Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie

Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki

PRACA MAGISTERSKA

PAWEŁ MICZKO

SYSTEM AUTOMATYCZNEGO ROZPOZNAWANIA WYBRANYCH OSÓB NA PODSTAWIE ANALIZY ICH AKTYWNEGO SŁOWNIKA SŁÓW I ZWROTÓW.

PROMOTOR:
dr Adrian Horzyk

Kraków 2009

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.

. . . . . .

PODPIS

AGH
University of Science and Technology in Krakow

Faculty of Electrical Engineering, Automatics, Computer Science and Electronics

MASTER OF SCIENCE THESIS

PAWEŁ MICZKO

SYSTEM OF AUTOMATIC USERS DETECTION BASED ON ANALYSIS OF THEIR ACTIVE DICTIONARY OF WORDS AND PHRASES.

SUPERVISOR:
Adrian Horzyk Ph.D

Krakow 2009

Bardzo serdecznie dziękuję promotorowi dr Adrianowi Horzykowi za cenne wskazówki, wyrozumiałość oraz wsparcie okazane przy pisaniu niniejszej pracy

Spis treści

1. Wstęp

W dobie społeczeństwa informacyjnego istotna stała się możliwość analizy pisanych przez użytkowników Internetu tekstów. Z biegiem lat, coraz większy odsetek społeczeństwa uzyskuje stały dostęp do Internetu, dla wielu ludzi staje się on głównym źródłem informacji i rozrywki. Tradycyjne biblioteki i księgarnie pustoszeją, a papierowe egzemplarze książek zastępowane są przez wiele osób ich wersjami elektronicznymi. Następuje coraz większy przyrost ilości stron internetowych, na których każdy może umieścić napisane przez siebie treści. Zgodnie z tzw. Web 2.0, główną rolę w zawartości serwisów internetowych powstałych po 2001 odgrywa treść generowana przez użytkowników danego serwisu [1]. Powszechne stały się portale społecznościowe, blogi, fora internetowe, serwisy edukacyjne, na których każdy może się wypowiedzieć bądź zamieścić stworzony przez siebie tekst. Niestety, ze względu na powszechność dostępu i względną anonimowość użytkowników Internetu, częste staje się kopiowanie, plagiatowanie i przypisywanie sobie autorstwa cudzych dzieł. Coraz większego znaczenia nabiera zatem możliwość rozpoznawania autorstwa tekstów pisanych.

Analizując częstotliwość używania przez daną osobę słów i zwrotów można stwierdzić, jakie są jej ulubione sformułowania, słowa unikalne dla tej osoby. Ze stopnia złożoności i długości budowanych zdań, możemy nawet wnioskować o jej inteligencji. Co za tym idzie, uzyskujemy cechy charakterystyczne dla jej sposobu mówienia/pisania, co może być potem wykorzystane jako czynnik wyróżniający tę osobę na tle innych, w celu ich rozpoznawania.

Analiza tekstu daje także możliwość tworzenia portretów psychologicznych użytkowników, na podstawie pewnych słów i zwrotów charakterystycznych dla określonych grup ludzi [2] [3] [4] [5]. Możliwe staje się zaklasyfikowanie tej osoby do jednego lub kilku programów osobowości o różnym stopniu intensywności. Stworzenie portretu psychologicznego osoby może być również bardzo pomocne przy rozpoznawaniu autorstwa tekstów.

Praca będzie unikalnym rozwiązaniem, łączącym standardowe metody frekwencyjne oraz próbę automatycznego stworzenia profilu psychologicznego danej osoby, w celu klasyfikowania i rozpoznawania autorstwa dzieł literackich i innych tekstów pisanych.

1.1. Główny cel

Głównym celem pracy magisterskiej było stworzenie internetowego systemu umożliwiającego rozpoznawanie wybranych osób na podstawie ich aktywnego słownika słów i zwrotów. Zakres pracy obejmuje skonstruowanie słowników frekwencyjnych słów i zwrotów dla wybranych mówców lub pisarzy oraz ich wykorzystanie do rozpoznania autorstwa innych tekstów wypowiadanych przez tych mówców lub pisarzy. Celem pracy jest skonstruowanie systemu pozwalającego na zidentyfikowanie dzieł literackich jak również dowolnych innych tekstów napisanych przez jedną z wybranej grupy osób. Jako metoda pomocnicza w rozpoznawaniu autorów wybrana została klasyfikacja typów osobowości według typologii stworzonej przez doktora Adriana Horzyka [2] [3] [4] (późniejsze uzupełnienie o typ Oszczędny [5]).

1.2. Postawienie problemu

Głównym problemem jest skonstruowanie algorytmu, dzięki któremu będzie można z jak najmniejszym prawdopodobieństwem błędu rozpoznać użytkownika spośród zgromadzonych w bazie tekstów. Należy więc stworzyć bogatą bazę danych, która będzie zawierała możliwie długie teksty różnych autorów, tak aby można było z dobrą dokładnością zebrać informacje na temat najczęściej używanych przezeń słów i zwrotów, a także spróbować stworzyć profil osobowości danego autora.

1.3. Zawartość pracy

W rozdziale drugim przedstawiono wprowadzenie teoretyczne do problemu oraz przegląd istniejących rozwiązań. Rozdział trzeci stanowi opis proponowanego rozwiązania oraz przebieg jego implementacji. Rozdział czwarty przedstawia efekt tej implementacji, przewodnik po aplikacji oraz wyniki testów systemu. Rodział piąty stanowi podsumowanie pracy.

2. Wprowadzenie teoretyczne do tematu

2.1. Użycie słów i zwrotów miarą ludzkiej inteligencji, kultury i osobowości

Analiza częstotliwościowa aktywnego słownika danej osoby może służyć uzyskaniu "podpisu" konkretnego autora, obrazu poziomu jego kultury osobistej, używania żargonu technicznego i innych środków językowych. Możliwa jest ekstrapolacja liczby słów użytych w pewnym tekście w kierunku całości słownictwa danej osoby. Analiza częstotliwościowa może wykazać, że niektórzy pisarze mają słownictwo dziesięciolatka lub zasób słów osoby uczącej się języka dopiero drugi rok.

Jest ona również sposobem na uniknięcie napisania identycznego tekstu dla tych, którzy muszą stworzyć tekst różny od źródła, np. autora stron internetowych, który musi wypełnić treścią wiele podobnych, lecz nie takich samych szablonów albo studenta chcącego uniknąć - słusznego lub też nie - oskarżenia o plagiat.

Detekcja plagiatów to kolejne z zastosowań analizy częstotliwościowej tekstów. Porównanie danego tekstu z całą zawartością Internetu to bardzo trudne zadanie, gdyż automatyczny system detekcji nie rozumie znaczenia tekstów. Analiza częstotliwości słów może dać wskazówkę na temat stylu piszącego, bez konieczności indeksowania całości.

Również wyszukiwarki internetowe korzystają z analiz częstotliwościowych, aby określić temat strony internetowej. Klasyfikują one witryny bez udziału człowieka. Np. użycie słowa z częstotliwością 3% jest dobrym wyznacznikiem tego, że o tym traktuje dany artykuł. Częstotliwość np. 10% podchodzi już pod "keyword stuffing", czyli technikę używaną przez webdeveloperów, polegającą na nadużywaniu słów kluczowych po to, aby strona znalazła się wyżej w wynikach wyszukiwania. Sztuczka ta jest jednak rozpoznawana i karana przez roboty indeksujące, musi zatem być obchodzona poprzez sprytne używanie synonimów, co wymaga albo dobrego słownika synonimów, albo dobrego redaktora. [6].

Analiza częstotliwościowa ogółu stworzonych przez ludzkość tekstów daje też dobry obraz historii człowieka, powiązań kulturalnych i społecznych. W projekcie Wordcount zostały wykorzystane różne źródła pisane i mówione, liczące łącznie ponad 100 milionów słów, w celu stworzenia rankingu 86800 najczęściej używanych słów w języku angielskim. Wyniki tych badań są bardzo ciekawe i mówią wiele o historii i kulturze człowieka, na przykład słowo "Bóg" dzieli w tym rankingu jedna pozycja od słowa "zaczynać się", dwie pozycje od słowa "początek" i sześć pozycji od "wojna". [7]

2.2. Używane słownictwo a osobowość człowieka

Każdy człowiek posiada pewien charakterystyczny, względnie stały sposób reagowania na otaczający go świat i wchodzenia z nim w interakcje, zwany osobowością [8]. Osobowość każdej żyjącej osoby jest unikalna i niepowtarzalna, różnie też będą się ujawniać poszczególne jej cechy w zależności od sytuacji. Możliwa jest jednak uogólniona kategoryzacja podobnych osobowościowo grup osób. Na ten temat powstało wiele teorii, które dzieliły osobowości według różnych kryteriów. Dość popularna jest na przykład teoria Hipokratesa [9], który podzielił ludzi na cztery grupy:

1. flegmatycy - ludzie spokojni, zrównoważeni, nieulegający emocjom.
2. cholerycy - ludzie wybuchowi, emocjonalni, działający w sposób gwałtowny, często nieprzemyślany
3. melancholicy - osoby refleksyjne, wrażliwe, które cechuje pesymizm i brak wiary w siebie
4. sangwinicy - ludzie żywi, aktywni, optymiści, weseli i tryskający energią

Inny podział opracowali na przykład specjaliści od programowania neurolingwistycznego (NLP) [10], którzy wyróżnili sześć tak zwanych metaprogramów:

1. wewnętrzny/zewnętrzny - określa źródło autorytetu dla osoby - czy jest nim ktoś inny czy też ona sama
2. od/do (unikanie/dążenie) - określa, czy dla osoby ważne jest to czego chce, czy to czego nie chce
3. podobieństwa/różnice - określa, czy osoba w procesie poznawczym koncentruje się na podobieństwach, czy też różnicach

4. ja/inni - określa stosunek do świata, wskazuje czy osoba koncentruje uwagę na sobie czy na innych

5. opcje/procedury (możliwość/konieczność) - określa czy osoba ceni wolność wyboru czy też woli być ograniczona procedurami

6. szczegół/ogół - określa, w jaki sposób osoba odbiera informacje płynące z otaczającego świata

O typie osobowości człowieka, jakkolwiek je skategoryzujemy i nazwiemy, możemy wnosić w poprzez interakcję z nim w dwojaki sposób: poprzez analizę jego postępowania (w tym mowę ciała) oraz poprzez analizę tego co mówi, czyli jego słownictwa. W tej pracy szczegółowo poruszony zostanie ten drugi aspekt, a więc próba rozpoznania osobowości człowieka na podstawie analizy jego aktywnego słownika słów i zwrotów. Zanim jednak to nastąpi, krótka prezentacja wybranej do realizacji typologii ludzkiej osobowości.

2.2.1. Opis typologii osobowości wg dr Adriana Horzyka

Do stworzenia portretów psychologicznych rozpoznawanych osób zostanie wykorzystana typologia dr Adriana Horzyka, zawarta w źródłach [2], [3], [4] i [5].

Typologia ta wyróżnia 12 podstawowych typów osobowości:

1. **Dominujący (DOM)**

   Osoba taka lubi rządzić, mówić i decydować w imieniu innych. Ma swoje zdanie na wiele tematów i nie boi się go otwarcie wyrażać. Lubi, gdy inni pytają ją o opinię, poddają się jej decyzjom. Nie lubi być nikomu podporządkowana, wykonywać cudzych poleceń, nie toleruje nakazów i zakazów. Nie znosi, gdy pozbawia się ją możliwości wyrażenia własnego zdania. Denerwuje się, gdy ktoś inny krytykuje jej zdanie bądź je podważa. Nade wszystko ceni wolność i niezależność.

   Charakterystyczne słowa i zwroty dla tego typu osobowości: *ja, mój, moje, my, nasze, według mnie, moim zdaniem, w mojej opinii, z mojego punktu widzenia, myślę że, uważam, postanowiłem, zdecydowałem, powinienem, muszę, chcę, wymyśliłem, zarządziłem, wybrałem, rozpatrzyłem, zdecydowałem, nie zgadzam się, wierz mi, potrzebuje, decyduję, kontroluję, jestem niezależny, sterować, prowadzić, kierować, rekomendować, polecać, itp.*

2. **Maksymalista (MAK)**

   Osoba o typie osobowości maksymalisty z reguły bardzo wysoko zawiesza sobie poprzeczkę, stawia przed sobą bardzo ambitne cele. Chce mieć to co największe, najlepsze, unikalne, ekstremalne. Jest perfekcjonistą - dąży do osiągnięcia doskonałości we wszystkim, w co się angażuje. Nie ma dla niego rzeczy niemożliwych, dlatego w swym otoczeniu uchodzi za optymistę. Chce też zarazić swym optymizmem innych - motywuje ich do działania, przekonując, że wszystko im się uda.

   Charakterystyczne słowa i zwroty dla tego typu osobowości: *wszystko się uda, wszystko będzie w porządku, nie będzie żadnych kłopotów, nie ma problemu, poradzimy sobie, dasz/damy radę, możemy osiągnąć więcej, musimy to ulepszyć, poprawić, podoba mi się, wielki, duży, ogromny, ekstremalny, olbrzymi, maksymalny, szybki, gwałtowny, prędki, błyskawiczny, ładny, cudowny, piękny, wyjątkowy, rzadki, lepszy, najlepszy, większy, największy, więcej, najwięcej, bardziej, najbardziej, wyższy, najwyższy, naj-, super-, hiper-*

3. **Inspirujący (INS)**

   Osoby o takim typie osobowości to zwykle artystyczne dusze. Odczuwają ciągłą potrzebę szukania inspiracji, poszukiwania rzeczy nowych, fascynujących i tajemniczych. Działają spontanicznie i nieszablonowo. Uwielbiają niespodzianki. Nie lubią działać według narzuconych norm, stereotypów. Nie potrafią trzymać się planu, preferują wolność i spontaniczność. Są bardzo kreatywne i pomysłowe. Potrafią rzucić świeże spojrzenie na daną sprawę i zaproponować nowe, niestandardowe, ciekawe rozwiązanie każdego problemu.

   Charakterystyczne słowa i zwroty dla tego typu osobowości: *wymyśliłem, wpadłem na pomysł, skojarzyło mi się, moja (nowa) koncepcja jest taka, nagle wpadł mi do głowy pomysł, zaraz coś wymyślę, zmieniłem zdanie, zmiana planów, spontanicznie, na poczekaniu, pomysł, idea, niespodzianka, nieoczekiwany, nieznany, nowy, rewolucja, inspirujący, fantastyczny, tajemniczy, magiczny, sztuka, nastrój, artystyczny, pomysłowo, stworzyć, twórczy, kreacja, kreatywny, konstruować*

4. **Odkrywczy (ODK)**

   Osoba taka jest ciekawa świata, dąży do jak dogłębnego poznania świata, przyczyn i zasad nim rządzących. Fascynuje ją to co nowe, nieznane, niezbadane. Ma naturalne zdolności obserwatora, badacza, odkrywcy. Zadaje setki pytań, jest dociekliwa, lubi obserwować i poznawać. Chce przebywać z osobami, od których może się czegoś nowego dowiedzieć, lubi z nimi dyskutować i rozprawiać na ciekawe dla siebie tematy. Nie lubi

zajęć rutynowych, podczas których nie może się rozwijać, odkrywać, badać. Chce cały czas pogłębiać wiedzę, rozwijać umiejętności i nabywać doświadczenie.

Charakterystyczne słowa i zwroty dla tego typu osobowości: *dlaczego, po co, w jakim celu, co przez to rozumiesz?, porozmawiajmy o tym, co sądzisz o..., powiedz mi, jakie jest twoje zdanie na temat..., wyjaśnij mi, rozwiń to, muszę się tego dowiedzieć, to ciekawe, interesujące, nie wiedziałem tego, wnioskuję że, z tego wynika, przeczytałem, dowiedziałem się, słyszałem, pytanie, wyjaśnienie, odkrycie, zrozumieć, sprawdzić, porównać, rozpoznać, relacja, związek, powody, przyczyny, penetrować, klasyfikować, łączyć, myśleć, zastanawiać się nad, rozmyślać, rozważać, dumać*

5. **Weryfikujący (WER)**

Osoba o takim typie osobowości lubi weryfikować, kontrolować i oceniać czyny innych. Udziela porad, dostrzega niedoskonałości, wady, błędy, wskazuje możliwości poprawek, zrobienia czegoś inaczej, lepiej. Jest osobą spostrzegawczą, o sporej dozie krytycyzmu zarówno względem innych jak i w stosunku do siebie. Często odbierana jest jako osoba krytyczna. Nie lubi, gdy ktoś ignoruje jej spostrzeżenia, nie stosuje się do jej rad, robi coś po swojemu.

Charakterystyczne słowa i zwroty dla tego typu osobowości: *nie, nienajlepszy, niepoprawnie, niewłaściwie, nieprawidłowo, niedokładnie, błąd, niedociągnięcia, uchybienia, braki, popraw, przyłóż się bardziej, postaraj się, skup się, można to było zrobić lepiej, źle, nie tak, radzę ci, zrób to lepiej tak, zwróć uwagę, nie masz racji, twój sposób jest błędny, wada, błąd, pomyłka, krzywo, fatalnie, okropnie, marnie, niedbale, niedobrze, paskudnie, kiepsko, zepsute, opuszczono, naprawić, poprawić, sprostować*

6. **Systematyczny (SYS)**

Typ systematyczny jest osobą poukładaną, lubiącą gdy wokół niej panuje ład i porządek. Lubi wszystko porządkować, sortować, układać w jakiejś kolejności. Wszystkie swoje działania stara się zaplanować, trzymać się z góry ustalonego terminu, rutynowych procedur. Lubi tworzyć harmonogramy, planować każdą chwilę swojego życia. Jest bardzo punktualny, nie toleruje też gdy ktoś inny krzyżuje jej plany spóźniając się na spotkanie. Nie lubi niespodziewanych zmian planów, reguł. Każde spontaniczne działanie wymaga u niego czasu aby móc przebudować swoje plany. W swoich wypowiedziach często stosuje wyliczenia, systematyzm, porządek chronologiczny.

Charakterystyczne słowa i zwroty dla tego typu osobowości: *po pierwsze, po drugie, trzecie, czwarte, piąte, szóste..., po kolei, stopniowo, przede wszystkim, najpierw, wpierw, na*

początku, na wstępie, potem, po tym, następnie, w dodatku, w końcu, na koniec, kończąc, podsumowując, reasumując, ponadto, w dodatku, poza tym, zaplanujmy to, nie wszystko na raz, bałagan, nieporządek, chaos, porządek, sortować, układać, sekwencja, układ, pozycja, ranking, systematyzować, poziom, etap, klasyfikować, grupa, chronologia, lista, plan, rozkład, termin, kalendarz, rozmieszczenie, ułożyć, kompozycja, struktura, model, organizacja, zaplanować, ułożyć, rozplanować, rozłożyć, podzielić, rozdzielić, czas, na czas, data, termin, deadline

7. **Asekuracyjny (ASE)**

   Typ asekuracyjny dąży do zabezpieczenia wszelkich swoich poczynań. Jest pesymistą, woli zakładać czarne scenariusze i odpowiednio się na nie przygotowywać. Lubi zawczasu tworzyć plany awaryjne w razie wszelkich niepowodzeń. Przywiązuje dużą wagę do bezpieczeństwa i ochrony przed zagrożeniami. Nie lubi podejmować ryzyka, często ostrzega przed nim innych. Wiedzie życie wolne od trosk, niepowodzeń i problemów, a to dlatego, że zwykle zawczasu je przewiduje i nie podejmuje nawet działań, które mogłyby je sprowokować. Na wszystko oczekuje możliwie długiej gwarancji.

   Charakterystyczne słowa i zwroty dla tego typu osobowości: *ale, przecież, na czarną godzinę, na wszelki wypadek, bo może się przydać, w zapasie, w rezerwie, w razie czego, dla bezpieczeństwa, zostaw/zachowaj trochę na później, gdyby, ryzyko, niebezpieczeństwo, zapamiętaj moje słowa, żebyś nie żałował, ostrzegam, a nie mówiłem, uważnie, ostrożnie, ochronić, uchronić, środki ostrożności, groźba, ochrona, zabezpieczenie, pewność, alarm, gwarancje*

8. **Oszczędny (OSZ)** [5]

   Typ ten lubi oszczędność i rozsądne gospodarowanie zasobami. Nie cierpi rozrzutności i gdy coś się marnuje. Wszystkie dokonywane przez siebie zakupy analizuje pod kątem powtórnego ich wykorzystania i zaoszczędzenia w ten sposób pieniędzy. Lubi także znajdować nowe zastosowania dla pozornie zużytych już i niepotrzebnych rzeczy. Rzadko zdarza mu się cokolwiek po prostu wyrzucić, woli raczej to przerobić, zutylizować. Nie lubi ludzi rozrzutnych, którzy lekkomyślnie gospodarują rzeczami, stara się ich przekonać by nie pozostawiali odpadów, wykorzystywali wszystko do końca. Często w parze z tym typem idzie skąpstwo, choć nie zawsze.

   Charakterystyczne słowa i zwroty dla tego typu osobowości: *oszczędny, oszczędność, oszczędnie, oszczędzać, ekonomicznie, ponownie wykorzystać, przetworzyć, odzyskać coś,*

z odzysku, zaadaptować, nie zostawiać resztek, odpadów, zutylizować, marnotrawić, marnotrawstwo, rozrzutność, rozrzutny

9. **Harmonijny (HAR)**

   Osoba o tym typie osobowości ceni spokój i harmonię międzyludzką. Przyjmuje postawę ugodową, nie angażuje się w żadne konflikty, a gdy już do nich dochodzi, stara się je jak najszybciej załagodzić. Często unika wyrażania własnej opinii na dany temat, jeśli wie, że mogłoby to prowadzić do potencjalnego sporu. Woli pójść na daleko idące ustępstwa, nawet kosztem własnego interesu, niż doprowadzić do nieprzyjemnych sytuacji. Podczas rozmowy zwykle przytakuje temu co mówi rozmówca, nawet jeśli w głębi duszy ma odmienne zdanie. Stosuje taktykę "ściemniania" i "owijania w bawełnę". Używa także zdrobnień i eufemizmów, aby nie sprawiać wrażenia osoby twardej, zdecydowanie walczącej o swoje.

   Charakterystyczne słowa i zwroty dla tego typu osobowości: *może rzeczywiście, tak jak mówisz, nie przeczę, zgadzam się, no dobrze, ok, tak, dobrze, dobry, nie ma problemu, zgoda, w zasadzie tak, niby racja, skoro tak mówisz, niech ci będzie, rób jak uważasz, trochę, troszkę, troszeczkę, nie za bardzo, niedużo, nie tak bardzo, mały, odrobina, odrobinka, prawie, jakby, jak gdyby, jakby to powiedzieć, że tak powiem, nie zrozum mnie źle, nie chcę sprawić przykrości, malutki problem*

10. **Empatyczny (EMP)**

    Osoby takie są bardzo emocjonalne, wrażliwe, czułe. Przywiązują dużą wagę do kontaktów międzyludzkich, są otwarte i wylewne. Lubią długie, szczere rozmowy z innymi na temat życia, problemów, przeżyć i uczuć. Lubią się zwierzać, odsłaniać swoje wnętrze i swe intencje. Potrafią wczuć się w sytuację innych, mają zdolność do empatii, współczucia. W rozmowach używają wielu zdrobnień, również starają się, by kogoś nie zranić. Ich wypowiedzi są bardzo długie, rozbudowane, pełne aluzji i odstąpień od głównego wątku. Nie lubią od razu zmierzać do celu. Starają się szybko przechodzić na "Ty", by spoufalić się z rozmówcą. Są bardzo rodzinne - rodzina to jeden z ich ulubionych tematów, podobnie jak zwierzęta, piękno przyrody, itp.

    Charakterystyczne słowa i zwroty dla tego typu osobowości: *jak się masz, co u Ciebie, opowiedz mi o swoim problemie/sytuacji, naprawdę?, ładny, miły, dzieci, rodzina, strapienie, troska, cierpienie, męka, ból, krzywda, niedola, współczuć, litość, przykro mi, żałować, pomóc, rozumieć, zamierzać, zamiar, z zamiarem, zamierzać, intencje, w celu, dlatego, ponieważ, bo, gdyż, jako że, z powodu*

11. **Zadaniowy (ZAD)**

Typ ten ukierunkowany jest na wykonywanie zadań, zmierzanie prosto do celu. Ma zawsze jasno wyznaczony cel, do którego zmierza. Przed rozpoczęciem jednego zadania, starają się zamknąć wszystkie poprzednie, tak by móc się skupić tylko nad tym jednym. Nie lubi tracić czasu na mało ważne, jego zdaniem, rzeczy. Niecierpliwi go owijanie w bawełnę, chce jak najszybciej wiedzieć wszystko czego potrzebuje i nic poza tym. Drażni go brak konkretów, zbędne uprzejmości i konwenanse. Zwykle nie potrafi nawiązywać bogatych relacji międzyludzkich, trudno jest mu się otworzyć i rozmawiać o sprawach, które są poza orbitą jego zainteresowań.

Charakterystyczne słowa i zwroty dla tego typu osobowości: *muszę, chcę, potrzebuję, konkrety, co konkretnie masz na myśli, dokładniej, przejdźmy do sedna, przejdźmy od razu do meritum, do rzeczy, szybciej, nie ma się nad czym zastanawiać, skończmy to jak najszybciej, chcę to już mieć z głowy, załatwmy to raz a dobrze, teraz następna sprawa, najpierw muszę zrobić to, potem muszę zrobić tamto, wydajność, osiągi, działanie, wydajny, szybki, konkretny, konkrety, fakty, przedmiot, zadanie, praktyczny, skończyć, skończony, gotowy, zmierzać do końca/do celu*

12. **Równoważący (RÓW)**

Osoby o takim typie osobowości dążą do równoważenia wszystkiego. Nadrzędnymi ich wartościami są równowaga i sprawiedliwość. Są wrażliwe na wszelkie odstępstwa od ogólnie przyjętych norm, zasad i regulaminów. Nie lubią niesprawiedliwości, bezprawia, dyskryminacji, poniżania innych. Starają się zawsze postępować sprawiedliwie. Są zwolennikami demokracji, równouprawnienia i równości wszystkich ludzi. Gdy komuś dzieje się krzywda, natychmiast reagują.

Charakterystyczne słowa i zwroty dla tego typu osobowości: *równowaga, balans, przeciwwaga, równoważyć, kompensować rekompensować, równo, konsekwentny, sprawiedliwy, sprawiedliwość, uczciwie, uczciwy, fair, nie fair, nieuczciwy, niesprawiedliwy, sądzić, zasada, reguła, odwzajemniać, odpłacać, wyrównać rachunki, tak będzie sprawiedliwie, tak będzie najlepiej dla wszystkich, sprawiedliwości stanie się zadość, najuczciwiej będzie, żeby nikt nie był pokrzywdzony, to nie fair, to niesprawiedliwe, to oburzające, to wbrew zasadom, nie mogę mu tego zrobić*

2.3. Istniejące rozwiązania

Istnieje szereg systemów wykorzystujących analizę tekstu w celu wyciągnięcia z niego użytecznych informacji, w celu ich późniejszego przetworzenia i zwrócenia odpowiedniego rezultatu.

Jednym z takich projektów jest praca magisterska [11], mająca za temat samoadaptacyjny sklep internetowy, obsługiwany przez inteligentnego cybersprzedawcę. Działa ona na zasadzie chatbota, prowadzącego rozmowę z klientem. Na podstawie słów zawartych w wypowiedziach klienta, rozpoznaje jego osobowość i dostosowuje do niej styl wygłaszanych przez siebie odpowiedzi, udziela też rad takich, aby odpowiadały typowi osobowości klienta, przez to zwiększając jego zadowolenie. Proces analizy tekstu w tejże pracy wygląda następująco:

1. Preprocessing - oczyszczenie tekstu ze znaków innych niż litery, cyfry, kropki i pytajniki. Podział zdania na tokeny - najmniejsze jednostki logiczne, rozdzielone spacjami.

2. Sprowadzenie do form podstawowych - w tym etapie każdy wyraz sprowadzany jest do swojej formy podstawowej, przy czym jeśli dany wyraz może wywodzić się od wielu form podstawowych, to sprawdzane są wszystkie możliwości, np. *nie* może być zarówno przeczeniem, jak i formą biernika zaimków *ona* i *ono*; *dam* to zarówno przyszła forma dokonana czasownika *dać*, jak i dopełniacz liczby mnogiej rzeczownika *dama*. W związku z tym sprawdzane są wszystkie możliwości, sensowne lub nie: np. *nie dam* $\rightarrow$ *nie dać*, *ona dać*, *ono dać*, *nie dama*, *ona dama*, *ono dama*.

3. Podmiana synonimów - zamiana wzorców, różniących się pojedynczymi wyrazami bliskoznacznymi poprzez zastępowanie tych wyrazów synonimami

4. Dopasowanie wygenerowanych zdań do wzorców znajdujących się w bazie danych. Aby nastąpiło dopasowanie, wszystkie wyrazy wzorca muszą znaleźć się w wypowiedzi klienta oraz ich kolejność musi być zgodna ze wzorcem. Następnie tworzony jest ranking dokładności dopasowań. Współczynnik dokładności definiowany jest iloczynem liczby dopasowanych wyrazów wzorca i ich wagi. Wzorzec, który zajmie pierwsze miejsce w tym rankingu, a więc ten, którego dopasowanie ma największą wartość współczynnika jest przekazywany do Zarządcy reagującego na wypowiedź klienta.

Dodatkowo aplikacja zawiera rozbudowane algorytmy rozmowy, która może być prowadzona nie tylko na temat zakupów, ale także na inne, niezwiązane tematy, z których również wnioskuje o osobowości badanej osoby.

Kolejnym projektem zajmującym się analizą tekstu, również pod kątem badania ludzkiej osobowości jest praca dyplomowa Maksymiliana Imioło [12]. W celu analizy programów osobowości używa on również typologii dr Horzyka. Teksty do analizy pozyskiwane są podczas wirtualnej rozmowy z użytkownikiem na jeden z grupy pięciu tematów: telefon komórkowy, wycieczka, dom, jedzenie, samochód.

Podczas analizy tekstu jego aplikacja wychwytuje charakterystycznie słowa i frazy występujące w bazie danych wraz z wagami do nich przypisanymi. Po wychwyceniu słowa, waga mnożona jest przez 1, natomiast po wychwyceniu frazy - przez 2, z uwagi na charakterystykę wystąpienia frazy. Kilkukrotne wystąpienie słowa lub frazy proporcjonalnie zwiększa intensywność danego programu. W wypadku gdy jedno słowo przypisane jest do więcej niż jednego typu osobowości, o tym, do którego typu go przypisać decyduje ilość wystąpień innych słów dla obu typów. Słowo zostaje przypisane do programu intensywniejszego.

Nowością w porównaniu z pracą Miklaszewski/Magierski jest wprowadzenie detekcji zdrobnień, które to są charakterystyczne dla typu Harmonijnego (HAR) i Empatycznego (EMP). Typowe końcówki wyrazów będących zdrobnieniami, to według tej pracy: *-czek, -szek, -czko, -czyk, -czki, -eńki, -sio, -tka, -tko, -sia, -ula, -chna, -uś, -unia, -unio, -ina, -lka, -utki, -cyk, -ek, -ka, -ko, -ik, -ak, -eńko, -aś, -uśki, -uchny*. Ponieważ zdrobnienia mogą być charakterystyczne dla dwóch typów, przypisywane są one temu typowi, dla którego wystąpień innych słów jest więcej. Do rozpoznawania typów wprowadzone zostały również wykrzyknienia, czyli frazy zakończone wykrzyknikiem. Analogicznie, w zależności od aktualnych wartości współczynników przypisywane są one do typu Dominującego (DOM) lub Weryfikującego (WER).

Niestety, nie wszystkie z tych sufiksów, wydają się być dobrane trafnie. Przykładowo, w języku polskim występuje mnóstwo słów zakończonych na *-tka, -ina, -ek, -lka, -ka, -ak*, nie będących zdrobnieniami, np. *matka, kurtka, godzina, malina, worek, stołek, pralka, belka, ławka, książka, ptak, barak*. W rozmowie na temat telefonów komórkowych czy urlopu częstość występowania tych wyrazów mogła być znikoma, jednak w tworzonej aplikacji trzeba będzie zmodyfikować tę listę.

Słowa pozyskiwane przez aplikację są przekształcane na postać bez tzw. "polskich znaków", co z jednej strony może być dobre, z uwagi na to, iż coraz więcej ludzi we współczesnym Internecie chce jak najbardziej uprościć sobie życie i pisze bez używania klawiszy Alt i Shift, czego efektem jest tekst pozbawiony polskich znaków diakrytycznych i wielkich liter. Tak więc zarówno słowo zapisane jako *dużą*, *duża* jak i *duza* zostaną prawidłowo rozpoznane i sklasyfikowane jako program Maksymalistyczny. Z drugiej jednak strony jako typ Asekuracyjny zostaną sklasyfikowane zarówno słowo *ale* (prawidłowo) jak i *Alę*, nawet jeśli wypowiadająca go osoba jedynie wspomina o swojej znajomej, a z typem Asekuracyjnym może mieć niewiele

wspólnego.

Niestety, aplikacja ma dość wąski zasób słownictwa, ograniczający się do form podstawowych, przy czym nie sprowadza słów znajdujących się w wypowiedzi do tych form. W związku z tym słowo *duża* zostanie rozpoznane prawidłowo, również słowo *dużą* (ale tylko przez szczęśliwy zbieg okoliczności polegający na tym, że formy te pozbawione znaków diakrytycznych są identyczne). Natomiast inne formy gramatyczne wyrazu, takie jak *dużej*, *dużymi*, *dużych* nie zostaną niestety sklasyfikowane jako reprezentujące odpowiedni typ osobowości. Jest to, w mojej ocenie, spory mankament.

3. Opis i implementacja proponowanego rozwiązania

Proponowany system stanowi połączenie metod frekwencyjnych oraz typologii ludzkiej osobowości w celu rozpoznawania autorów.

3.1. Projekt rozwiązania

Na początku należy się zastanowić, jakiego typu aplikacja będzie najlepiej realizować postawiony cel - czy ma być to aplikacja stacjonarna, czy webowa. Na korzyść aplikacji webowej przemawia łatwość dostępu z każdego miejsca na świecie oraz brak konieczności instalowania po stronie klienta jakiegokolwiek oprogramowania, z wyjątkiem przeglądarki internetowej, która i tak zwykle wchodzi w skład systemu operacyjnego. Można więc taką aplikację uruchomić na każdym rodzaju komputera. Jedynym wymaganiem jest dostęp do Sieci, co w dobie Internetu nie stanowi większego problemu. Te właśnie czynniki zadecydowały o wyborze do realizacji aplikacji webowej.

Następne pytanie, jakie się nasuwa, to wybór technologii w jakiej realizowany będzie projekt. Jeśli aplikacja webowa, to na pewno potrzebne będą elementy hipertekstowego języka znaczników HTTP, w którym zrealizowana jest zdecydowana większość obecnie występujących stron internetowych. Ponadto, do ostylowania strony zostaną wykorzystane kaskadowe arkusze stylów CSS, standardowa obecnie technologia niemal nieodłącznie towarzysząca HTML. Do realizacji logicznej części całej aplikacji wykorzystany zostanie obiektowy, skryptowy język programowania zaprojektowany do generowania stron internetowych w czasie rzeczywistym - PHP. Argumentem, jaki zadecydował o wyborze tego właśnie języka była jego względna popularność oraz chęć połączenia realizacji pracy dyplomowej z podniesieniem poziomu znajomości tego języka, zgodnie z zasadą, że nic tak dobrze nie uczy jak praktyka. Pomocniczo zostaną użyte także skrypty obiektowego języka programowania JavaScript, wykonywane po stronie klienta. Dane niezbędne do prawidłowego funkcjonowania aplikacji przechowywane będą w

relacyjnej bazie danych MySQL.

Po wyborze technologii, czas na znalezienie odpowiedniego serwera, na którym działała będzie aplikacja. Ze względu na zerowy budżet projektu w grę wchodziły jedynie darmowe serwisy hostingowe. Wybór padł na polski `http://www.ugu.pl`, oferujący za darmo m. in.:

- hosting WWW
- 250 MB dla bazy MySQL 5
- dostęp do PHP 5.2
- 150MB dla konta
- nieograniczony transfer - brak limitów na generowany ruch
- FTP
- DNS

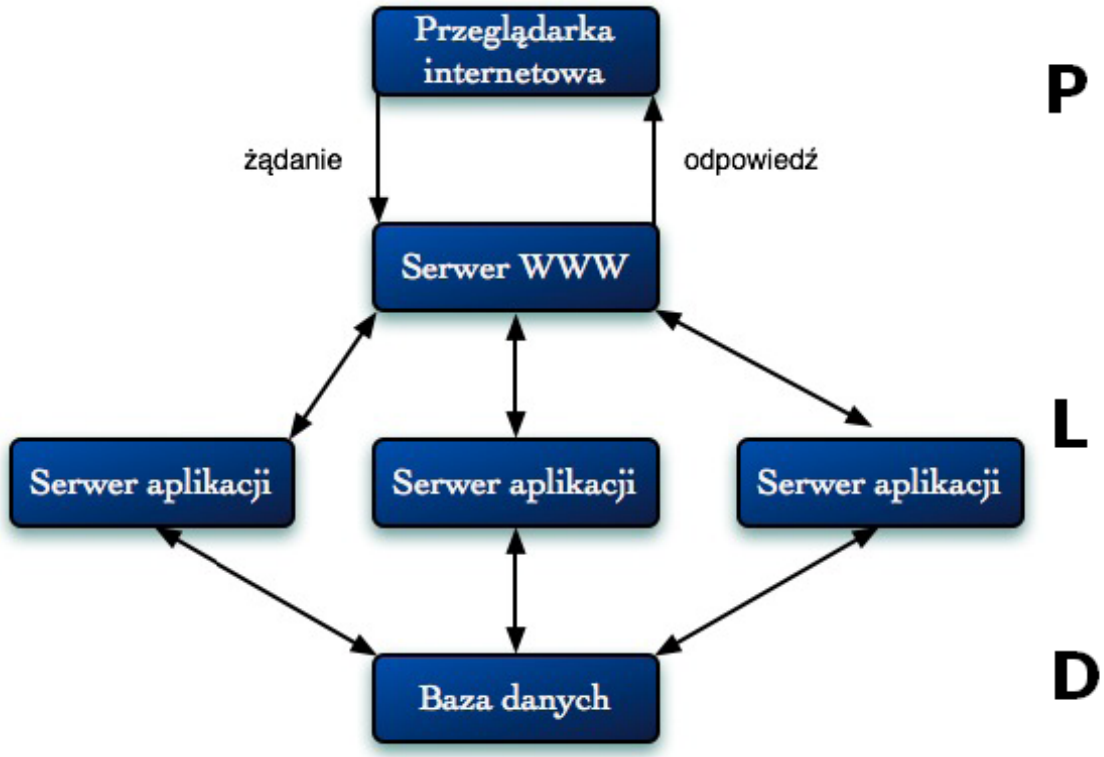
Usługi te są więcej niż wystarczające dla tworzonej aplikacji, a sam serwer wydaje się szybki i stabilny.

3.1.1. Architektura systemu

Architektura trzywarstwowa (ang. three-layer architecture) jest sposobem takiej budowy systemu informatycznego, w którym jest on podzielony na warstwy:

1. warstwa danych - informacje, które system wykorzystuje najczęściej składowane w bazie danych. Warstwę tę stanowić będzie relacyjna baza danych SQL.
2. warstwa przetwarzania danych - odpowiedzialna za logikę biznesową, czyli wszelkie operacje przetwarzające dane pomiędzy warstwami sąsiadującymi. Technologią odpowiadającą za działanie tej części będzie język skryptowy PHP.
3. warstwa prezentacji, czyli interfejs użytkownika odpowiedzialny za wizualne przekazanie mu informacji i pozwalający na interakcję z systemem. Interfejs użytkownika zbudowany jest z użyciem technologii HTML i CSS.

Na rysunku 3.1 widoczna jest architektura typowej aplikacji webowej. Poszczególne warstwy oznaczone zostały literami.



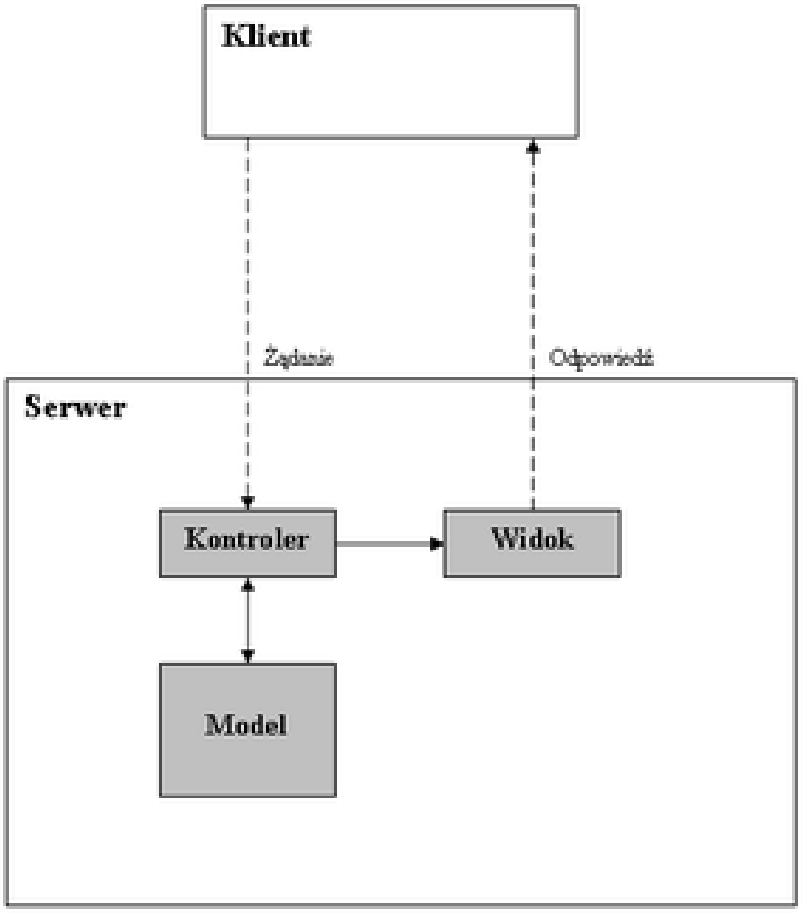
Rysunek 3.1: Struktura aplikacji internetowej. Literami oznaczono: D - warstwa danych, L - wartwa logiki, P - warstwa prezentacji [11]

Z trójwarstwową architekturą systemu powiązany jest wzorzec projektowy MVC (ang. *Model - View - Controller*), czyli Model - Widok - Kontroler [13]. Jest to wzorzec zakładający istnienie trzech powiązanych ze sobą modułów:

- modelu - czyli danych, na których program operuje
- widoku - część programu prezentująca dane użytkownikom
- kontrolera - czyli komponentu odpowiedzialnego za wykonywanie akcji w odpowiedzi na działania użytkownika

Zastosowanie takiego wzorca projektowego przynosi następujące korzyści:

- dekompozycję programu na mniejsze moduły, z których każdy ma inny zakres odpowiedzialności
- możliwość zmiany implementacji każdego z modułów osobno, bez większego wpływu na pozostałe



Rysunek 3.2: Wzorzec MVC

Schemat wzorca przedstawiono na rysunku 3.2.

Powstający system zbudowany zostanie w oparciu o taki właśnie model.

Przeprowadzone w rozdziale 2 badania literaturowe wyznaczyły przybliżony kierunek, w którym podążać powinna realizacja aplikacji, a więc parser, który będzie wczytywał wprowadzany tekst do bazy danych, rozbijając go na atomy w postaci słów i fraz oraz część logiczna, która przeanalizuje wczytany tekst i zwróci oczekiwany rezultat.

W porównaniu ze źrodłem [11] nową jakością będzie praca na słowach w takiej formie, w jakiej występują, a więc bez sprowadzania ich do formy podstawowej. Implikuje to konieczność przygotowania takiej bazy danych, która umożliwi operowanie na dowolnej formie gramatycznej wyrazu, w szczególności rozpoznawanie typu osobowości, niezależnie od przypadku, liczby czy rodzaju w jakim zostało użyte słowo. Wyeliminowany zostanie także mankament

z pracy [12], polegający na parsowaniu słów bez polskich znaków diakrytycznych i rozpoznawaniu tylko podstawowych form gramatycznych.

Sam temat pracy zakłada rozpoznawanie autorów, czyli konieczność porównywania jednego nieznanego tekstu ze zbiorem tekstów znanych, w celu wyłonienia tego, którego autor z największym prawdopodobieństwem jest autorem tekstu nieznanego. Takie postawienie zadania implikuje podział aplikacji na dwa moduły:

- moduł związany z rozbudową bazy danych znanych autorów
- moduł związany stricte z rozpoznawaniem nieznanego tekstu

W związku z wykorzystaniem typologii osobowości wg dr Horzyka opisanej w podrozdziale 2.2.1, celowe wydaje się stworzenie również modułu służącego do zarządzania słowami i frazami charakterystycznymi dla poszczególnych typów osobowości.

Ogólny plan modułów aplikacji przedstawia się więc następująco:

1. Rozbudowa bazy autorów
    - dodawanie autora do bazy
    - statystyki autorów z bazy
    - usuwanie autora z bazy
2. Rozbudowa bazy słów i fraz
3. Rozpoznawanie
    - dodawanie tekstu do rozpoznania
    - rozpoznawianie autorstwa
    - usuwanie tekstu
    - przesuwanie tekstu do bazy

3.1.2. Algorytm działania aplikacji

Pierwszym krokiem będzie pozyskanie odpowiedniej ilości tekstów źródłowych pewnej grupy autorów. Aby ich analiza przyniosła miarodajne dane, powinny być one możliwie długie. Następnym krokiem jest wprowadzenie tekstów do bazy danych oraz rozbicie ich na atomy, czyli najmniejsze cząstki, które będą podlegać analizie. Będą to pojedyncze słowa oraz

frazy składające się z dwóch i trzech wyrazów. Następnie dla takich atomów przeprowadzona zostanie analiza pod kątem częstotliwości występowania słów i zwrotów, a także analiza programów osobowości autora powiązanych ze słowami jakich użył. Uzyskane w ten sposób wyniki stanowić będą podstawę do późniejszego porównania z nowym, nie znajdującym się w bazie tekstem. Zostanie teraz przedstawiona autorska metoda zastosowana w aplikacji do rozpoznawania autorstwa takiego tekstu.

3.1.3. Opis matematyczny używanej metody

W celu porównania dwóch tekstów, należy mieć jakąś metrykę, którą mierzyć będziemy jak bardzo są one do siebie podobne, względnie jak bardzo się od siebie różnią. W wypadku porównywania pewnego tekstu z nim samym metryka taka powinna dawać wynik 100% podobieństwa (0% różnic). W tym wypadku teksty porównywane są pod względem aktywnych słów i zwrotów jak również pod względem osobowości autora. Stworzona metryka powinna więc brać pod uwagę obydwa te czynniki. Dodatkowo, ze względu na to, iż aspekt osobowościowy używany jest jako pomocniczy, powinien mieć jednak nieco mniejszą wagę od frekwencyjnego. W toku prac nad projektem wyklarował się ogólny wzór 3.1 będący miarą podobieństwa dwóch tekstów. Zostanie on teraz przedstawiony, a następnie jego równania składowe zostaną dokładnie omówione.

Miarę podobieństwa między dwoma porównywanymi autorami - znanym (Z) i rozpoznawanym (X) opisać można ogólnym wzorem:

$$\Delta = 100 - (A + B + C + D + E) \tag{3.1}$$

w równaniu tym:

$$A = 0.4 \cdot \sum_{i=1}^{12} |\underline{x}_i - \underline{z}_i| \tag{3.2}$$

$$B = 0.5 \cdot \sum_{i=1}^{20} |x_i - z_i| \tag{3.3}$$

przy warunku

$$x_i \geq 15 \vee z_i \geq 15$$

$$C = 100 \cdot (1 - 2 \cdot \frac{\overline{X_w \cap Z_w}}{\overline{X_w}}) \tag{3.4}$$

$$D = 100 \cdot (1 - 5 \cdot \frac{\overline{X_{f2} \cap Z_{f2}}}{\overline{X_{f2}}}) \tag{3.5}$$

$$E = 100 \cdot (1 - 10 \cdot \frac{\overline{X_{f3} \cap Z_{f3}}}{\overline{X_{f3}}}) \tag{3.6}$$

Użyte symbole oznaczają:

$\underline{x_i}, \underline{z_i}$ - znormalizowane wartości poszczególnych programów osobowości

x_i, z_i - znormalizowane częstotliwości występowania słów

$\overline{X_w}, \overline{Z_w}$ - moc zbioru wszystkich wyrazów występujących w wypowiedziach autora

$\overline{X_{f2}}, \overline{Z_{f2}}$ - moc zbioru wszystkich fraz dwuwyrazowych występujących w wypowiedziach autora

$\overline{X_{f3}}, \overline{Z_{f3}}$ - moc zbioru wszystkich fraz trójwyrazowych występujących w wypowiedziach autora

Wzór (3.1) zależy od pięciu składników A, B, C, D i E, wraz ze wzrostem których miara podobieństwa maleje.

Składnik A (3.2) określa podobieństwo programów osobowości autorów X i Z. Jest to suma modułów różnic znormalizowanej intensywności występowania wszystkich dwunastu programów osobowości u obu autorów. Innymi słowy, im bardziej rozbieżne są osobowości autora znanego i rozpoznawanego, tym suma ta jest większa. Moduł dodany został, aby upewnić się, iż każdy ze składników sumy będzie dodatni. W szczególnym przypadku porównywania tekstu z nim samym wartość składnika A wyniesie 0. W późniejszych badaniach wartość A została skalibrowana o czynnik 0.4.

Składnik B (3.3) operuje na maksymalnie dwudziestu słowach, które znajdują się w słownikach obydwu autorów jednocześnie, przy dodatkowym warunku, że u przynajmniej jednego z nich występowanie tego słowa jest na poziomie 15 % względem słowa najczęstszego. Poza tymi warunkami metryka jest tu identyczna jak w przypadku A, to znaczy suma modułów różnic. Wartość B informuje zatem jak często jeden z autorów używa wyrazów ulubionych przez drugiego i vice versa. Jeśli te ulubione wyrazy będą miały zbliżone częstotliwości występowania, może to prowadzić do konkluzji, że napisała te teksty jedna i ta sama osoba (oczywiście teksty muszą być odpowiednio długie). W przypadku identycznych tekstów znów wartość B wynosić będzie dokładnie 0. Podobnie jak B, wartość została skalibrowana o współczynnik 0.5.

Składnik C (3.4) określa procentowy stosunek słów wspólnych dla obydwu autorów do ogólnej liczby słów w słowniku autora rozpoznawanego. Analogicznie D (3.5) oznacza tę samą proporcję dla fraz 2-wyrazowych, a E - dla 3-wyrazowych. Podobnie jak poprzednio, dla identycznych tekstów składniki te również będą równe 0. W związku z tym wskaźnik Δ dla tego

samego tekstu będzie miał wartość 100, co będzie oznaczało 100 - procentową pewność, iż autorem obydwu tekstów jest jedna i ta sama osoba. Nie działa to jednak w przeciwną stronę - nie ma dolnego ograniczenia na wartość Δ, gdyż nigdy nie można z całą pewnością wykluczyć autorstwa tekstu, nawet jeżeli wszystkie znaki wskazują przeciwko temu. Teksty różnić mogą się problematyką, sytuacją w jakiej powstawały, w pewnych przypadkach mogło być nawet tak, że autor nie chciał, aby go rozpoznano, dlatego znacząco odbiegł od swego stylu. Tak więc malejąca w stronę zera, lecz nie osiągająca go wartość Δ jest jedynie sygnałem malejącego prawdopodobieństwa, w obliczu czego bardziej prawdopodobni stają się inni domniemani autorzy.

3.2. Implementacja rozwiązania

3.2.1. Zbieranie i przechowywanie danych

Pierwszym etapem będzie pozyskiwanie tekstów i zapisywanie ich do bazy. Stworzony w tym celu moduł insert_text.php umożliwia wprowadzenie tekstu do bazy w dwojaki sposób. Po pierwsze, tekst może być wpisany do formularza ręcznie lub przeklejony z jakiegoś źródła za pomocą kombinacji klawiszy Ctrl+C, Ctrl+V. Po wysłaniu formularza, tekst zapisywany jest w zmiennej $data. Drugą możliwością jest wczytanie tekstu z pliku tekstowego. W tym drugim przypadku plik pobierany jest na serwer, po czym tekst w nim zawarty kopiowany jest do zmiennej $data, a następnie poddawany jest dalszemu przetwarzaniu, identycznie jak w przypadku pierwszym, natomiast sam plik jest po tym fakcie kasowany, aby nie zajmować niepotrzebnie miejsca na serwerze.

Gdy już tekst zostanie wczytany, następuje jego oczyszczenie ze znaków interpunkcyjnych, symboli i innych. Usuwane są zeń znaki takie jak: - " ( ) @ # $ % & * _ = + ; ? ! . ,

Przed usunięciem z tekstu wykrzykników, zostaje zliczona ich liczba. Będzie ona później składową rozpoznania u autora dominującego typu osobowości.

Następnie, w celu ujednolicenia wyrazów pisanych wielką i małą literą, ze względów gramatycznych (początek zdania) czy też grzecznościowych (Ci, ci) dokonywane jest przekształcenie całego tekstu na małe litery.

Tak oczyszczony tekst można podzielić na wyrazy i zwroty. Używana jest do tego phpowska funkcja explode, przyjmująca za znak podziału spację.

Następnie tworzone są dynamicznie trzy tabele, nazywane od nazwiska bądź pseudonimu osoby, którą dodajemy do bazy. Przykładowo, dla autora 'szymborska', tabele nazywać się będą 'szymborska1', 'szymborska2', 'szymborska3'. Tabelę pierwszą wypełniają wszystkie po-

jedyncze słowa zawarte w podzielonym przed chwilą tekście. Jeśli słowo się powtarza, nie jest dodawany nowy wpis, tylko wpis już istniejący ma inkrementowaną wartość pola 'ilosc'. Tabela druga zawiera wszystkie możliwe dwójki słów następujących po sobie w tekście, czyli wszystkie (sensowne lub nie) zwroty dwuwyrazowe, wraz z ilością powtórzeń w tekście. Podobnie, tabela trzecia zawiera wszystkie frazy trójwyrazowe i ich ilość.

W tym momencie z tabeli pierwszej (słów) zostają usunięte przyimki i niektóre spójniki. Są to zależne części mowy, nie stanowiące samodzielnych słów. Występują zawsze w towarzystwie innych słów, a ich częstotliwość jest tak duża, że mogłyby zaburzać dane dla słów charakterystycznych dla konkretnych osób. Pełna lista przyimków i spójników usuwanych w tym kroku: *i, z, w, na, a, się, od, do, po, za, o, we, to, co*

Następnie tabele zostają znormalizowane. Ponieważ może się zdarzyć, iż będziemy dysponować próbkami tekstów o diametralnie różnych długościach, należy w jakiś sposób uczynić możliwym ich porównanie. Temu właśnie służy normalizacja. Słowo lub zwrot powtarzające się najczęściej w danej tabeli otrzymuje współczynnik 100, pozostałe ilości wystąpień przeliczane są względem niego. Współczynnik nazwany został 'iloscna100'.

Tabele 3.1 3.2 i 3.3 obrazują powstałą bazę słów i fraz dla krótkiego tekstu literackiego [14].

To ja, Kasandra.
A to jest moje miasto pod popiołem,
A to jest moja laska i wstążki prorockie,
A to jest moja głowa pełna wątpliwości.

Tabela 3.1: Przykładowa tabela zawierająca tekst podzielony na słowa

słowo	**ilosc**	**iloscna100**
to	4	100
ja	1	25
kasandra	1	25
a	3	75
jest	3	75
moje	1	25
miasto	1	25
pod	1	25
popiołem	1	25

c.d. na następnej stronie

Tabela 3.1 – c.d. z poprzedniej strony

słowo	ilosc	iloscna100
moja	2	50
laska	1	25
i	1	25
wstążki	1	25
prorockie	1	25
głowa	1	25
pełna	1	25
wątpliwości	1	25

Tabela 3.2: Przykładowa tabela zawierająca tekst podzielony na frazy dwuwyrazowe

słowo1	słowo2	ilosc	iloscna100
to	ja	1	33
ja	kasandra	1	33
a	to	3	100
to	jest	3	100
jest	moje	1	33
moje	miasto	1	33
miasto	pod	1	33
pod	popiołem	1	33
jest	moja	2	67
moja	laska	1	33
laska	i	1	33
i	wstążki	1	33
wstążki	prorockie	1	33
moja	głowa	1	33
głowa	pełna	1	33
pełna	wątpliwości	1	33

Tabela 3.3: Przykładowa tabela zawierająca tekst podzielony na frazy trójwyrazowe

słowo1	słowo2	słowo3	ilosc	iloscna100
to	ja	kasandra	1	33
a	to	jest	3	100
to	jest	moje	1	33
jest	moje	miasto	1	33
moje	miasto	pod	1	33
miasto	pod	popiołem	1	33
to	jest	moja	2	67
jest	moja	laska	1	33
moja	laska	i	1	33
laska	i	wstążki	1	33
i	wstążki	prorockie	1	33
jest	moja	głowa	1	33
moja	głowa	pełna	1	33
głowa	pełna	wątpliwości	1	33

Kolejnym krokiem po normalizacji jest zliczenie słów i zwrotów charakterystycznych dla 12 podstawowych typów osobowości, według typologii z rodziału 2.2.1. W bazie danych znajdują się tabele tblSlowaNaTypy (3.4), tblSlowanaTypy2 (3.5) oraz tblSlowaNaTypy3 (3.6), zawierające zestaw ok. 2400 słów i kilkaset zwrotów pogrupowanych według typów osobowości.

Tabela 3.4: Schemat tabeli "tblSlowaNaTypy1"

Pole	Typ
slowo	nvarchar(50)
typ_id	int
waga	float

Każde słowo i zwrot ma parametr oznaczający wagę, z jaką jest liczone. Z uwagi na to, że całe zwroty trudniej spotkać w tekście niż pojedyncze słowa, mają one odpowiednio większą wagę gdy już wystąpią. Wagi dla zwrotów 3-wyrazowych kształtują się w granicach 5-6, dla dwuwyrazowych: 3-4, dla pojedynczych słów: 1-2, z wyjątkiem słowa "nie", które może

Tabela 3.5: Schemat tabeli "tblSlowaNaTypy2"

Pole	Typ
slowo1	nvarchar(50)
slowo2	nvarchar(50)
typ_id	int
waga	float

Tabela 3.6: Schemat tabeli "tblSlowaNaTypy3"

Pole	Typ
slowo1	nvarchar(50)
slowo2	nvarchar(50)
slowo3	nvarchar(50)
typ_id	int
waga	float

charakteryzować typ Weryfikujący (WER), jednak poza tym występuje tak często w różnych kontekstach, że użycie go z tą samą wagą co innych powodowałoby zaburzenie systemu i uznanie wszystkich autorów za osoby weryfikujące. Słowu temu została nadana zatem waga 0.2. Dodatkowo do typów Harmonijnego (HAR) i Empatycznego (EMP) doliczane są zdrobnienia wychwytywane na podstawie końcówek z tabeli 3.7. W porównaniu z pracą [12] usunięte zostały niektóre końcówki, co do których wątpliwości przedstawiono w rozdziale 2.3.

Tabela 3.7: Zdrobnienia rozpoznawane przez aplikację

końcówka	przykładowe słowa
-czek	koteczek, Jureczek
-szek	łańcuszek, paluszek
-szki	maluszki, pieluszki
-czyk	mieczyk, haczyk
-czko	słoneczko, Aneczko
-eńki	maleńki, Oleńki
-sio	milusio, Czesio
-sia	pięknisia, Gosia
-utka	ładniutka, milutka
	c.d. na następnej stronie

Tabela 3.7 – c.d. z poprzedniej strony

końcówka	przykładowe słowa
-utko	malutko, króciutko
-ątko	kurczątko, jagniątko
-ątka	cielątka, zawiniątka
-ula	matula, babula
-uchna	córuchna, matuchna
-uś	pracuś, Wojtuś
-unia	babunia, Agunia
-unio	wnunio, tatunio
-ulka	koszulka, Urszulka
-utki	malutki, milutki
-ik	pokoik, lufcik
-yk	kocyk, piecyk
-eńko	króciuteńko, maleńko
-uchny	piękniuchny, bieluchny

Z kolei w tabeli 3.8 przedstawione zostały cząstki (prefiksy i sufiksy) charakterystyczne dla typu Maksymalistycznego (MAK). Te słowa również są zliczane i powiększają wynik dla tego programu.

Tabela 3.8: Prefiksy i sufiksy dla programu Maksymalista (MAK)

cząstka	przykładowe słowa
super-	superszybki
hiper-	hiperdokładność
naj-	najwspanialsza
-szy	lepszy
-ejszy	najładniejszy

Ponieważ poprzednia operacja była wykonywana na nieznormalizowanych wartościach

bazy, należy więc i tutaj znormalizować wynik, aby dalsze porównywanie miało sens. W tym celu wybrany zostaje ten z 12 programów, dla którego suma wag jest najwyższa, a więc program, który zaznacza się najmocniej. Jemu zostaje przypisana wartość 100, a pozostałe programy są przeskalowywane względem niego. Wynik normalizacji zapisywany jest w tabeli tblStatsPerc (3.9).

Tabela 3.9: Schemat tabel "tblStatswords", "tblStatsPhrases2", "tblStatsPhrases3", "tblStatsPerc", "tblStatsPercNew"

Pole	Typ
user	nvarchar(50)
DOM	int
MAK	int
INS	int
ODK	int
WER	int
SYS	int
ASE	int
OSZ	int
HAR	int
EMP	int
ZAD	int
RÓW	int

Po realizacji tego kroku otrzymujemy wreszcie końcowy rezultat, w postaci znormalizowanej bazy częstotliwości słów i zwrotów oraz znormalizowanego przekroju osobowości autora. Wynik ten można obejrzeć w postaci tabel lub też wykresu radarowego, jak na rysunku 4.5.

3.2.2. Rozpoznawanie nieznanego autora

Przejdźmy teraz do automatycznego rozpoznawania autorów.

Na początku trzeba wprowadzić rozpoznawany tekst do systemu. Dzieje się to w identyczny sposób jak przy tworzeniu bazy w paragrafie 3.2.1: najpierw wypełniany jest, ręcznie bądź z pliku, formularz. Następnie tekst zapisywany jest do zmiennej, na której dokonywany jest szereg operacji: usuwanie symboli i znaków przestankowych, zamiana całego tekstu na małe litery, wreszcie podział na słowa. Tak podzielony tekst zapisywany jest do trzech tablic,

tym razem o nazwach 'temp1', 'temp2' i 'temp3'. Tak jak poprzednio, do pierwszej z tablic trafiają słowa, do drugiej frazy dwuwyrazowe, a do trzeciej - frazy trójwyrazowe. Utworzone tabele są normalizowane. Następnie podliczone zostają wskaźniki dla każdego z 12 programów osobowości i znormalizowane względem najwyższego z nich. Wynik normalizacji zapisywany jest w tabeli tblStatsPercNew (3.9).

W tym momencie rozpoczyna się właściwe rozpoznawanie. Jak już wspominano w rozdziale 3, odbywa się ono dwutorowo, przy użyciu metod frekwencyjnych oraz porównania programów osobowości.

Matematyczny opis algorytmu wyliczania miary podobieństwa między autorem znanym, a rozpoznawanym podano w rozdziale 3.1.3, w równaniach 3.1 - 3.6.

Wszystkie te operacje dokonywane są poprzez joinowanie tabel należących do rozpoznawanego autora oraz każdego kolejnego autora z bazy, po czym na tak złączonych tabelach wykonywane są operacje matematyczne - sumowania, zliczania wierszy, itp.

Po otrzymaniu wszystkich pięciu wskaźników są one sumowane i odejmowane od 100, aby uzyskać rosnącą miarę podobieństwa tekstów.

Wyliczony rezultat Δ obrazuje zgodność pomiędzy porównywanymi tekstami. Im jest on wyższy, bliższy 100, tym większe prawdopodobieństwo, iż sprawdzana osoba jest autorem rozpoznawanego tekstu. W skrajnym przypadku, jeżeli do rozpoznawania weźmiemy tekst identyczny z istniejącym w bazie, wyznaczona Δ wyniesie dokładnie 100, co oznacza 100-procentową pewność. W praktyce, nawet jeśli weźmiemy dwa różne teksty napisane przez tę samą osobę, współczynnik Δ nigdy nie osiągnie 100.

Opisany powyżej algorytm postępowania przeprowadzany jest dla wszystkich autorów znajdujących się w bazie. W zależności od ilości autorów, wielkości próbek tekstów znajdujących się w bazie oraz rozpoznawanego tekstu, cały proces może trwać dosyć długo. Gdy się zakończy, zwracana jest lista potencjalnych autorów posortowana według malejącego współczynnika Δ. Dostępne są także statystyki porównawcze dla każdej pary autorów, w postaci tabeli lub wykresu porównawczego 4.9.

Ze względu na złożoność operacji i długość obliczeń przy dużej ilości danych w bazie został opracowany plan optymalizacji aplikacji poprzez pozbycie się już na którymś z wstępnych etapów tych słów i zwrotów, które występują w całej wypowiedzi tylko raz. Niosą one tym samym stosunkowo najmniej informacji o ulubionych słowach autora. Mogą natomiast zawierać istotne informacje o osobowości autora (nawet występując jednokrotnie). Dlatego najlepiej taką optymalizację będzie przeprowadzić na etapie po rozpoznawaniu osobowości, usuwając z bazy nadmiarowe wyrazy. O efektach takiej optymalizacji będzie mowa w rozdziale 4, przy okazji testów aplikacji.

4. Interfejs użytkownika i testy aplikacji

4.1. Opis interfejsu użytkownika

Interfejs webowy aplikacji został przygotowany przede wszystkim z myślą o funkcjonalności, dlatego jest dość prosty, jednak przyjemny dla oka, utrzymany w niejarzących zbytnio barwach.

Po wpisaniu w przeglądarce adresu aplikacji `http://www.sarwo.ugu.pl` użytkownika wita strona powitalna (rys. 4.1).

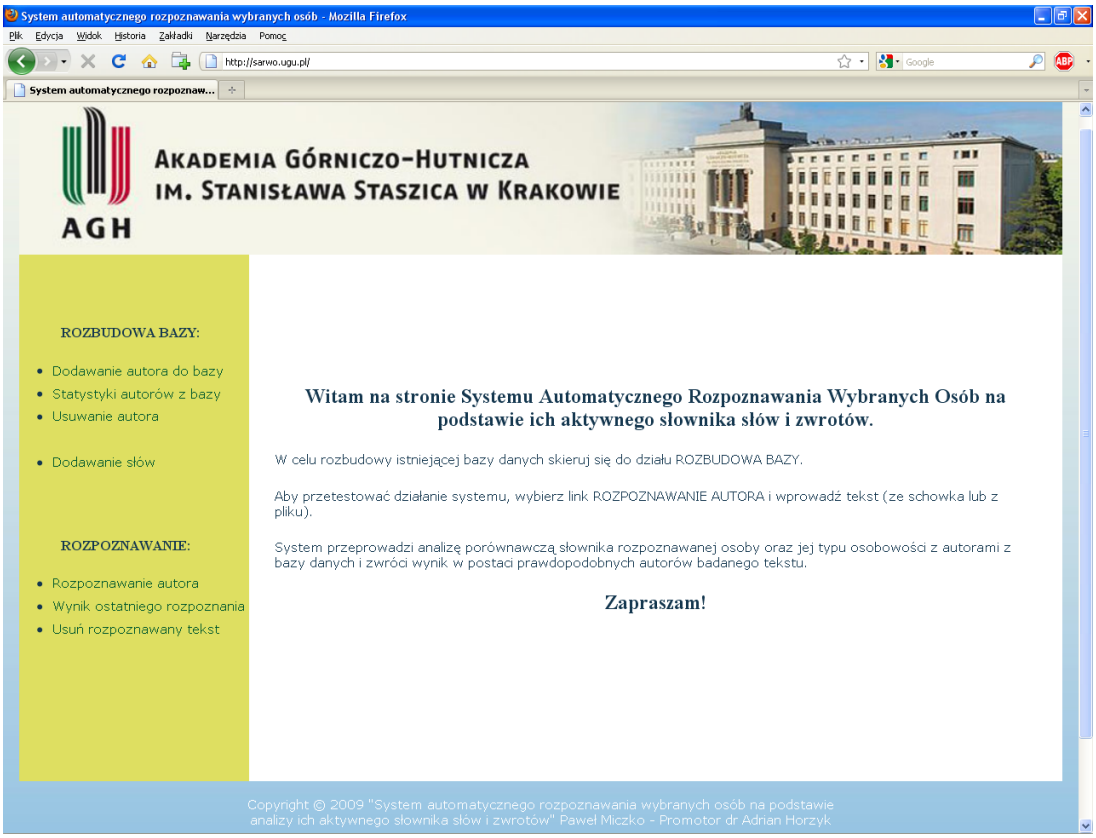
Layout utrzymany jest w klasycznej konwencji nagłówek - menu - treść - stopka.

Menu nawigacyjne znajduje się po lewej stronie. Zawiera ono odnośniki do wszystkich dostępnych modułów programu. Zostało ono podzielone na dwie części:

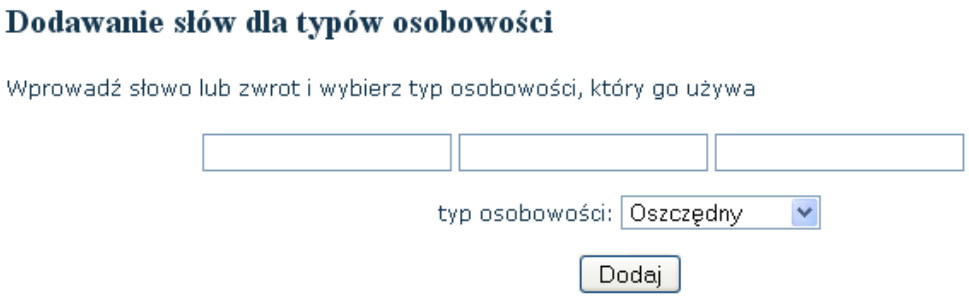
- Rozbudowa bazy danych
- Rozpoznawanie

Funkcjonalność modułu *Rozbudowa bazy danych*:

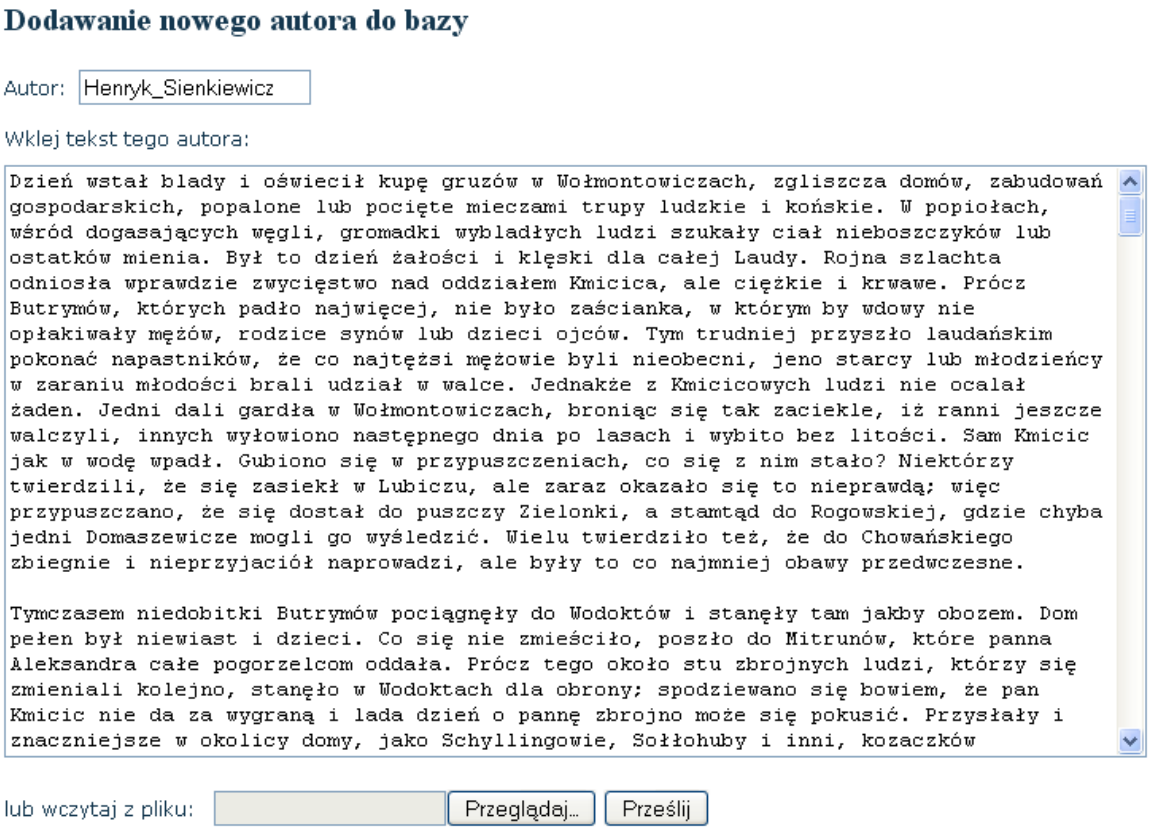
- Dodawanie słów - czyli rozbudowa bazy słów i zwrotów dla poszczególnych typów osobowości. Stworzona baza danych zawiera ok. 2400 słów i kilkaset zwrotów, na pewno jednak jej rozbudowa przyczyni się do lepszego funkcjonowania tej aplikacji, jak również może mieć znaczenie w szeregu innych projektów. Dlatego też strona daje możliwość rozbudowy istniejącej bazy. Formularz przedstawiony na rysunku 4.2 jest uniwersalny. Wypełnienie jedynie pierwszego pola spowoduje dodanie do bazy pojedynczego słowa, wypełnienie dwóch pól - doda zwrot dwuwyrazowy, a wszystkich trzech - trójwyrazowy. Z listy natomiast wybiera się jeden z dwunastu możliwych typów znajdujących się w bazie.
- Dodawanie autora - czyli wprowadzanie nowego autora do bazy danych w postaci jego tekstu lub zbioru tekstów. Formularz przedstawiony na rysunku 4.3 umożliwia wpisanie



Rysunek 4.1: Strona główna serwisu



Rysunek 4.2: Moduł dodawanie słów



Rysunek 4.3: Moduł dodawanie autora

bądź wklejenie tekstu albo jego wczytanie z pliku tekstowego. Rezultatem którejkolwiek z tych operacji jest dodanie tektu do bazy, natomiast dodającemu wyświetlają się statystyki słów (20 najpopularniejszych), zwrotów (po 10 najpopularniejszych) i typów osobowości dla dodanego autora (rys. 4.4). Czas trwania operacji wczytywania tekstu jest różny, zależnie od jego objętości.

W ramach statystyk dostępny jest również wykres radarowy programów osobowości autora. Na każdej z dwunastu osi odłożona jest wartość dla odpowiedniego programu. Punkty połączone są zieloną linią, a wyznaczony obszar oznaczony jest zielonym tłem (rys. 4.5).

- Statystyki autorów - zbiorcze statystyki programów osobowości dla autorów z bazy przedstawione w postaci tabeli (rys. 4.6). Dostępne są również linki do statystyk każdego autora z osobna (jak na rys. 4.4, 4.5)

Autor: Henryk_Sienkiewicz

typ	słów	zwroty 2-wyr	zwroty 3-wyr	suma	intensywność programu
Dominujący	103	0	0	103	73%
Maksymalista	93	0	0	93	65%
Inspirujący	13	0	0	13	9%
Odkrywczy	39	0	0	39	27%
Weryfikujący	28	0	0	28	20%
Systematyczny	39	0	0	39	27%
Asekuracyjny	91	0	0	91	64%
Oszczędny	0	0	0	0	0%
Harmonijny	77	0	0	77	54%
Empatyczny	130	12	0	142	100%
Zadaniowy	9	3	0	12	8%
Równoważący	43	0	0	43	30%

Najczęściej używane słowa i zwroty

Pojedyncze wyrazy

słowo	ilość
nie	108
że	58
ale	40
tak	25
było	24
go	22
ich	19
tym	17
mu	16
jest	16
gdy	16
tylko	16
był	16
jeszcze	16
ze	15
też	15
on	14
im	14
mnie	14
więc	13

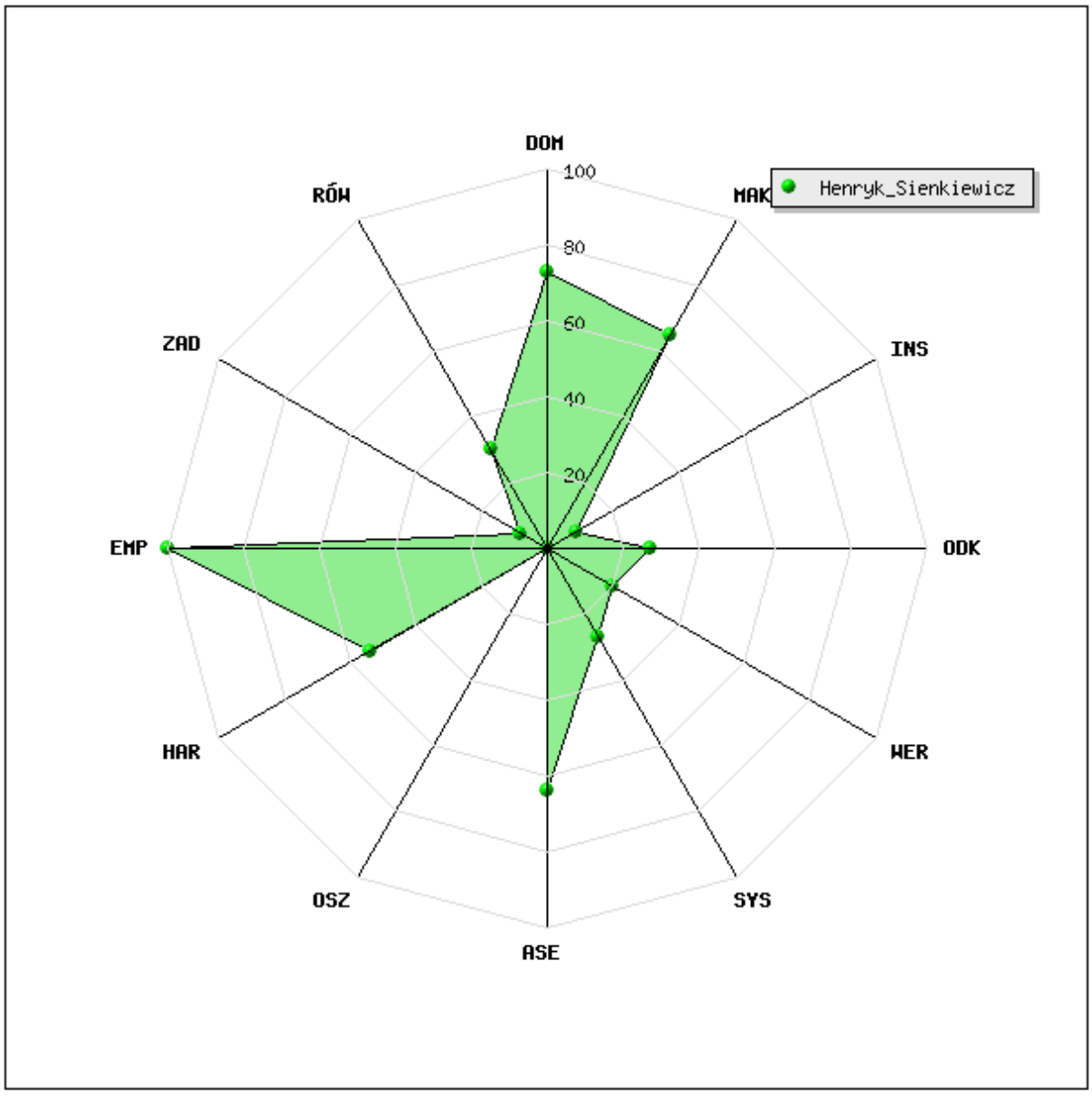
Zwroty 2-wyrazowe

zwrot	ilość
i w	10
i na	9
się w	9
panna aleksandra	6
nie było	6
się z	6
rzekł	5
się to	5
nie tylko	5
się nie	5

Zwroty 3-wyrazowe

zwrot	ilość
z nimi wojować	3
i w innych	2
w miarę jak	2
się cały dwór	2
się z nim	2
w tej chwili	2
święty benedykt z	2
przy wielkim ołtarzu	2
w tym królestwie	2
nie mógł i	2

Rysunek 4.4: Statystyki przykładowego autora w formie tabelarycznej



Rysunek 4.5: Przykładowy wykres osobowości autora

Autorzy i ich programy osobowości

Osoba	DOM	MAK	INS	ODK	WER	SYS	ASE	OSZ	HAR	EMP	ZAD	RÓW	Statystyki	
Henryk_Sienkiewicz	73	65	9	27	20	27	64	0	54	100	8	30	tabele	wykres
Marek_Hłasko	100	13	2	45	13	49	36	0	47	55	10	3	tabele	wykres
Maria_Konopnicka	100	24	8	44	16	22	38	0	57	83	4	5	tabele	wykres
Stephen_King	100	65	16	42	19	65	55	0	72	52	7	6	tabele	wykres
Władysław_Reymont	33	22	12	19	19	46	46	0	77	100	3	12	tabele	wykres

Rysunek 4.6: Statystyki autorów

Analiza typów osobowości dla wprowadzonego tekstu

typ	słów	zwroty 2-wyr	zwroty 3-wyr	suma	intensywność programu
Dominujący	134	0	0	134	94%
Maksymalista	80	3	0	83	58%
Inspirujący	8	0	0	8	6%
Odkrywczy	53	0	0	53	37%
Weryfikujący	24	6	0	30	21%
Systematyczny	49	9	0	58	41%
Asekuracyjny	107	0	0	107	75%
Oszczędny	0	0	0	0	0%
Harmonijny	82	0	0	82	57%
Empatyczny	127	6	10	143	100%
Zadaniowy	10	3	0	13	9%
Równoważący	14	0	0	14	10%

Wynik rozpoznawania wśród autorów z bazy

Autor	Programy osobowości	POWT	Słowa	Zwroty 2-wyr	Zwroty 3-wyr	Suma	% dopasowania	Porównanie	
Henryk_Sienkiewicz	36	73	72	94	99	374	62.62%	tabele	wykres
Stephen_King	57	75	81	96	100	409	59.07%	tabele	wykres
Marek_Hłasko	72	62	81	95	100	426	57.36%	tabele	wykres
Maria_Konopnicka	55	99	82	95	100	432	56.84%	tabele	wykres
Władysław_Reymont	170	107	81	96	100	459	54.13%	tabele	wykres

Co chcesz teraz zrobić?

Usunąć rozpoznawany tekst

Zapisać do bazy jako autor: Zapisz

Rysunek 4.7: Wynik rozpoznawania - widok ogólny

Funkcjonalność modułu *Rozpoznawanie*:

- Rozpoznawanie autora - pierwszym jego krokiem jest wpisanie lub zaimportowanie z pliku rozpoznawanego tekstu. Formularz jest tutaj analogiczny jak ten z rys. 4.3 - brakuje tylko rzecz jasna pola Autor, gdyż ten jest nieznany. Następnie ma miejsce wczytanie i przetworzenie danego tekstu, po czym wyniki zostają porównane z każdym tekstem znajdującym się w bazie z osobna. Rozpoznawanie przebiega według algorytmu opisanego w rozdziale 3.1.3.

Statystyki porównawcze autorów

rozpoznawany	typ	Henryk_Sienkiewicz
79%	Dominujący	67%
51%	Maksymalista	60%
5%	Inspirujący	8%
31%	Odkrywczy	25%
21%	Weryfikujący	18%
40%	Systematyczny	25%
63%	Asekuracyjny	59%
0%	Oszczędny	0%
49%	Harmonijny	50%
100%	Empatyczny	100%
9%	Zadaniowy	10%
8%	Równoważący	28%

Najczęściej używane słowa i zwroty

Pojedyncze wyrazy

rozpoznawany		Henryk_Sienkiewicz	
słowo	ilość	słowo	ilość
nie	120	nie	108
że	67	że	58
ale	47	ale	40
tak	35	tak	25
jak	30	było	24
maćko	29	go	22
go	28	ich	19
zbyszko	27	tym	17
rzekł	26	mu	16
mu	26	jest	16
powała	24	gdy	16
był	21	tylko	16
by	21	był	16
ja	21	jeszcze	16
też	20	ze	15
który	19	też	15
księżna	18	on	14
przez	18	im	14
zaś	18	mnie	14
było	16	więc	13

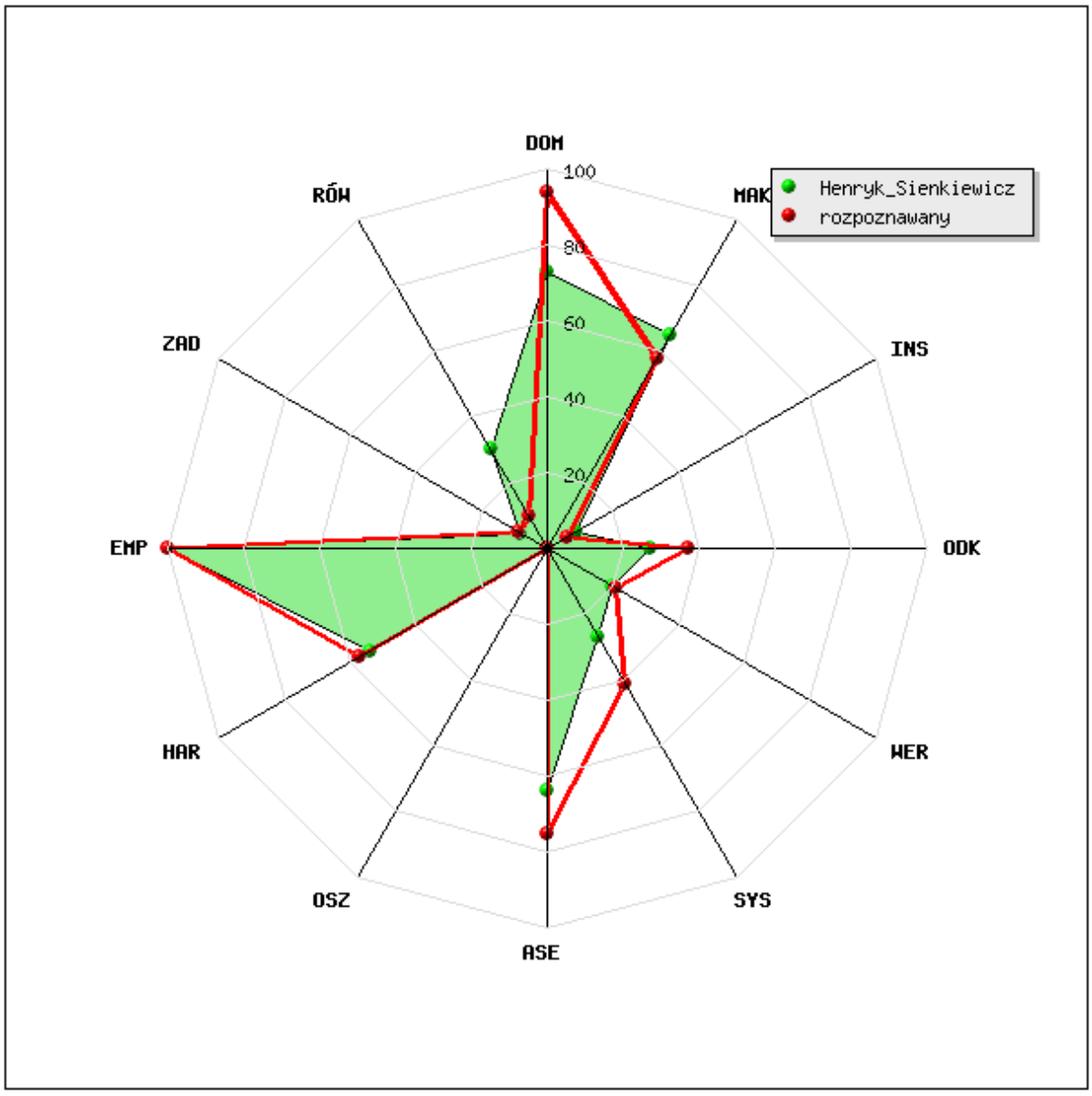
Zwroty 2-wyrazowe

rozpoznawany		Henryk_Sienkiewicz	
zwrot	ilość	zwrot	ilość
się do	15	i w	10
rzekł	12	i na	9
z długolasu	10	się w	9
z taczewa	9	panna aleksandra	6
się na	9	nie było	6
się w	9	się z	6
ozwał się	8	rzekł	5
mikołaj z	7	się to	5
na to	7	nie tylko	5
i rzekł	6	się nie	5

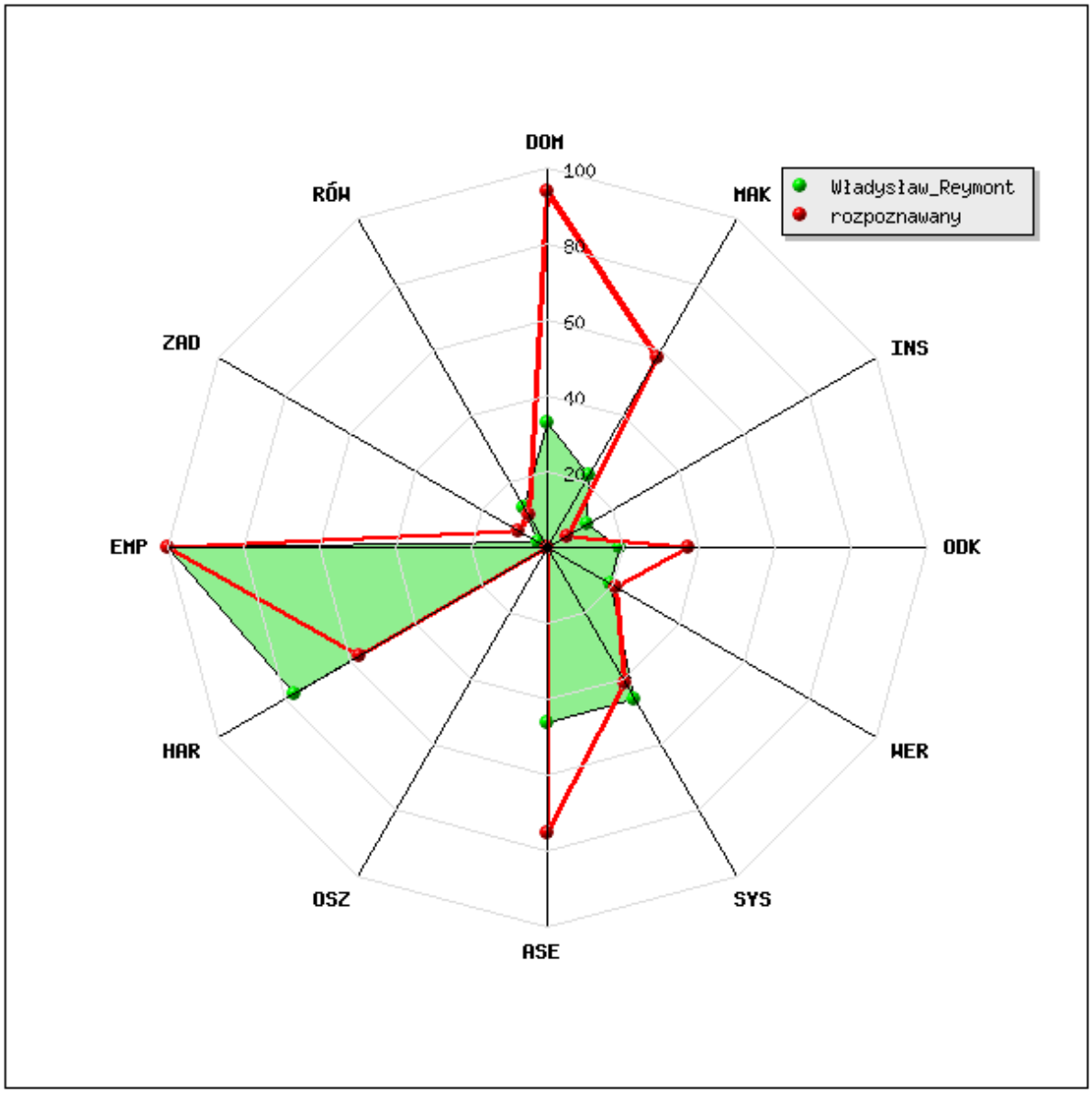
Zwroty 3-wyrazowe

rozpoznawany		Henryk_Sienkiewicz	
zwrot	ilość	zwrot	ilość
mikołaj z długolasu	7	z nimi wojować	3
powała z taczewa	4	i w innych	2
maćka i zbyszka	4	w miarę jak	2
i rzekł	4	się cały dwór	2
mikołaja z długolasu	3	się z nim	2
zwrócił się do	3	w tej chwili	2
z długolasu który	3	święty benedykt z	2
maćko i zbyszko	3	przy wielkim ołtarzu	2
dowiedziawszy się o	2	w tym królestwie	2
z wielką chwałą	2	nie mógł i	2

Rysunek 4.8: Wynik rozpoznawania - porównanie dwóch autorów



Rysunek 4.9: Przykładowy wykres porównawczy - ten sam autor



Rysunek 4.10: Przykładowy wykres porównawczy - różni autorzy

W wyniku wyświetlane są statystyki dla rozpoznawanego tekstu oraz wynik rozpoznania - tabela zawierająca autorów posortowanych od najbardziej do najmniej podobnego pod względem używanego słownictwa i programu osobowości (rys. 4.7). Widoczne jest wszystkie pięć kryteriów tej oceny, wraz z wartościami poszczególnych składników, ich suma oraz sumaryczny "procent" podobieństwa. Dostępne są statystyki porównawcze autora rozpoznawanego oraz wybranego autora z bazy (rys. 4.8). Można także zobaczyć wykres porównawczy dla tych dwóch autorów. Wykres autora z bazy oznaczony jest tak jak poprzednio - zielona linia i zielone tło, podczas gdy wykres autora rozpoznawanego określa czerwona linia, bez żadnego tła (rys. 4.9, 4.10). Wszystkie wykresy generowane są z pomocą biblioteki jpgraph [15]

- Wynik ostatniego rozpoznania - rezultat rozpoznawania jest przechowywany w bazie i jest dostępny przez cały czas aż do momentu wykonania jednej z następujących operacji:
  - usunięcie rozpoznawanego tekstu - tekst kasowany jest z bazy i można wprowadzić kolejny tekst do rozpoznawania
  - zapisanie rozpoznawanego tekstu do bazy - jeśli chcemy rozbudować bazę danych, np. dany autor nie jest żadną z osób znajdujących się w bazie, można go do niej dodać pod dowolną nazwą. Rozpoznawany tekst wraz ze wszystkimi jego statystykami zostanie wówczas przeniesiony do bazy danych tekstów, a zwolnione w ten sposób miejsce dostępne jest do rozpoznawania kolejnego tekstu.

W przedstawionym na rysunkach 4.3 - 4.10 przykładzie rozpoznawaniu poddano fragment *Krzyżaków* Henryka Sienkiewicza, natomiast w bazie znajdowały się fragmenty różnych tekstów M. Konopnickiej, W. Reymonta, M. Hłaski, S. Kinga i właśnie Sienkiewicza - dokładniej były to fragmenty *Potopu* i *Krzyżaków* (inny rozdział niż rozpoznawany). Krótki rzut oka na rys. 4.7, 4.9 i 4.10 wręcz narzuca pytanie na temat skuteczności działania aplikacji. Analiza tego zagadnienia będzie więc przedmiotem kolejnego rozdziału.

4.2. Testy aplikacji

Do testowania aplikacji został pozyskany zbiór dzieł polskich twórców (pisarzy i nowelistów) z różnych epok literackich, udostępniony publicznie w Internecie pod adresem [16]. Do bazy wprowadzone zostały fragmenty utworów pisarzy takich jak H. Sienkiewicz, W. Reymont, B. Prus, A. Mickiewicz, J. I. Kraszewski, M. Konopnicka, E. Orzeszkowa, S. Żeromski, T. Dołęga-Mostowicz, B.Schulz, M. Hłasko. Do testowania były używane inne teksty tych autorów, w szczególności inne fragmenty tych samych tekstów.

Poniżej przedstawiono kilka przykładów działania systemu dla tego zbioru autorów.

- Rozpoznawany tekst - fragment *Chłopów* W. Reymonta

  Tekst został rozpoznany poprawnie - co widać na rys. 4.11. "Przewaga punktowa" nad kolejnym tekstem jest stosunkowo duża; wynika to przede wszystkim z używania przez autora charakterystycznej gwary ludowej (rys. 4.12), która nie występuje w takim nasyceniu w żadnym innym tekście bazy. Również profile osobowościowe obydwu tekstów są bardzo podobne (rys. 4.13).

Wynik rozpoznawania wśród autorów z bazy

Autor	Programy osobowości	POWT	Słowa	Zwroty 2-wyr	Zwroty 3-wyr	Suma	% dopasowania	Porównanie	
Chłopi	27	51	48	67	93	285	71.45%	tabele	wykres
Sienkiewicz	55	70	63	83	98	369	63.09%	tabele	wykres
Quo_Vadis	50	78	64	82	97	371	62.88%	tabele	wykres
Szkice_węglem	50	104	64	82	98	398	60.22%	tabele	wykres
Prus	72	105	67	82	98	423	57.67%	tabele	wykres
Mickiewicz	68	104	66	89	99	426	57.41%	tabele	wykres
DołęgaMostowicz	67	115	64	82	98	426	57.41%	tabele	wykres
Konopnicka	72	115	65	80	98	430	56.96%	tabele	wykres
Hłasko	79	118	63	80	98	437	56.33%	tabele	wykres
Kraszewski	93	119	60	80	98	449	55.08%	tabele	wykres
Orzeszkowa	90	159	67	86	99	501	49.87%	tabele	wykres
Krasicki	100	155	74	87	99	515	48.55%	tabele	wykres
Przedwiośnie	102	167	70	87	99	524	47.59%	tabele	wykres
Schulz	120	151	80	94	100	545	45.51%	tabele	wykres

Rysunek 4.11: Wynik rozpoznawania dla *Chłopów* B.Prusa

Najczęściej używane słowa i zwroty

Pojedyncze wyrazy

rozpoznawany		Chłopi	
słowo	ilość	słowo	ilość
nie	137	nie	133
że	110	że	103
już	52	już	46
ale	38	tak	38
tak	34	jeno	38
jeszcze	29	kiej	34
aż	28	ale	33
antek	28	jej	28
zaś	26	było	25
ino	25	zaś	22
jeno	24	jak	19
kiej	23	te	18
ze	21	jeszcze	17
go	21	ten	16
było	20	nawet	16
mu	20	jakby	16
jak	19	był	16
też	19	ją	16
był	18	ano	15
przy	18	jako	15

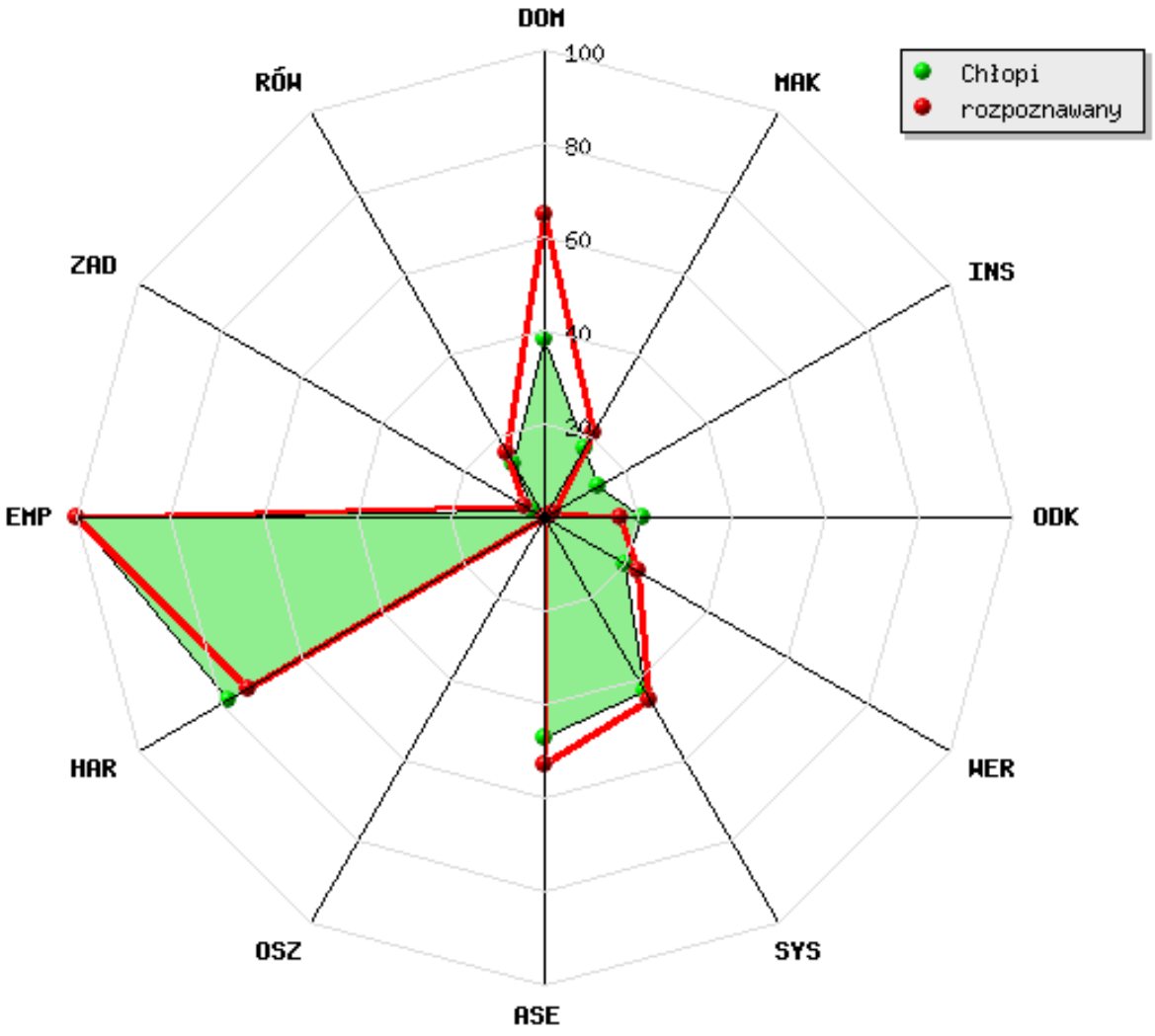
Zwroty 2-wyrazowe

rozpoznawany		Chłopi	
zwrot	ilość	zwrot	ilość
i nie	12	to i	16
i w	12	się jej	13
to i	11	i nie	11
się do	11	i tak	10
że to	11	że i	8
aż się	10	kiej te	8
się z	10	się już	8
się nie	9	się do	8
juści że	9	się w	8
się w	8	że już	7

Zwroty 3-wyrazowe

rozpoznawany		Chłopi	
zwrot	ilość	zwrot	ilość
z całej mocy	3	że i nie	4
w ten mig	3	i bez to	4
bartek z tartaku	3	się w sobie	3
się do niej	3	raz po raz	3
że i nie	3	to juści że	3
jest tylko jedna	2	a i bez	3
tylko jedna rada	2	nie dziwota że	3
i nie wypowiedzieć	2	i nie dziwota	3
i tany szły	2	że się już	3
a za nim	2	jakby z musu	2

Rysunek 4.12: Porównanie rozpoznawanego fragmentu *Chłopów* ze znajdującym się w bazie



Rysunek 4.13: Wykres typów osobowości dla obu tekstów

- Rozpoznawany tekst - fragment *Lalki* B.Prusa

  Kolejny tekst rozpoznany poprawnie (rys. 4.14). Również i w tym wypadku w bazie znajdował się inny fragment tego samego utworu, co mogło ułatwić rozpoznanie. W następnym przykładzie przeanalizowany zostanie zatem inny tekst Prusa - *Faraon*.

Wynik rozpoznawania wśród autorów z bazy

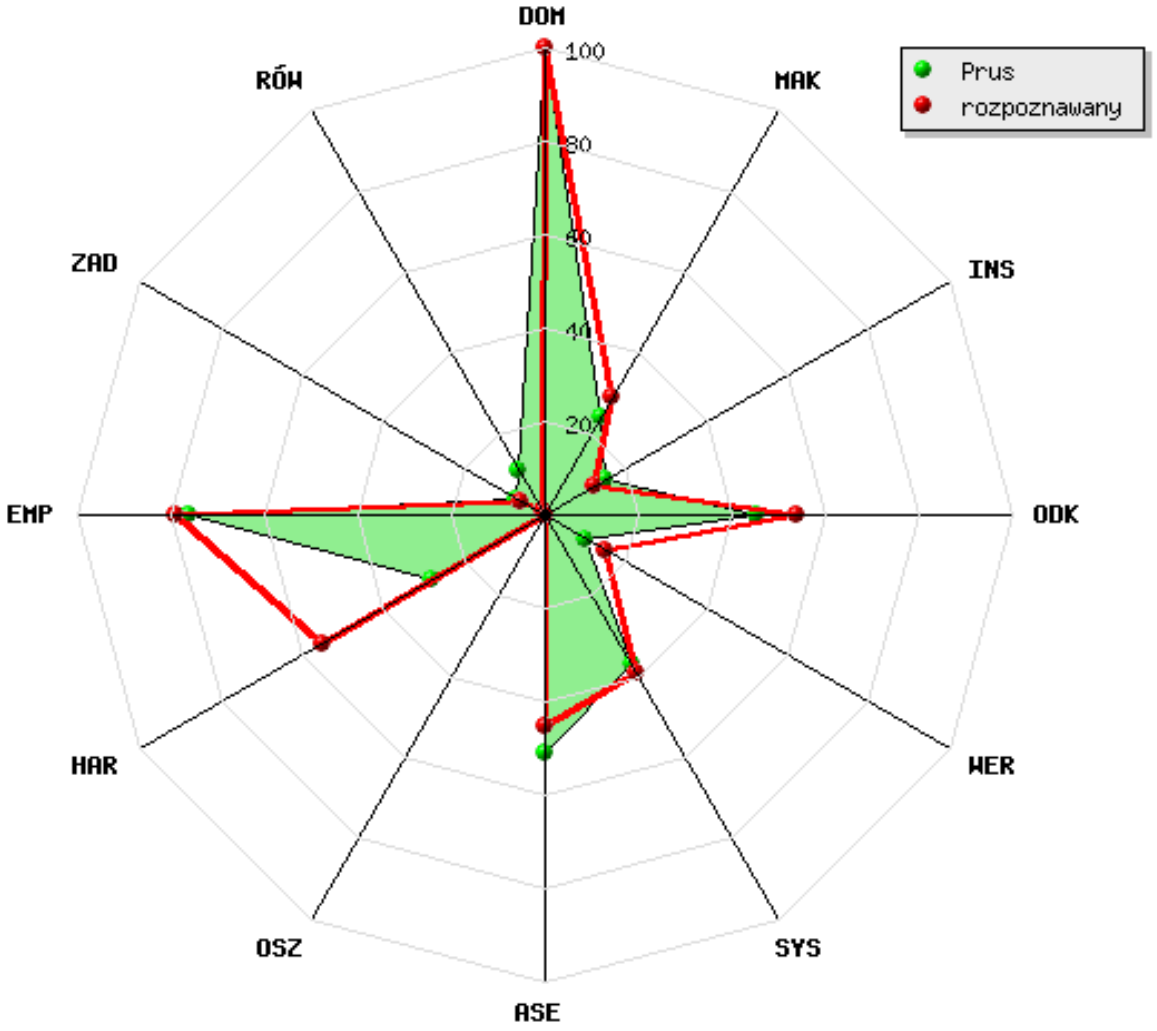
Autor	Programy osobowości	POWT	Słowa	Zwroty 2-wyr	Zwroty 3-wyr	Suma	% dopasowania	Porównanie	
Prus	28	78	51	67	90	314	68.65%	tabele	wykres
DołęgaMostowicz	44	76	55	73	95	343	65.74%	tabele	wykres
Hłasko	40	84	57	72	96	349	65.09%	tabele	wykres
Konopnicka	29	95	65	80	98	367	63.28%	tabele	wykres
Sienkiewicz	73	90	61	81	98	401	59.85%	tabele	wykres
Mickiewicz	59	96	64	86	99	404	59.59%	tabele	wykres
Szkice_węglem	78	89	63	82	97	408	59.23%	tabele	wykres
Kraszewski	67	118	57	77	97	416	58.42%	tabele	wykres
Quo_Vadis	81	112	58	76	96	422	57.83%	tabele	wykres
Chłopi	71	112	70	83	98	434	56.59%	tabele	wykres
Orzeszkowa	106	97	61	84	99	447	55.35%	tabele	wykres
Krasicki	67	123	71	88	100	448	55.19%	tabele	wykres
Schulz	57	139	77	91	99	463	53.74%	tabele	wykres
Przedwiośnie	102	144	64	84	99	493	50.75%	tabele	wykres

Rysunek 4.14: Wynik rozpoznawania dla *Lalki* B.Prusa

Statystyki porównawcze autorów

rozpoznawany	typ	Prus
100%	Dominujący	100%
29%	Maksymalista	24%
13%	Inspirujący	19%
59%	Odkrywczy	42%
15%	Weryfikujący	9%
45%	Systematyczny	47%
45%	Asekuracyjny	47%
0%	Oszczędny	0%
56%	Harmonijny	25%
82%	Empatyczny	71%
6%	Zadaniowy	8%
1%	Równoważący	10%

Rysunek 4.15: Porównanie rozpoznawanego fragmentu *Lalki* ze znajdującym się w bazie



Rysunek 4.16: Wykres typów osobowości dla obu tekstów

- Rozpoznawany tekst - fragment *Faraona* B.Prusa

  Tym razem automatyczna analiza wskazała jako najbardziej podobne dwa teksty H. Sienkiewicza. Prawdziwy autor - B. Prus znalazł się dopiero na trzecim miejscu (rys. 4.17). Warto jednak zauważyć, iż *Faraon* jako tekst historyczny ma wspólną tematykę z *Quo Vadis* i innymi tekstami Sienkiewicza (*Potop*, *Krzyżacy*) niż z bardziej współczesną *Lalką*. Analiza programów osobowości w obydwu tekstach wypadła podobnie (rys. 4.18, 4.19).

Wynik rozpoznawania wśród autorów z bazy

Autor	Programy osobowości	POWT	Słowa	Zwroty 2-wyr	Zwroty 3-wyr	Suma	% dopasowania	Porównanie	
Sienkiewicz	48	102	54	81	99	385	61.47%	tabele	wykres
Quo_Vadis	75	82	54	80	96	387	61.26%	tabele	wykres
Prus	74	91	53	80	99	397	60.31%	tabele	wykres
DołęgaMostowicz	62	100	56	81	97	397	60.30%	tabele	wykres
Mickiewicz	66	87	59	91	99	401	59.87%	tabele	wykres
Konopnicka	66	97	63	82	100	407	59.26%	tabele	wykres
Hłasko	71	106	56	81	99	413	58.75%	tabele	wykres
Kraszewski	72	109	53	81	99	414	58.56%	tabele	wykres
Szkice_węglem	69	120	59	82	97	426	57.43%	tabele	wykres
Orzeszkowa	87	102	59	85	98	433	56.68%	tabele	wykres
Chłopi	72	120	66	84	98	438	56.20%	tabele	wykres
Schulz	84	100	76	92	99	452	54.85%	tabele	wykres
Krasicki	87	126	69	87	99	467	53.26%	tabele	wykres
Przedwiośnie	92	144	61	84	98	479	52.15%	tabele	wykres

Rysunek 4.17: Wynik rozpoznawania dla *Faraona* B.Prusa

Statystyki porównawcze autorów

rozpoznawany	typ	Prus
78%	Dominujący	100%
75%	Maksymalista	24%
11%	Inspirujący	19%
22%	Odkrywczy	42%
10%	Weryfikujący	9%
49%	Systematyczny	47%
29%	Asekuracyjny	47%
0%	Oszczędny	0%
49%	Harmonijny	25%
100%	Empatyczny	71%
14%	Zadaniowy	8%
7%	Równoważący	10%

Rysunek 4.18: Porównanie rozpoznawanego fragmentu *Faraona* ze znajdującym się w bazie fragmentem *Lalki* tego autora

Statystyki porównawcze autorów

rozpoznawany	typ	Sienkiewicz
78%	Dominujący	68%
75%	Maksymalista	60%
11%	Inspirujący	8%
22%	Odkrywczy	25%
10%	Weryfikujący	18%
49%	Systematyczny	25%
29%	Asekuracyjny	59%
0%	Oszczędny	0%
49%	Harmonijny	50%
100%	Empatyczny	100%
14%	Zadaniowy	10%
7%	Równoważący	28%

Rysunek 4.19: Porównanie rozpoznawanego fragmentu *Faraona* z tekstami H. Sienkiewicza

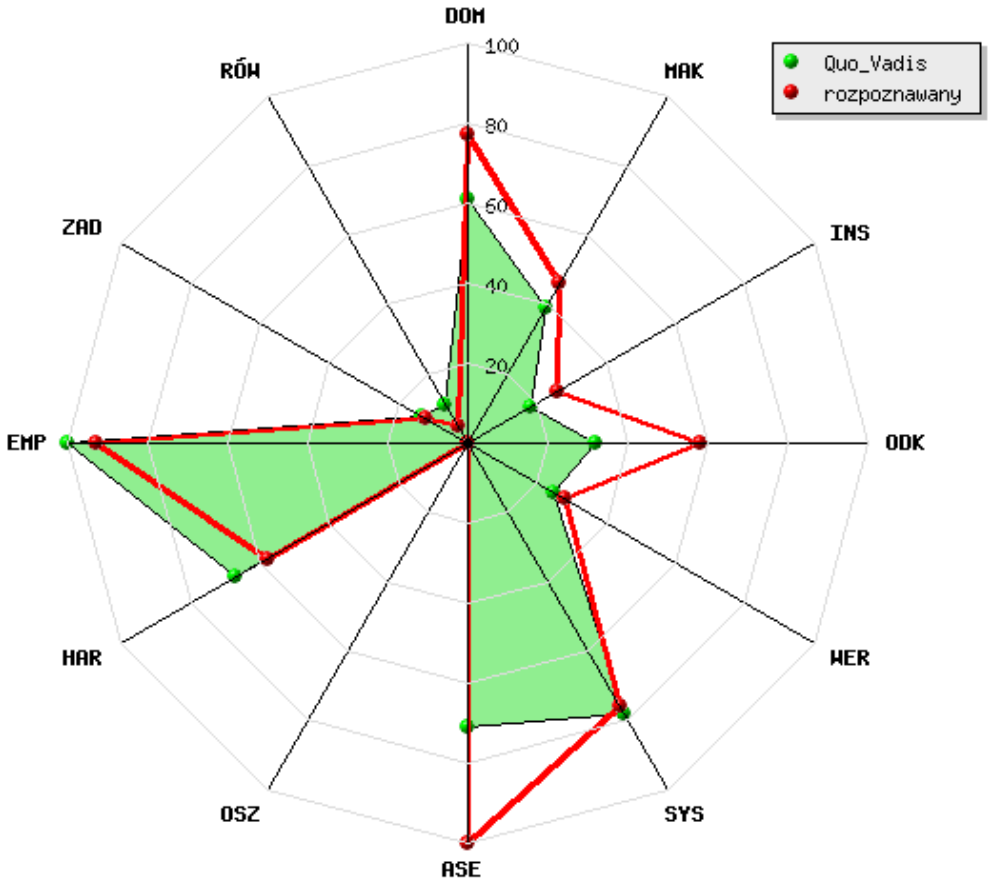
- Rozpoznawany tekst - fragment *Quo Vadis* H. Sienkiewicza

  O niebagatelnym znaczeniu jakie ma w całym rozpoznawaniu tematyka i kontekst powstawania danego tekstu, świadczy poniższy, ostatni już przykład. Fragment *Quo Vadis* został idealnie rozpoznany w pierwszym rzędzie jako *Quo Vadis*, a dopiero w dalszej kolejności jako tekst Henryka Sienkiewicza 4.20. Również programy osobowości dla *Quo Vadis* i innych tekstów Sienkiewicza dosyć wyraźnie się różnią (rys. 4.21, 4.22). Przykład ten świadczy o tym, że wielu autorów stylizuje używany przez siebie język w zależności od tego, o czym piszą, np. w powieści o czasach starożytnych używany będzie inny styl niż w powieści o polskich rycerzach epoki średniowiecza. Również osobowość kreowanych postaci może być różna w różnych tekstach literackich. Tego typu zabiegi czynią rozpoznawanie trudniejszym niż np. w wypadku neutralnego tekstu, w którym autor wypowiada się jako on sam, używając języka, którym posługuje się naturalnie na co dzień.

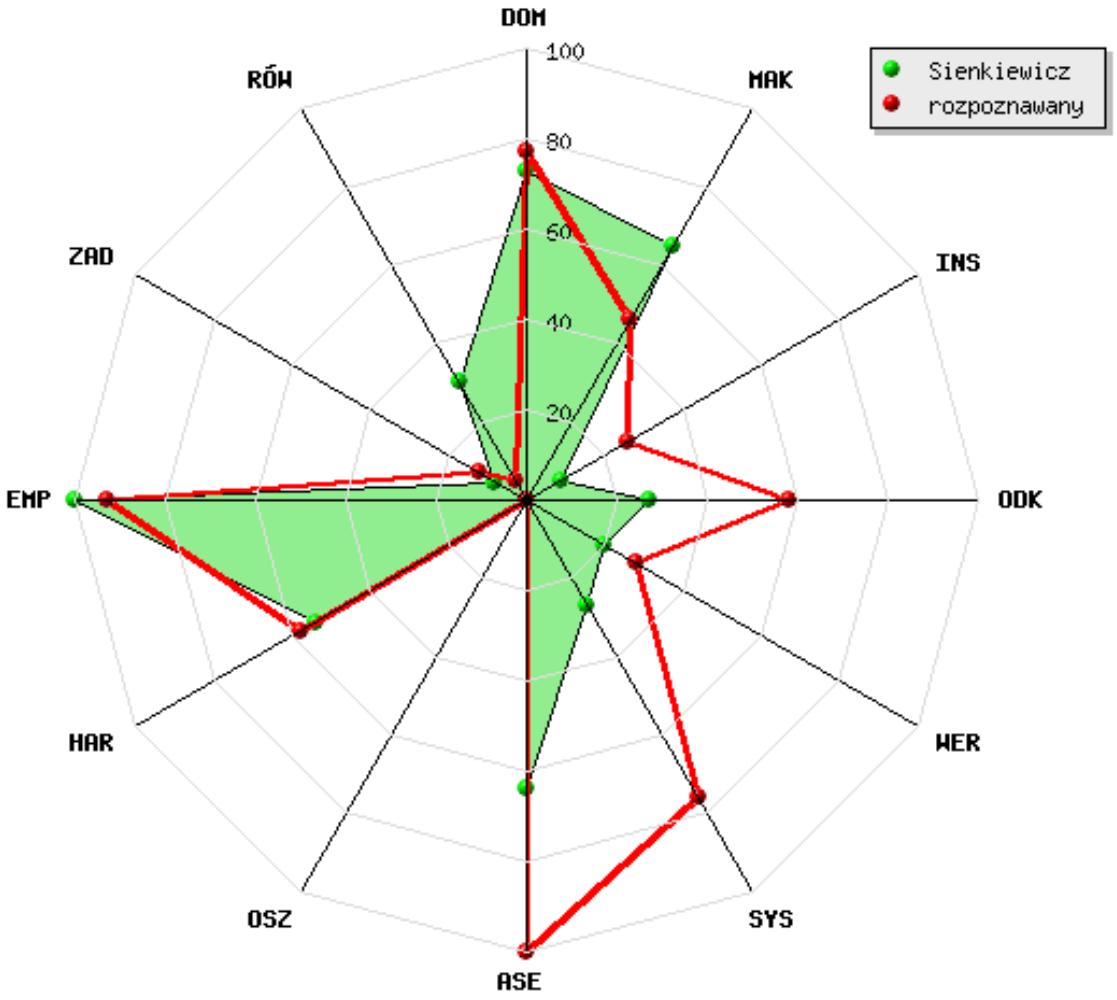
Wynik rozpoznawania wśród autorów z bazy

Autor	Programy osobowości	POWT	Słowa	Zwroty 2-wyr	Zwroty 3-wyr	Suma	% dopasowania	Porównanie	
Quo_Vadis	46	76	31	58	93	305	69.53%	tabele	wykres
Sienkiewicz	82	64	48	72	96	363	63.68%	tabele	wykres
DołęgaMostowicz	74	86	53	76	97	386	61.45%	tabele	wykres
Kraszewski	68	91	54	80	98	391	60.91%	tabele	wykres
Mickiewicz	52	93	61	86	99	392	60.80%	tabele	wykres
Szkice_węglem	54	112	55	78	98	396	60.38%	tabele	wykres
Prus	93	96	55	76	96	416	58.38%	tabele	wykres
Chłopi	102	90	63	80	99	434	56.57%	tabele	wykres
Hłasko	103	112	55	75	98	444	55.55%	tabele	wykres
Konopnicka	89	129	61	81	98	458	54.24%	tabele	wykres
Przedwiośnie	98	133	57	78	99	465	53.50%	tabele	wykres
Krasicki	140	112	64	83	99	498	50.21%	tabele	wykres
Orzeszkowa	110	152	59	82	99	501	49.87%	tabele	wykres
Schulz	96	166	74	91	99	526	47.41%	tabele	wykres

Rysunek 4.20: Wynik rozpoznawania dla *Quo Vadis* H. Sienkiewicza



Rysunek 4.21: Porównanie rozpoznawanego fragmentu *Quo Vadis* ze znajdującym się w bazie



Rysunek 4.22: Porównanie rozpoznawanego fragmentu *Quo Vadis* z innymi tekstami Sienkiewicza z bazy

Łącznie przeprowadzono 17 rozpoznań dla różnych tekstów autorów z bazy. Prawidłowe rozpoznanie otrzymano w 13 z nich, co daje skuteczność na poziomie 76%. Błędnie zidentyfikowane zostały 4 teksty.

Ostatnim aspektem, nad którym należało się zastanowić, była optymalizacja aplikacji w wypadku analizy długich tekstów, sięgających tomów czy nawet całych powieści. W takim wypadku, przy analizie dużej ilości autorów rozmiar bazy rośnie, gdyż dodawane są do niej wszystkie słowa znajdujące się w tekście, nawet te występujące jednokrotnie, które nie dają wielkiej informacji o autorze. W takim przypadku nakład obliczeniowy rośnie, co przedłuża czas rozpoznawania. Zaproponowana optymalizacja polegała na usunięciu z bazy wszystkich słów i zwrotów pojawiających się jednokrotnie przed etapem rozpoznawania.

Usunięcie tych słów i fraz spowodowało przyspieszenie działania aplikacji. Zmniejszeniu uległ rozmiar tabel, które były ze sobą porównywane, przez co cała operacja odbywa się szybciej. Tracona przy tym ilość informacji okazała się, zgodnie z przewidywaniami akceptowalnie mała. Optymalizacja aplikacji umożliwiła wprowadzanie długich tekstów, zarówno do bazy danych, jak i do rozpoznawania i otrzymywanie wyników w możliwym do przyjęcia czasie.

5. Podsumowanie

5.1. Wnioski

Analiza wyników zaprezentowanych w rozdziale 4.2 prowadzi do konkluzji, iż zaproponowany algorytm działa poprawnie. Jakość rozpoznawania jest bardzo dobra w obrębie jednego dzieła literackiego, nieco słabsza, lecz wciąż dobra w przypadku różnych dzieł literackich.

Ogólnie można uznać, iż budowa aplikacji realizującej postawione w zadaniu cele zakończyła się pomyślnie. Stworzony został zatem unikalny system rozpoznawania wybranych osób na podstawie analizy ich aktywnego słownika słów i zwrotów. Analizowane słowa i frazy dwu- i trzywyrazowe okazały się wystarczającym materiałem, aby automatyczny system mógł dokonać identyfikacji autora dzieła literackiego. Aplikacja tworzy słowniki frekwencyjne dla dowolnych autorów, których tekstami dysponuje, a następnie wykorzystuje je do rozpoznania autorstwa innych tekstów wypowiadanych przez tych mówców lub pisarzy. Pozwala to na zidentyfikowanie dzieł literackich, jak również dowolnych innych tekstów napisanych przez jedną z wybranej grupy osób. Pomocniczo została stworzona baza prawie 3 tysięcy słów i zwrotów charakterystycznych dla 12 typów ludzkiej osobowości według typologii dr Adriana Horzyka. Połączenie tych dwóch metod pozwoliło osiągnąć zadowalające wyniki.

Zaimplementowany system może mieć szereg ciekawych zastosowań. Może posłużyć do identyfikacji nieznanych wcześniej tekstów literackich, odnajdywanych czasem przypadkowo przez badaczy przeszłości. Może być pomocny w analizie plagiatów, gdyż po wprowadzeniu dwóch tekstów, z których jeden jest częściowo lub całkowicie skopiowany z drugiego, system natychmiast to wykaże. Można wreszcie za pomocą tego systemu wysnuć wiele interesujących wniosków na temat osobowości autorów dzieł literackich, pisarzy, publicystów czy nawet ludzi z naszego otoczenia, o ile tylko dysponujemy odpowiednio obszerną próbką stworzonego przez nich tekstu. Poznając ich osobowość, poznajemy również ich samych. Możemy również sprawdzić, kto inny używa podobnego do nich stylu, a czyj styl jest całkowicie odmienny.

5.2. Możliwości rozbudowy

Po zakończonej realizacji projektu pozostaje wiele furtek - możliwości rozbudowy. System można by uczynić bardziej uniwersalnym, wprowadzając alternatywny sposób pozyskiwania tekstów - nie tylko dzieła literackie, ale także proste wypowiedzi dowolnych osób na codzienne tematy, na zasadzie chatbota, tak jak w pracy dyplomowej [12]. System taki prowadziłby rozmowę z zalogowanym użytkownikiem na wybrany przez niego temat, a udzielone odpowiedzi zapisywał w bazie, podobnie jak obecnie wprowadzane teksty literackie. Po zgromadzeniu odpowiedniej ilości wypowiedzi można by przeprowadzać rozpoznawanie, prowadząc rozmowę z jedną z tych osób na inny niż poprzednio temat i analizując jej dobór słów i zwrotów starać się ją zidentyfikować.

Inną możliwością rozbudowy byłoby niejako odwrócenie zasady działania programu w części osobowościowej. Na podstawie rozpoznanych typów osobowości u grupy autorów, przeprowadzana byłaby analiza najczęściej używanych słów i zwrotów przez osoby o danym typie osobowości i dodawanie do bazy tych, które jeszcze się tam nie znajdują, jako charakterystycznych dla osób o tym właśnie typie. Aplikacja o rozbudowanej w ten sposób bazie byłaby w stanie jeszcze optymalniej dokonywać rozpoznawania autorstwa tekstów.

A.Dodatek A

Spis zawartości płyty CD:

- pliki z kodami źródłowymi poszczególnych modułów aplikacji
- pliki SQL zawierający zrzut pustej bazy danych (bez wprowadzonych autorów)
- baza słów i zwrotów dla poszczególnych typów osobowości w formacie MS Excel
- treść pracy w postaci zbioru .pdf

Bibliografia

[1] http://pl.wikipedia.org/wiki/Web_2.0.

[2] Horzyk A., Tadeusiewicz R.: *Cechy osobowości użytkownika w systemach sztucznej inteligencji. Ich automatyczne rozpoznawanie, rozumienie i reagowanie na wynikające z nich potrzeby.* Rozdział w pracy zbiorowej Grzech A., Juszczyszyn K., Kwaśnicka H., Nogoc Thanh Nguyen (red.): *Inżyniera wiedzy i systemy ekspertowe.* Akademicka Oficyna Wydawnicza EXIT, Proc. of IWSE 2009 - REFERAT PLENARNY, Warszawa 2009, pp. 3-18.

[3] Horzyk, A., Magierski S., Miklaszewski G.: *An Intelligent Internet Shop-Assistant Recognizing a Customer Personality for Improving Man-Machine Interactions*, Kłopotek M.A. at all (Eds.), Proc. of IIS 2009, EXIT, ISBN 978-83-60434-59-8, 2009, pp. 13-26.

[4] Horzyk, A., Tadeusiewicz, R.: *A Psycholinguistic Model of Man-Machine Interactions Based on Needs of Human Personality, Man-Machine Interactions*, K.A. Cyran (Eds.), Proc. of ICMMI 2009, Springer, Advances in Intelligent and Soft Computing 59, 2009, pp. 55-67.

[5] Horzyk, A.: Wykłady z sekretów negocjacji z satysfakcją. Dostępne w internecie: http://home.agh.edu.pl/~horzyk/lectures/snzs/sekretyosobowosci.pdf.

[6] Samoilovich, S. R.: *Word Frequency Analysis as a Way to improve writing quality.* Dostępny w Internecie: http://www.usingenglish.com/articles/word-frequency-analysis-as-way-to-improve-writing-quality.html.

[7] Harris, J.: http://www.wordcount.org, 2003.

[8] http://pl.wikipedia.org/wiki/Osobowość.

[9] http://pl.wikipedia.org/wiki/Hipokrates.

[10] Łabuz, P., Urbański, M.: *Mistrz manipulacji*. Yans Cron Consulting Group Sp. z o.o. 2003/04.

[11] Magierski, S., Miklaszewski, G.: *Samoadaptacyjny sklep internetowy obsługiwany przez inteligentnego cybersprzedawcę realizującego postulaty CRMu.* - praca magisterska 2009.

[12] Imioło, M.: *Internetowy system psycholingwistyczny umożliwiający automatyczne rozpoznawanie programów osobowości człowieka.* - praca magisterska 2009.

[13] http://pl.wikipedia.org/wiki/MVC.

[14] Szymborska, W.: *Monolog dla Kasandry*, z tomu *Sto pociech*, 1967.

[15] http://www.aditus.nu/jpgraph.

[16] http://univ.gda.pl/~literat/books.htm.