogramowanie – komputera i oprogramowania(334), komputera z oprogramowaniem(294),
komputery i oprogramowanie(237), komputer i oprogramowanie(215), komputerów z oprogramowaniem(161),
126. kobieta – komputer jest lepszy od kobiety(25),

Wnioski

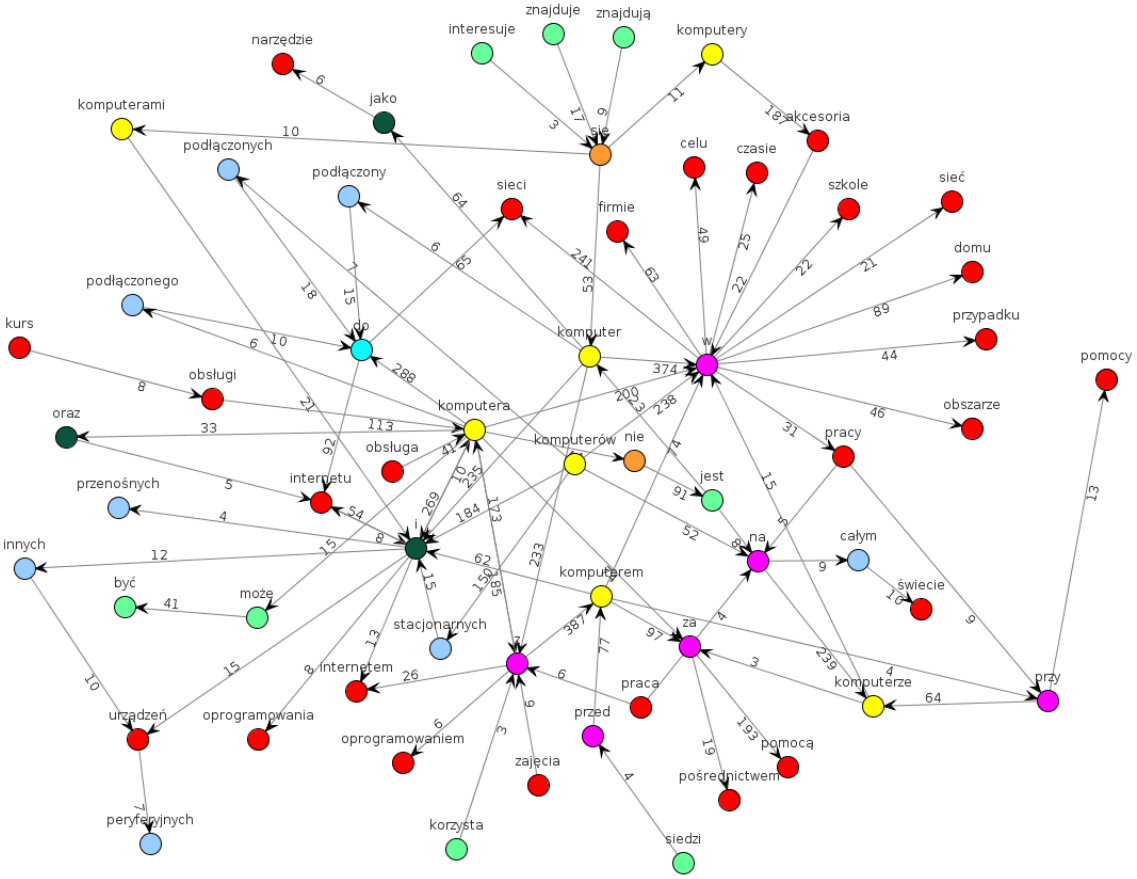
Powyżej zaprezentowana została kolejna definicja dla krótkiego i popularnego hasła. Pająk internetowy działał o kilka godzin krócej, niż dla wcześniejszych definicji, aby sprawdzić uzyskany rezultat, jaki zostanie uzyskany w takim wypadku. Zgodnie z oczekiwaniami, przełożyło się to na liczbę pozyskanych fraz i jest ich mniej. Co więcej, liczba wyszukanych przymiotników jest stosunkowo nieduża, natomiast zwiększyła się liczba wyszukanych czasowników. Rzeczowniki nadal stanowią najliczniejszą kategorię.

Podobnie jak dla wcześniejszych definicji, pojawia się problem niedokończonych ciągów słów. Zdarzają się także informacje błędnie dopasowane - np. dla słowa *zielony* i frazy *komputery zielona góra*. Wyszukane słowo nie definiuje w tym wypadku hasła, a zostało dodane przez algorytm, bo nie potrafił on stwierdzić, że jest to nazwa własna miasta. Wyeliminowanie takiego zachowania wymaga przede wszystkim stwierdzenia, że jest to część nazwy, czego komputer nie jest w stanie samodzielnie wykonać bez dodatkowych informacji.

Czytając wyszukane słowa i odpowiadające im frazy, można bez wątpienia dowiedzieć się podstawowych informacji o definiowanym haśle. Pomimo zmniejszonego czasu działania algorytmu powstała definicja nadal jest bardzo obszerna i zawiera wiele trafnie dopasowanych słów. Można wnioskować, w oparciu o wyniki dla poprzednich haseł, że uruchomienie algorytmu na kilka dodatkowych godzin, z pewnością wpłynęłoby na wygenerowanie bardziej szczegółowego opisu.

Podsumowując uzyskane wyniki, należy podkreślić fakt, że zbudowany algorytm potrafił wyszukać bardzo dużą liczbę fraz opisujących definiowane hasło, uwzględniając przy tym różne formy fleksyjne wyszukanych słów. Dzięki uzyskanym wynikom można odpowiedzieć na podstawowe pytania dotyczące opisywanego hasła, co zostało wykonane w sposób całkowicie automatyczny, bez ingerencji człowieka.

Graf LHG



Rysunek 5.3. Fragment grafu LHG dla hasła: komputer.

5.2.4. Definicja formy hasłowej: komputer kwantowy

Parametry algorytmu

Liczba przeglądanych stron przez pająka:	40
Czas działania pająka internetowego w godzinach:	12
Liczba najpopularniejszych fraz na grafie:	20
Minimalna popularność frazy:	2
Minimalna ilość wystąpień połączenia:	2
Minimalna długość prezentowanych fraz:	2
Minimalna popularność czasowników:	2
Minimalna popularność rzeczowników:	5
Minimalna popularność przymiotników:	2

Przetworzone dane

Liczba przetworzonych adresów:	81870
Liczba zapisanych wyrazów w bazie danych:	1757
Liczba zapisanych połączeń w bazie danych:	2998
Liczba zapisanych zdań w bazie danych:	613

Prezentacja definicji

Hasło: komputer kwantowy

Jakie to jest? - zestaw przymiotników opisujących definiowane hasło.

1. komercyjny - określone jako pierwszy komercyjny komputer kwantowy(2),
2. domowy - domowe komputery kwantowe(4),
3. zwany - zwanych komputerami kwantowymi(2),
4. możliwy - komputerach kwantowych możliwe byłoby wykonanie
tych operacji w bardziej realnym okresie(2),
5. mały - mały komputer kwantowy miałby(2),

Co można z tym robić? Jaki można mieć na to wpływ?

1. zbudować - zbudować komputer kwantowy szybki(2),

Co to robi? Co się z tym dzieje?

1. być - komputer kwantowy będzie(23), komputery kwantowe są(17),
zbudowanie komputera kwantowego jest(4), komputera kwantowego są(3),
komputer kwantowy jest zdolny(2),
2. mają - komputery kwantowe mają(12),
3. móc - komputery kwantowe mogą(30), komputer kwantowy mógłby(12),
komputer kwantowy powstać może w ciągu kilku lat(4),
4. iść - komputerami kwantowymi idą więc dwoma torami(3),
5. rozwiązują - komputery kwantowe rozwiązują zagadnienia niedostępne naszym umysłom(9),
6. wykonać - komputer kwantowy może więc jednocześnie wykonać wiele rachunków równolegle(3),
7. działać - komputery kwantowe działają(2),

Z czym to jest powiązane? - rzeczowniki połączone z definiowanym hasłem.

1. dostęp - dostęp do komputera kwantowego(2),
2. użycie - użyciu komputera kwantowego mógłby szybko rozkładać bardzo
duże liczby na iloczyny liczb pierwszych(2),
3. teoria - teoria komputerów kwantowych(5),
4. praca - prace nad komputerami kwantowymi(4),
5. budowa - budowę komputera kwantowego(5),
6. przyszłość - komputery kwantowe przyszłość informatyki(11),
komputerem kwantowym jest przyszłość(2),
7. algorytm - algorytmy wykonywane przez komputer kwantowy
są algorytmami probabilistycznymi(5),

8. marzenie - komputer kwantowy marzenie a realizacja(6),
komputer kwantowy marzenie szpiegów(3),
9. maszyna - maszyna ta będzie komputerem kwantowym(2),
10. krok - komputer kwantowy krok naprzód komentarzy(3),
11. obliczenie - komputer kwantowy obliczenia(4), pojedynczy wynik
obliczeń komputera kwantowego będzie niepewny(3),
12. rozwiązania - równoległość osiągana w komputerach
kwantowych wystarczy do rozwiązania(2),
13. układ - komputer kwantowy układ fizyczny do opisu którego wymagana
jest mechanika kwantowa zaprojektowany tak aby wynik ewolucji tego układu
reprezentował rozwiązanie określonego problemu obliczeniowego(5),
14. realizacja - sieci optycznej dla realizacji komputerów kwantowych(5),
15. działania - kompletną teorię działania komputera kwantowego stworzył
w połowie lat(4), podstawy działania komputera kwantowego(4),

Wnioski

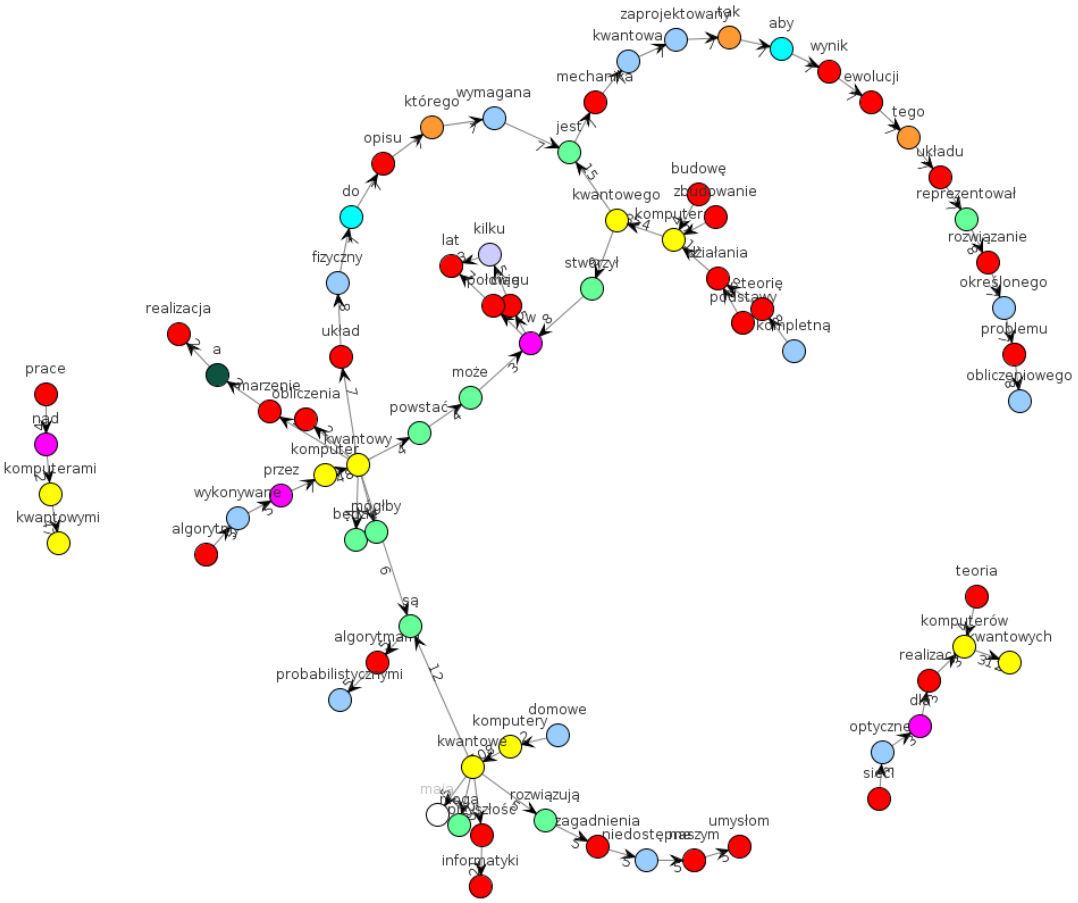
Przedstawiony został tutaj przykład definicji dla kwerendy zbudowanej z dwóch słów. Nieprzypadkowo jest definiowany specyficzny rodzaj “komputera“ czyli ”komputer kwantowy“. Chodziło o sprawdzenie, czy algorytm zbuduje równie długą definicję dla szczególnego podtypu ”komputera“.

Jak można zauważyć, spowodowało to pozyskanie małej liczby zdań pomimo, że algorytm przetworzył znacznie więcej stron internetowych. Przyczyną tego jest konieczność analizowania treści zawierających informacje dotyczące różnych ”komputerów“. Algorytm musiał samodzielnie stwierdzić, które z nich są charakterystyczne dla definiowanej kwerendy. To spowodowało, że zbudowana definicja jest krótsza niż wcześniejsze. System wyszukał mniej przymiotników i czasowników. Zawiera ona jednak, dość charakterystyczne dla definiowanego znaczenia, informacje. Jednak na pewno nie są one wystarczające do pełnego scharakteryzowania opisywanego hasła.

Niestety, ograniczenia komputerów powodują, że konieczne jest sekwencyjne przetwarzanie dużej ilości danych, aby wyszukać informacje charakterystyczne dla pewnych specyficznych pojęć. Jest to zagadnienie czasochłonne z powodu bardzo dużej liczby zadań, jakie musi wykonać komputer.

Pocieszające jest to, że otrzymany rezultat zawiera słowa bardzo charakterystyczne dla definiowanego pojęcia. Co więcej, zostały one bardzo dobrze dopasowane do poszczególnych kategorii. Należy wnioskować, że dostarczenie większej liczby zdań, pozwoli na zbudowanie dokładniejszej definicji.

Graf LHG



Rysunek 5.4. Fragment grafu LHG dla hasła: komputer kwantowy.

5.2.5. Definicja formy hasłowej: systemy wizyjne

Parametry algorytmu

Liczba przeglądanych stron przez pająka:	40
Czas działania pająka internetowego w godzinach:	18
Liczba najpopularniejszych fraz na grafie:	50
Minimalna popularność frazy:	2
Minimalna ilość wystąpień połączenia:	2
Minimalna długość prezentowanych fraz:	2
Minimalna popularność czasowników:	2
Minimalna popularność rzeczowników:	5
Minimalna popularność przymiotników:	2

Przetworzone dane

Liczba przetworzonych adresów:	47479
Liczba zapisanych wyrazów w bazie danych:	3229
Liczba zapisanych połączeń w bazie danych:	6587
Liczba zapisanych zdań w bazie danych:	1009

Prezentacja definicji

Hasło: systemy wizyjne

Jakie to jest? - zestaw przymiotników opisujących definiowane hasło.

1. elastyczny - elastyczne systemy wizyjne(4),
2. oparty - systemów wizyjnych opartych(7), systemy wizyjne oparte(5),
3. zaawansowany - zaawansowane systemy wizyjne(26), zaawansowanych systemów wizyjnych(13),
4. stosowany - systemów wizyjnych stosowanych(6), systemy wizyjne stosowane są(5),
5. kompletny - kompletne systemy wizyjne(10),
6. przemysłowy - przemysłowe systemy wizyjne(31),
przemysłowych systemów wizyjnych i technik identyfikacji(5),

Co można z tym robić? Jaki można mieć na to wpływ?

1. stosuje - stosuje się systemy wizyjne(4),

Co to robi? Co się z tym dzieje?

1. być - systemy wizyjne są(26), systemów wizyjnych są(8), system wizyjny będzie(7),
system wizyjny jest w stanie(2), system wizyjny jest wykorzystywany w wielu dziedzinach
przemysłu motoryzacja opakowania farmacja kosmetyki elektronika tworzywa sztuczne do kontroli
jakości produkowanych elementów jest to tzw wizyjna kontrola jakości szczególnie przydatna jest
w produkcji wielkoseryjnej taśmowej zautomatyzowana jakości pozwala na wytwarzanie produktów
najwyższej klasy będących kluczem do sukcesu nowoczesnej firmy dzięki szybkości działania
i niezawodności pozwalają na maksymalne wykorzystanie możliwości przerobowych nowoczesnych
urządzeń produkcyjnych przyczyniając się do zwiększenia zysków(2),
2. przetwarzać - system wizyjny przetwarza(3),
3. służyć - systemy wizyjne służą(3),
4. móc - system wizyjny może(10),
5. obejmuje - systemów wizyjnych obejmuje(4),
6. kontrolują - systemy wizyjne coraz częściej kontrolują krajowe produkty i procesy produkcyjne(2),
7. dokonuje - program systemu wizyjnego dokonuje(2),
8. wymagać - systemu wizyjnego nie wymaga(2),

Z czym to jest powiązane? - rzeczowniki połączone z definiowanym hasłem.

1. produkcja - systemów wizyjnych w produkcji(3), systemy wizyjne w produkcji(2),
2. pomiar - system wizyjny do pomiarów(8),

3. sterowanie – systemu wizyjnego do sterowania(2), system wizyjny do sterowania(2),
4. komputer – możliwość integracji systemu wizyjnego z posiadanym komputerem(9),
system wizyjny z komputerem(3),
5. wiatr – systemach wizyjnych wiatr(20), systemach wizyjnych kazimierz wiatr(13),
6. kamera – systemy wizyjne kamery przemysłowe(16), kamery i systemy wizyjne(8),
kamery w systemach wizyjnych(8), system wizyjny składa się z kamery(2),
7. oświetlacz – systemach wizyjnych są oświetlacze(2),
8. połączeniu – połączeniu z systemem wizyjnym(2),
9. projektowanie – projektowanie systemów wizyjnych(5),
10. integracja – integracją systemów wizyjnych(9),
11. maszyna – systemów wizyjnych maszyn(3),
12. zastosowanie – zastosowanie systemu wizyjnego(23),
zastosowanie systemów wizyjnych w systemach(2),
13. rynek – rynku systemów wizyjnych(12), rynek systemów wizyjnych wzrośnie(3),
rynek systemów wizyjnych jest(2), polski rynek systemów wizyjnych oraz kamer
termowizyjnych druga edycja(2),
14. wykorzystanie – wykorzystanie systemu wizyjnego(21),
możliwości wykorzystania systemów wizyjnych w układach(2),
15. zastosowania – zastosowania systemów wizyjnych(16),
16. montaż – montaż systemów wizyjnych(10),
17. oferta – systemy wizyjne w ofercie(4),
18. firma – systemy wizyjne firmy(22), systemów wizyjnych visionlab firmy(2),
wykorzystanie systemów wizyjnych firmy cognex zwiększa konkurencyjność i możliwości
produkcyjne nawet najsprawniej działających przedsiębiorstw dzięki wyeliminowaniu pomyłek(2),
systemów wizyjnych renomowanej firmy(2),
19. kierowca – systemy wizyjne kierowcy(8),
20. czujnik – czujniki i systemy wizyjne(32),
21. kontrola – systemy wizyjnej kontroli jakości(17), system wizyjnej kontroli(11),
systemy wizyjne kontroli(3),
22. przemysł – systemy wizyjne w przemyśle(7), systemów wizyjnych w przemyśle(7),
23. zabezpieczenia – systemu wizyjnego zabezpieczenia(6),
24. robot – systemy wizyjne robotów przemysłowych(26), systemach wizyjnych robotów(6),
systemów wizyjnych robotów przemysłowych(5), systemy wizyjne robotów mobilnych(3),
systemu wizyjnego z robotem przemysłowym(3),
25. technologia – oparte na technologiach systemów wizyjnych(5),
26. określenia – systemu wizyjnego do określenia(2),
27. zadaniem – zadaniem systemu wizyjnego jest(3),
28. część – systemy wizyjne część i podstawy(4),
29. nadzór – systemów wizyjnego nadzoru(21), systemy wizyjnego nadzoru i ochrony mienia(18),
30. instalacja – instalacje systemów wizyjnych(2),
31. możliwości – posiadającym możliwości pomiarowe systemem wizyjnym(3),
32. programowanie – projektowanie i programowanie systemów wizyjnych(3),
33. wielofunkcyjność – system wizyjny wielofunkcyjność i pełny zdalny dostęp(7),
34. pomoc – pomocą systemu wizyjnego(13), pomocą systemów wizyjnych(6),
35. telewizja – systemy wizyjne telewizja przemysłowa(7),
systemy wizyjne telewizja użytkowa telewizja przemysłowa tv użytkowa zabezpieczenia techniczne(6),
36. oprogramowanie – oprogramowanie dla systemów wizyjnych(8),
systemy wizyjne i oprogramowanie dedykowane(7),
37. proces – systemy wizyjne w procesach(3),
38. robota – system wizyjny robota(17), system wizyjny dla robota(7),
systemu wizyjnego i robota(2),
39. automatyka – systemy wizyjne w automatyce(14),
automatyka omron systemy wizyjne(5),
40. robotach – systemy wizyjne w robotach przemysłowych(4),
41. miasto – systemu wizyjnego miasta(7), system wizyjny miasta(2),
42. oparciu – systemu wizyjnego w oparciu(3),
43. lokalizacja – lokalizacji i systemom wizyjnym(3),
44. typ – typu systemów wizyjnych(2),
45. analiza – system wizyjny do analizy(10), systemu wizyjnego do analizy(3),

46. rozwój – rozwój systemów wizyjnych w montażu płytek w technologii(2),

Wnioski

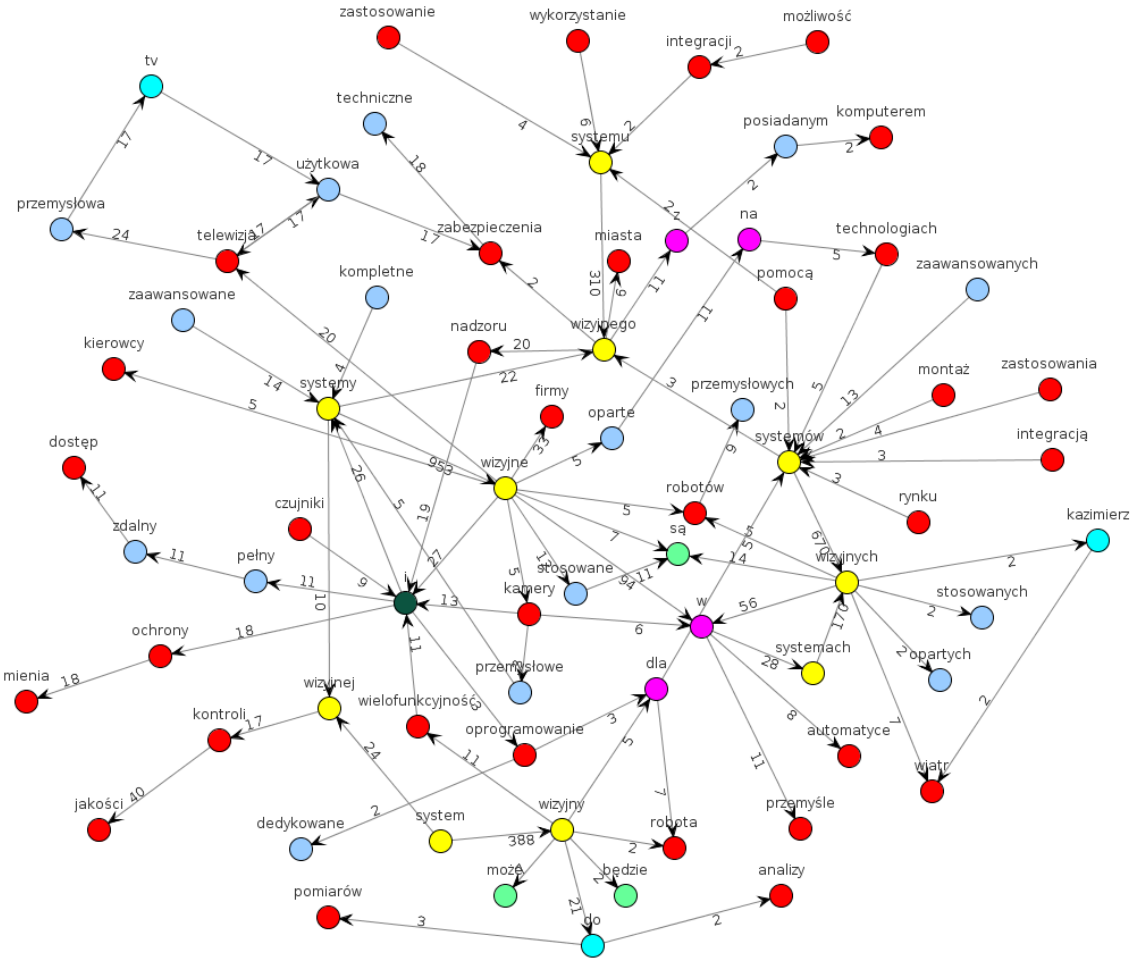
Zaprezentowana wyżej definicja, podobnie jak poprzednia, opisuje pewien charakterystyczny podtyp jakiegoś znaczenia. W tym wypadku chodzi o ”system“. Z racji, że ”systemy wizyjne“ są dość popularną odmianą ”systemu“, to zagadnienie częściej występuje w Internecie. Przekłada się to na powstały opis, który jest dłuższy, bo pająk internetowy dostarczył większą liczbę wyekstrahowanych fraz. Działał on jednak dłużej, co potwierdza przypuszczenie, że ilość pozyskanych fraz jest ściśle zależna od czasu działania algorytmu.

Również i w tym przypadku, definicja zawiera mniej przymiotników i czasowników powiązanych z definiowanym znaczeniem, bo ilość pozyskanych fraz jest mniejsza. Nie przeszkodziło to jednak w stworzeniu dość długiej definicji, zawierającej wiele informacji o opisywanym haśle.

Podsumowując należy stwierdzić, że ilość wyekstrahowanych przez pająka internetowego fraz jest mniejsza dla specyficznych znaczeń. Jest to zgodne z intuicją, bo bardziej szczegółowe znaczenia występują rzadziej na stronach internetowych. Powoduje to konieczność dłuższej pracy algorytmu, w celu wyszukania jak największej liczby zdań, aby stworzyć maksymalnie rozbudowaną definicję. Na rozmiar uzyskanego wyniku, ma także wpływ popularność danego tematu w zasobach internetowych. Dla zagadnień popularniejszych jest on bardziej szczegółowy.

W większości wypadków uzyskany rezultat jest zadowalający, a dla bardziej popularnych fraz otrzymujemy obszerny zbiór słów wraz z dopasowanymi frazami. Ograniczenia, spowodowane zbyt mała mocą obliczeniową komputerów oraz brakiem bazy stron internetowych, zostały tutaj zrekompensowane koniecznością dłuższej pracy algorytmu niż w aplikacji **TextRunner**. Zagadnienie poruszane jednak przez niniejszą pracę jest trudniejsze, bo język polski zawiera bardzo obszerny zbiór form fleksyjnych dla różnych słów, które są uwzględniane przez zbudowany algorytm, przy pomocy biblioteki lingwistycznej CLP.

Graf LHG



Rysunek 5.5. Fragment grafu LHG dla hasła: systemy wizyjne.

6. Podsumowanie

6.1. Wnioski

Główne zadanie, jakie zostało podjęte w ramach niniejszej pracy, to budowa ogólnego algorytmu, który powinien pozyskiwać dane dotyczące pewnego pojęcia ze stron internetowych, ekstrahować z nich szczegółowe informacje i zapisywać je na diagramie LHG. Dalej powstały algorytm, na bazie pozyskanej wiedzy, miał budować sensowną definicję dla zadanego hasła.

W ramach tej oto pracy, przeprowadzono wiele testów i prób, co przyczyniło się do uzyskania ciekawych wniosków i spostrzeżeń.

1. Zaprezentowany algorytm jest dopiero pierwszą próbą rozwiązania zagadnienia automatycznego generowania definicji pewnych pojęć, podjętą pod nadzorem promotora pracy dra Adriana Horzyka. Pomimo, że problem jest bardzo ciekawy, to jest on kłopotliwy, z racji na swoją specyfikę. Zmiany wprowadzane w algorytmie wymagają często wielu testów dla różnych parametrów, a to powoduje, że są one bardzo czasochłonne. Potrzeba jeszcze dużo czasu, aby dokładnie przeanalizować i wypróbować w praktyce nowe podejścia i pomysły.

2. Wykorzystane w tej pracy podejście bazuje na danych ilościowych, zgromadzonych za pomocą specjalnie zbudowanego pająka internetowego, który wykorzystywał wyszukiwarkę google.com do pozyskiwania stron internetowych. Następnie wyszukiwane były ciągi słów, zawierające szukane hasło i za ich pomocą tworzono grafy LHG, jednocześnie prowadząc statystyki przetworzonych słów oraz połączeń. Kolejno na bazie zgromadzonych danych wyszukiwane były listy słów, a następnie ponownie za pomocą wyszukiwarki, sprawdzana była popularność pojawiania się danego ciągu w Internecie. Dalej, z tak potwierdzonych fraz, wybierane były czasowniki, rzeczowniki i przymiotniki. Jak się okazało, zastosowane podejście dało bardzo ciekawe wyniki.

3. Problematyczne okazało się być wyszukiwanie pełnych zdań, które opisują dane znaczenie. Sam problem ekstrakcji pełnych zdań z Internetu, dostarcza sporo

kłopotów. Nie możemy analizować częstości pojawiania się całych zdań, ponieważ język polski jest tak bogaty i dostarcza tak dużo możliwości konstruowania zdań, że spotkanie zdania o takiej samej budowie w rożnych źródłach jest bardzo mało prawdopodobne. Co więcej to, że pewna forma definiowanego hasła występuje gdzieś w środku zdania, nie wystarcza do stwierdzenia, że całe zdanie ją opisuje. To spowodowało, że w ramach niniejszej pracy, wyszukiwane były jedynie frazy, a nie całe zdania, zachowując przy tym pewność, że słowa nie zostały przypadkowo połączone usuwając znaczniki HTML.

4. Zaprezentowane definicje pokazały, że dane statystyczne pozwalają na wyszukanie najważniejszych informacji związanych z danym hasłem. Na ich podstawie możemy powiedzieć, jakie może być to, co definiujemy. Można także, w większości wypadków, wyróżnić orzeczenia charakterystyczne dla danego hasła oraz rzeczowniki bezpośrednio z nim związane. Pozostaje nadal jednak problem wieloznaczności językowych, który powoduje, że takie słowa mogą być niepoprawnie dopasowane.

5. Ilość fraz pobranych przez robota, zależy od popularności danego hasła w Internecie. Jeżeli robot wyszuka więcej fraz w Internecie, to powstała definicja jest ostatecznie bardziej rozbudowana.

6. Zaproponowany algorytm nie rozwiązuje wszystkich problemów. Pozwala on na zgromadzenie pewnych informacji na dany temat. Jednak czasami znalezione frazy mogą być urwane, błędne lub mogą zawierać niepełne informacje. Spowodowane jest to zastosowanym podejściem, które zakłada, że połączenia mniej popularne są obcinane i uznawane za błędne. Związane jest to z brakiem mechanizmu, który potrafiłby stwierdzić, czy usuwane informacje są ważne.

7. Należy tutaj podkreślić fakt, że badania nad tym zagadnieniem dla języka polskiego są bardzo potrzebne. Jeżeli sami nie będziemy rozwijać tej dziedziny, to zostaniemy w tyle za innymi krajami. Wiąże się to przede wszystkim z charakterystyką języka polskiego, która powoduje, że nie można dla niego stosować metod i formalizmów zbudowanych dla innych języków.

6.2. Zrealizowane cele

W ramach pracy udało się uzyskać bardzo dobre rezultaty, które z pewnością należy uznać za sukces, zwłaszcza, że jest to jedno z pierwszych podejść do tego problemu dla języka polskiego:

- W ramach pracy udało się zbudować specjalistycznego, wielowątkowego pająka internetowego, wykorzystującego wyszukiwarkę google.com, zdolnego ekstrahować frazy, zawierające definiowane pojęcie. Wykorzystano do tego celu algorytm opisany w rozdziale 4.

- Powstał mechanizm potrafiący, za pomocą danych statystycznych oraz asocjacji słownych, zgromadzonych na diagramie LHG, ekstrahować frazy i filtrować je, wykorzystując częstotliwość wystąpienia hasła w rezultatach zwracanych przez wyszukiwarkę google.com.

- Zbudowany został mechanizm potrafiący, z pobranych fraz, wyszukać najważniejsze przymiotniki, rzeczowniki i czasowniki, podzielić je na odpowiednie kategorie i zaprezentować je w przystępnej formie.

- Powstały algorytm działa w sposób całkowicie automatyczny i jest w pełni ogólny. Z taką samą skutecznością wyszukuje on informacje dla różnych pojęć, a rozmiar wyniku uzależniony jest wyłącznie od popularności definiowanej kwerendy w Internecie i ilości danych na ten temat na stronach internetowych.

- Wygenerowane przez algorytm wyniki są bardzo obszerne. Większość wyszukanych słów została bezbłędnie dopasowana do konkretnej kategorii i są one poparte wyszukanymi w tekście frazami. Dzieje się to pomimo znacznych ograniczeń badawczych, które są efektem zbyt jeszcze małych możliwości obliczeniowych komputerów.

6.3. Możliwości rozbudowy

Z racji, że zagadnienie podjęte w tej pracy jest jeszcze bardzo słabo rozwinięte, jest tutaj bardzo dużo możliwości rozbudowy:

- W pierwszej kolejności, warto byłoby rozwiązać problem efektywnego wyszukiwania zdań w Internecie, które zawierają informacje na temat definiowanego pojęcia. Jest to spory problem i ograniczenie, a poprawa tego modułu z pewnością poprawiłaby efektywność całego systemu.

- Na pewno, ciekawa byłaby próba połączenia metod statystycznych z metodami bazującymi na gramatyce języka. Rozbiór zdania pozwalający na wyodrębnianie najważniejszych jego części, jak podmiot i orzeczenie, z pewnością pomogłyby w

ekstrakcji najważniejszych informacji ze zdania i stwierdzenie, które z nich są w danym kontekście najważniejsze.

- Warto byłoby także, poprawić sposób prezentacji wyszukanych informacji. Wiąże się to jednak, z kolejnymi przekształceniami wyszukanych fraz i próbą dopasowania ich do pewnych, bardziej szczegółowych grup. W większości przypadków, wiązałoby się to z potrzebą rozwiązywania problemu niejednoznaczności językowych występujących w tekście i koniecznością ich rozróżnienia.

- Co więcej, warto byłoby wzbogacić system o możliwość podglądania źródła, z którego pochodzi wyszukane zdanie, co pozwoliłoby na prześledzenie szerszego kontekstu.

Bibliografia

[1] Wiesław Lubaszewski. *Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu*. AGH, 2009.

[2] Marcin Gadamer. Automatyczna kontekstowa korekta tekstów na podstawie grafu przyzwyczajeń lingwistycznych (lhg) zbudowanego przez robota internetowego dla języka polskiego. Praca magisterska, AGH, 2009.

[3] Bartosz Zarzyczny. Budowa słownika frekwencyjnego na bazie polskojęzycznych stron internetowych. Praca magisterska, AGH, 2005.

[4] Raymond J. Mooney, Un Yong Nahm. Text mining with information extraction, 2005.

[5] Agnieszka Mykowiecka. Narzędzia inżynierii lingwistycznej w analizie dialogów, 2009.

[6] Sergey Brin. Extracting patterns and relations from the world wide web, 2009.

[7] Marek Gajęcki. Słownik fleksyjny języka polskiego clp - opis użytkowy, 2008.

[8] Jan Czajkowski. Ekstrakcja informacji. Seminarium z przetwarzania języka naturalnego w Instytucie Informatyki Uniwersytetu Wrocławskiego.

[9] Marcin Junczys-Dowmunt. Zastosowanie automatów skończonych stanów w przetwarzaniu języków naturalnych.

[10] Michał Marcinczuk. Ekstrakcja informacji o zdarzeniach z tekstów dziedzinowych, 2008.

[11] Aleksander Buczyński. Pozyskiwanie z internetu tekstów do badań lingwistycznych. Praca magisterska, Uniwersytet Warszawski, 2004.

[12] Przemysław Sołdacki. Wprowadzenie do eksploracji tekstu i technik płytkiej analizy tekstu. Wykład z przedmiotu Inteligentne Systemy Informacyjne.

[13] Sychniak Marcin, Wiak Sławomir. Lingwistyczny system informatyczny oparty o mechanizmy przetwarzania języka naturalnego, 2007.

[14] Włodzisław Duch. Analiza języka naturalnego.

[15] The linguistic engineering group(http://nlp.ipipan.waw.pl/).

[16] Grupa lingwistyki komputerowej(http://winnie.ics.agh.edu.pl/).

[17] Grzegorz Jagodziński. Gramatyka języka polskiego(http://grzegorj.w.interia.pl/gram/gram00.html).

[18] Textrunner search (experimental) (http://www.cs.washington.edu/research/textrunner/indextrtypes.html).

[19] Adrian Horzyk. Wykłady podstaw informatyki(http://home.agh.edu.pl/ horzyk/lectures/pi-/ahdydpiwykl0.html).

[20] Janusz Majewski, Marcin Kuta, Marta Majewska. Teoria automatów i języków formalnych, teoria kompilacji(http://kompilatory.agh.edu.pl/).

[21] Swt widgets(http://www.eclipse.org/swt/widgets/).

[22] Latex(http://en.wikibooks.org/wiki/latex).

[23] Apache log4j(http://logging.apache.org/ log4j/).

[24] Java universal network/graph framework(http://jung. sourceforge.net/).

[25] Jericho html parser(http://jericho.htmlparser.net/docs/index.html).

[26] Korpus języka polskiego wydawnictwa naukowego pwn(http://korpus.pwn.pl/).

[27] Agata Filipowska. Automatyczne tworzenie abstraktów z dokumentów. *Gazeta IT*, 2003.

[28] Douglas E. Appelt, David Israel. Introduction to information extraction technology(http://www.ai.sri.com/ appelt/ie-tutorial/).