**AKADEMIA GÓRNICZO-HUTNICZA**

**IM. STANISŁAWA STASZICA W KRAKOWIE**



**Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej**

Katedra Automatyki

PRACA MAGISTERSKA

Opracowanie narzędzi informatycznych wspomagających analizę statystyczną języka włoskiego

Gabriela Pastuszka  
Kierunek studiów: Informatyka Stosowana

Opiekun pracy: dr inż. Mirosław Gajer

**Oświadczenie autora**

Oświadczam, świadoma odpowiedzialności karnej za poświadczenie nieprawdy, że niniejszą pracę dyplomową wykonałam osobiście i samodzielnie, oraz że nie korzystałam ze źródeł innych niż wymienione w pracy.

Spis treści

[1. Wstęp 7](#_Toc393888670)

[1.1. Cel pracy 7](#_Toc393888671)

[1.2. Opis dokumentu 7](#_Toc393888672)

[2. Przetwarzanie języka naturalnego – wprowadzenie 9](#_Toc393888673)

[2.1. Odkrywanie wiedzy w bazie danych (KDD) 9](#_Toc393888674)

[2.2. Statystyczna analiza języka 11](#_Toc393888675)

[2.2.1. Uczenie maszynowe 11](#_Toc393888676)

[2.2.2. Eksploracja danych 12](#_Toc393888677)

[2.3. Główne zastosowania NLP 13](#_Toc393888678)

[2.4. Problemy i wyzwania NLP 15](#_Toc393888679)

[3. Zagadnienia statystyczne – omówienie 18](#_Toc393888680)

[3.1. Kwantytatywne prawa językowe 18](#_Toc393888681)

[3.1.1. Prawa Zipfa 18](#_Toc393888682)

[3.2. Metody wyszukiwania kolokacji 20](#_Toc393888683)

[3.2.1. Zliczanie wystąpień 20](#_Toc393888684)

[3.2.2. Testy statystyczne 21](#_Toc393888685)

[3.2.3. Test t Studenta 21](#_Toc393888686)

[3.2.4. Test Pearsona 21](#_Toc393888687)

[3.2.5. Test Pearsona 22](#_Toc393888688)

[3.2.6. Wskaźnik prawdopodobieństwa 22](#_Toc393888689)

[3.2.7. Punktowa wzajemna informacja 23](#_Toc393888690)

[3.2.8. Miara Poisssona-Sterlinga 24](#_Toc393888691)

[3.2.9. Współczynnik Dice 24](#_Toc393888692)

[3.2.10. Indeks Jaccarda 24](#_Toc393888693)

[4. Język włoski –charakterystyka pod kątem analizy statystycznej 26](#_Toc393888694)

[4.1. Wyrazy pochodzenia obcego 26](#_Toc393888695)

[4.2. Oznaczanie części mowy (POS tagging) 27](#_Toc393888696)

[4.2.1. Przyrostki 27](#_Toc393888697)

[4.2.2. Składnia 28](#_Toc393888698)

[5. Implementacja rozwiązania 30](#_Toc393888699)

[5.1. Środowisko programistyczne 30](#_Toc393888700)

[5.2. Interfejs graficzny 30](#_Toc393888701)

[5.3. Struktura aplikacji 32](#_Toc393888702)

[5.4. Rozpoczęcie pracy z programem 33](#_Toc393888703)

[5.4.1. Rozpoznanie kodowania znaków 33](#_Toc393888704)

[5.4.2. Obliczenie średniej długości wyrazów i zdań 33](#_Toc393888705)

[5.4.3. Sporządzenie rozkładu częstotliwości wyrazów 34](#_Toc393888706)

[5.4.4. Tokenizacja 34](#_Toc393888707)

[5.5. Analiza zjawisk związanych z częstotliwością występowania wyrazów. 35](#_Toc393888708)

[5.5.1. Ranking częstotliwości wyrazów 35](#_Toc393888709)

[5.5.2. Prawo Zipfa 35](#_Toc393888710)

[5.6. Wyszukiwanie wzorców w tekście 37](#_Toc393888711)

[5.6.1. Wyrazy pochodzenia obcego 38](#_Toc393888712)

[5.6.2. Wyrazy spełniające wyrażenia regularne 38](#_Toc393888713)

[5.7. Oznaczanie części mowy 39](#_Toc393888714)

[5.7.1. Tagger ręczny 40](#_Toc393888715)

[5.7.2. Tagger wykorzystujący wyrażenia regularne 41](#_Toc393888716)

[5.7.3. Tagger składniowy 41](#_Toc393888717)

[5.7.4. Tagger prawdopodobieństwa 42](#_Toc393888718)

[5.7.5. Ocena rezultatów tagowania 43](#_Toc393888719)

[5.7.6. Korpus wzorcowo otagowany częściami mowy 43](#_Toc393888720)

[5.8. Wyszukiwanie kolokacji 44](#_Toc393888721)

[5.8.1. Przygotowanie listy bigramów 44](#_Toc393888722)

[5.8.2. Ocena bigramów korpusu 45](#_Toc393888723)

[5.9. Kontekst słów 46](#_Toc393888724)

[6. Przykład użycia 47](#_Toc393888725)

[6.1. Podstawowe informacje o korpusie 47](#_Toc393888726)

[6.2. Prawo Zipfa 48](#_Toc393888727)

[6.3. Wyszukiwanie wzorców 48](#_Toc393888728)

[6.4. Określanie części mowy 48](#_Toc393888729)

[6.5. Wyszukiwanie kolokacji 48](#_Toc393888730)

[6.6. Podgląd kontekstu 48](#_Toc393888731)

[7. Podsumowanie 50](#_Toc393888732)

[8. Załącznik A 51](#_Toc393888733)

[9. Bibliografia 52](#_Toc393888734)

**Spis rysunków**

[Rysunek 1 Przebieg procesu odkrywania wiedzy w bazie danych 9](#_Toc393888599)

[Rysunek 2 Główne okno aplikacji 30](#_Toc393888600)

[Rysunek 3 Zależności pomiędzy elementami wzorców MVC i MVP 31](#_Toc393888601)

[Rysunek 4 Narzędzia badające ranking częstotliwości wyrazów 34](#_Toc393888602)

[Rysunek 5 Opcje związane z wykresem ilustrującym Prawo Zipfa 35](#_Toc393888603)

[Rysunek 6 Wyszukiwanie słów pochodzenia obcego 37](#_Toc393888604)

[Rysunek 7 Parametry wyszukiwania słów spełniających wyrażenie regularne 38](#_Toc393888605)

[Rysunek 8 Zakładka z narzędziami do określania części mowy 39](#_Toc393888606)

[Rysunek 9 Zakładka poświęcona wyszukiwaniu kolokacji w tekście 43](#_Toc393888607)

[Rysunek 10 Wyszukiwarka kontekstu wyrazów 45](#_Toc393888608)

[Rysunek 12 Wykres punktowy rozkładu częstotliwości i wykres linii trendu 47](#_Toc393888609)

**Spis tabel**

[Tabela 1 Zestawienie charakterystyk dla tekstów różnego pochodzenia 47](#_Toc393888633)

1. Wstęp
   1. Cel pracy

Celem niniejszej pracy magisterskiej jest opracowanie aplikacji wspomagającej analizę statystyczną języka włoskiego oraz studium przypadku użycia.

Aplikacja ta ma umożliwiać prostą i intuicyjną analizę statystyczną tekstu w języku włoskim w oparciu o dostarczony korpus języka. Użytkownik nie musi posiadać wiedzy technicznej, by wykorzystać wszystkie możliwości oferowane przez program.

Głównymi funkcjonalnościami aplikacji mają być: sporządzenie charakterystyki statystycznej dostarczonego korpusu, weryfikacja prawa Zipfa, określanie części mowy wyrazów za pomocą różnych algorytmów oraz analiza bigramów korpusu pod kątem tworzenia związków frazeologicznych i kolokacji językowych z wykorzystaniem narzędzi i testów statystycznych.

* 1. Opis dokumentu

Niniejszy dokument jest podzielony na następujące rozdziały:

Rozdział pierwszy – wstęp – nakreśla cel pracy oraz plan dokumentu.

W rozdziale drugim znajdują się informacje o przetwarzaniu języka naturalnego – możliwości i zastosowania oraz problemy związane z tą dziedziną nauki.

Rozdział trzeci opisuje zagadnienia statystyczne – zjawiska i narzędzia przydatne w dalszej części pracy.

Rozdział czwarty przybliża charakterystykę języka włoskiego pod kątem analizy statystycznej języka.

W rozdziale piątym została przedstawiona szczegółowo implementacja aplikacji będącej przedmiotem pracy.

Rozdział szósty poświęcony jest analizie przykładu użycia oraz wnioskom wyciągniętym z osiągniętych rezultatów działania aplikacji.

Rozdział siódmy podsumowuje osiągnięte cele i opisuje dalsze możliwości rozwoju aplikacji.

1. Przetwarzanie języka naturalnego – wprowadzenie

Przetwarzanie języka naturalnego (NLP) to szeroka gałąź nauki, łącząca zagadnienia informatyki, sztucznej inteligencji i lingwistyki, badająca zjawiska zachodzące w językach naturalnych. Jej zastosowanie można sprowadzić do tłumaczeń z jednego języka naturalnego na drugi, do analizy próbek w językach naturalnych i przekształcania ich na język formalny, zrozumiały przez maszyny oraz przedstawiania danych pobranych z bazy danych w języku naturalnym, zrozumiałym przez człowieka. Zwłaszcza analiza i rozumienie języka naturalnego jest tu bardzo złożonym zagadnieniem, wymagającym od komputera szerokiej wiedzy o świecie rzeczywistym, umiejętności wychwytywania sensu słów z kontekstu, interpretacji metafor, ironii i innych środków stylistycznych.

Analiza języka naturalnego dotyczy zarówno języka mówionego jak i pisanego, jednak są to dość odrębne kwestie, zmagające się z innymi problemami. Przekształcenie sygnału analogowego w symbole języka formalnego może być bardzo trudnym procesem. Dużą rolę odgrywa tutaj gwara, intonacja (świadcząca np. o emocjach mówiącego, o charakterze prowadzonej rozmowy), akcent padający na określony wyraz (może zmienić znaczenie całego zdania). W mojej pracy zajmuję wyłącznie językiem pisanym.

* 1. Odkrywanie wiedzy w bazie danych (KDD)

KDD (Knowledge Discovery in Databases) to interaktywny i iteracyjny proces przetwarzania i ekstrakcji wiedzy (wzorców, reguł, informacji) z dużych baz danych. Powstał w latach 90, kiedy w obliczu dramatycznie rosnącej ilości danych cyfrowych (dotyczących wielu dziedzin, zarówno nauki jak i biznesu) dostrzeżono potrzebę opracowania nowych narzędzi i sposobów pozwalających na szybką ekstrakcję wiedzy z baz danych. Jest to proces składający się z kilku etapów, które często wymagają od badacza podejmowania subiektywnych decyzji – nie jest to proces w pełni zautomatyzowany.

Mimo szerokiego zastosowania KDD i idącej za tym rozmaitej specyfiki danych, można wyróżnić następujące podstawowe etapy:

* zrozumienie dziedziny zastosowania i sprecyzowanie celu całego procesu;
* utworzenie zestawu danych źródłowych – wybór odpowiednich próbek;
* wstępne przetworzenie i ujednolicenie danych wejściowych – np. usunięcie szumu, obsłużenie niepełności danych (brakujące pola rekordów);
* redukcja i odpowiednie przedstawienie danych pod kątem określonego celu;
* dobór najlepszej metody ekstrakcji wiedzy z danych do określonego celu, np. streszczenie, klasyfikacja, regresja;
* wybór hipotezy, wybór modelu i parametrów do przedstawienia pod kątem użytkownika końcowego;
* eksploracja wiedzy – poszukiwanie interesujących wzorców w określonych reprezentacjach danych. Poprzez poprawne wykonanie poprzednich kroków badacz może znacząco poprawić efekty poszukiwań;
* interpretacja wykrytych wzorców, ewentualny powrót do kroków poprzednich w celu poprawy wyników;
* wykorzystanie zdobytej wiedzy – sporządzenie raportów, sprawdzenie, czy nie zachodzi konflikt z wcześniej pozyskaną wiedzą.

Uproszczony przebieg KDD uwzględniający powroty do poprzednich etapów został przedstawiony na rysunku 1.



Rysunek Przebieg procesu odkrywania wiedzy w bazie danych

W kontekście przetwarzania języka naturalnego, pierwszym etapem jest wybór odpowiednich tekstów (korpusu języka). Surowy tekst korpusu należy ujednolicić, np. poprzez zamianę dużych liter na małe, usunięcie niepożądanych znaków. Następnie, w zależności od charakteru poszukiwań, trzeba przygotować odpowiednią prezentację danych, np. rozkład częstotliwości wyrazów czy zestawienie par wyrazów często występujących razem. W tak przygotowanych danych można wyszukiwać wzorce za pomocą różnych metod, opisanych szerzej w rozdziale 2.2.2.

* 1. Statystyczna analiza języka

W celu zautomatyzowania procesów przetwarzania oraz umożliwienia analizy dużych ilości danych, stosuje się metody stochastyczne, statystyczne oraz probabilistyczne. Szczególnie przydatne są one w przypadku długich, wieloznacznych zdań, do których analizy nie wystarczają proste zasady gramatyczne. Statystyczne NLP opiera się na podejściu ilościowym – modelach probabilistycznych, teorii informacji i algebrze liniowej. Główne techniki działania to uczenie maszynowe i eksploracja danych.

* + 1. Uczenie maszynowe

Uczenie maszynowe jest cechą nowoczesnych algorytmów NLP. W przeciwieństwie do algorytmów opartych na z góry narzuconych regułach, polega ono na automatycznym uczeniu się tych reguł przez system poprzez analizę dostarczonego korpusu języka (bazy tekstów w języku naturalnym), opatrzonego prawidłowymi wartościami (np. tagami części mowy, wyjątkami specyficznymi dla danego języka, itd.). Algorytmy oparte na uczeniu maszynowym posiadają wiele zalet w porównaniu do algorytmów opartych na ręcznie wprowadzanych regułach:

* są efektywne, gdyż skupiają się na najczęściej występujących przypadkach, podczas gdy przy ręcznym definiowaniu reguł trudno określić, na co położyć największy nacisk,
* korzystając z metod wnioskowania statystycznego (estymacje, weryfikacje hipotez), są odporne na błędne lub niespotkane wcześniej dane wejściowe,
* dokładność wyników można zwiększyć poprzez dodanie nowych danych wejściowych – większa próba oznacza lepsze działanie algorytmu; ręcznie wprowadzane reguły muszą zaś być coraz bardziej skomplikowane dla polepszenia wyniku.
  + 1. Eksploracja danych

Eksploracja danych (data mining) to etap analizy w odkrywaniu wiedzy z baz danych (KDD). Polega na wykrywaniu wzorców w dużych zbiorach danych z wykorzystaniem różnych metod, głównie wywodzących się z badań nad sztuczną inteligencją, np. . Idea eksploracji danych sprowadza się do wykorzystania szybkości komputera w celu wykrywania prawidłowości ukrytych dla człowieka ze względu na ograniczone możliwości czasowe.

W przypadku NLP, wzorce te mogą dostarczyć nam wiedzy o różnych aspektach językowych danego tekstu. Przykładowo, różnorodność leksykalna czy średnia długość wyrazu może być diametralnie różna dla napisów filmowych i dla tekstu rozprawy naukowej. Duża liczba wystąpień słowa typowego dla danego dialektu, slangu czy żargonu może zdradzić informację o pochodzeniu i środowisku autora tekstu.

Istnieją różne metody przetwarzania stosowane w eksploracji danych:

* wykrywanie anomalii – nietypowe dane, mogą świadczyć o błędzie;
* modelowanie zależności – wykrywanie związków pomiędzy pozornie niezależnymi od siebie zmiennymi. Istnieje tutaj ryzyko nadinterpretacji (korelacja nie zawsze oznacza związek przyczynowo – skutkowy);
* klasteryzacja – odkrywanie grup i struktur danych mających jakieś cechy wspólne;
* klasyfikacja – stosowanie istniejących (wykrytych wcześniej) grup na nowych danych (przypisywanie im klas)
* regresja – próba znalezienia funkcji obrazującej zestaw danych z najmniejszym błędem
* streszczenie – opracowanie zwięzłej formy zestawu danych poprzez wizualizację i raporty.
  1. Główne zastosowania NLP

Przetwarzanie języka naturalnego znajduje bardzo różnorodne zastosowania, od praktycznych (np. tłumaczenie tekstów) do czysto teoretycznych (badanie zjawisk zachodzących w językach naturalnych). A oto niektóre z nich:

* automatyczne streszczanie tekstu – tworzenie krótkiego streszczenia, zawierającego najważniejsze informacje oryginalnego tekstu. Istnieją dwa rodzaje streszczeń: ekstrakt i abstrakt. Ekstrakt składa się ze słów i zdań wybranych z oryginalnego tekstu, mających największe (według oceny systemu) znaczenie. Abstrakt powstaje poprzez utworzenie semantycznej reprezentacji danych i wygenerowanie na jej podstawie tekstu w języku naturalnym. Jest to trudniejsze zadanie, wymagające zarówno trafnej intepretacji danych, jak i spójnej prezentacji w formie tekstu w języku naturalnym. Analogicznie do streszczenia można także dokonywać upraszczania – przedstawiania danego tekstu prostszym językiem, np. za pomocą zdań pojedynczych.
* tłumaczenie tekstu z jednego języka naturalnego na drugi – to trudne zadanie, a jego złożoność jest silnie zależna od stopnia skomplikowania gramatyki poszczególnych języków podlegających tłumaczeniu. Ważną kwestią jest odróżnienie idiomów i innych metaforycznych wyrażeń, których nie należy tłumaczyć dosłownie. Potrzebna jest więc dostarczona uprzednio baza takich zwrotów.
* generowanie języka naturalnego – przekształcanie treści zapisanych w bazach danych do formy czytelnej dla człowieka. Ważnym zadaniem jest dobór odpowiednich słów, poprawność gramatyczna i naturalność.
* rozumienie języka naturalnego – proces odwrotny do generacji języka o znacznie większej złożoności, wymagający trafnej analizy kontekstu, anafor, odniesień do świata zewnętrznego, a także umiejętności rozstrzygania wieloznaczności. Stanowi podstawę wielu wymienionych tutaj zastosowań, a także bardzo atrakcyjny sposób porozumiewania się z komputerem.
* rozróżnianie części mowy – zadanie pośrednie, pomocne m. in. w wykrywaniu konkretnych kolokacji (np. par czasownik + rzeczownik) i w tłumaczeniu. Automatyczne rozstrzygnięcie części mowy danego słowa nie zawsze jest możliwe, często można tego dokonać tylko z pewnym prawdopodobieństwem. Algorytmy tagujące działają iteracyjnie – najpierw oznaczane są wyrazy, których przynależność do danej części mowy jest oczywista. W kolejnych iteracjach łatwiej można określić część mowy wyrazów sąsiadujących.
* podział morfologiczny – wyodrębnienie z tekstu i klasyfikacja morfemów (najmniejszej jednostki gramatycznej). Kluczowy wpływ ma tu złożoność morfologiczna (struktura wyrazów) danego języka. Nie jest to zastosowanie samo w sobie, stanowi jeden z etapów rozumienia języka naturalnego czy dokonywania tłumaczeń.
* wykrywanie kolokacji – odnajdywanie par wyrazów często występujących razem, mogących stanowić utarte związki frazeologiczne, powiedzenia, memy ulegające modzie. Pozwalają one śledzić zmiany zachodzące w języku na przełomie lat.
* badanie specyfiki tekstów różnego pochodzenia – na podstawie zróżnicowania leksykalnego, średniej długości słowa, liczbie pomyłek w tekście i wielu innych czynników można się wiele dowiedzieć na temat danej próbki tekstu – czy jest to np. zapis nieoficjalnej rozmowy telefonicznej, czy też przemówienie polityka. Informacje te są bardzo przydatne w generowaniu tekstu w języku naturalnym, gdy ważne jest nadanie wypowiedzi odpowiedniego stylu.
* dostarczanie informacji o języku – każdy język naturalny posiada swoje specyficzne cechy, uwarunkowane kulturowo i historycznie. Stosując proste metody statystyczne można badać język pod kątem podatności na wpływ wyrazów pochodzenia obcego, neologizmów, tworzenia się dialektów i gwar.
* pomoc w nauce języka – osoba chcąca szybko nauczyć się danego języka może położyć nacisk na wyrazy najczęściej występujące w całym języku lub jego podzbiorze odpowiadającym jakiemuś zagadnieniu, np. nomenklatura prawnicza. Bardzo przydatne jest też wyszukiwanie kontekstów czy związków frazeologiczych w jakich często występuje dane słowo. Istnieją także programy pomagające w nauce wymowy lub pisma, oceniające podobieństwo próbek dźwiękowych i graficznych.
* automatyczne ocenianie wypracowań, esejów – jako pomoc dla nauczyciela, program taki może szybko ocenić różnorodność leksykalną, liczbę błędów, a także wartość merytoryczną wypracowań, np. zgodność z faktami historycznymi.
* rozpoznawanie i definicja nazw własnych – wyszukiwanie w tekście wyrazów będących nazwami własnymi i szerszych informacji o nich (przydatne zwłaszcza w przypadku języków, w których nazw własnych nie pisze się wielką literą).
  1. Problemy i wyzwania NLP

Przetwarzanie języka naturalnego, a zwłaszcza jego rozumienie natrafia na wiele problemów i barier, dlatego zagadnienie to jest często nazywane problemem AI-complete (problemem sztucznej inteligencji). Aby z powodzeniem rozwikłać liczne wieloznaczności języka naturalnego, potrzebna jest bowiem szeroka wiedza o świecie zewnętrznym, a także rozumienie aluzji i skojarzeń – sposób myślenia właściwy dla ludzkiego mózgu.

Główne problemy NLP można przedstawić następująco:

* wieloznaczność składniowa zdania – zwana też amfibologią - zwłaszcza długie, wielokrotnie złożone zdania dopuszczają mnogość interpretacji. W niektórych przypadkach poprawna interpretacja wymaga elementarnej wiedzy o świecie i nie stanowi problemu dla człowieka. Np. w przypadku zdań *Daliśmy małpom banany, bo były głodne* i *Daliśmy małpom banany, bo były przejrzałe* czasownik *były* odnosi się za każdym razem do innego rzeczownika. Poprawne ich zintepretowanie wymaga wiedzy na temat małp i bananów.
* wieloznaczność zakresu – rodzaj błędu logicznego, wynikający z nieprawidłowego uzycia kwantyfikatorów lub podobnych określeń, takich jak każdy, wiele, ktoś, nikt. Np. angielskie zdanie *Many people visit Europe every month* może oznaczać zarówno fakt, iż istnieje grupa wielu ludzi zwiedzających Europę co miesiąc, oraz że Europa jest odwiedzana co miesiąc przez wielu ludzi (niekoniecznie tych samych).
* rozstrzyganie odniesień – w języku angielskim problematyczne jest jednoznaczne określenie, jakiego słowa tyczy się dany przymiotnik, np. wyrażenie *pretty little girls’ school* można rozumieć aż na 4 sposoby (przymiotniki *pretty* i *little* mogą dotyczyć zarówno słowa *girls*, jak i *school*).
* błędy językowe, logiczne, stylistyczne - w języku polskim wieloznaczności rodzi np. rzeczownik w funkcji przydawki lub rzeczownik mogący spełniać zarówno funkcję podmiotu jak i dopełnienia, np. *Samochód wyprzedził motocykl*. Kolejny rodzaj wieloznaczności powodują homonimy, np. Problemy tego typu są cechą języków naturalnych i nie tylko komputer miałby trudności z ich interpretacją.
* homonimy – wiele wyrazów ma więcej niż jedno znaczenie, więc tylko kontekst, w jakim dane słowo występuje, może pomóc w określeniu jego znaczenia. Filozof języka, Ludwig Wittgenstein, zajmował się m. in. problemem znaczenia słowa i ujął to w ten sposób: „*Znaczeniem słowa jest sposób użycia go w języku.”*
* niemożność określenia części mowy – problem ten dotyczy głównie języków analitycznych, posiadających znikomą morfologię fleksyjną, która w innych językach pomaga odróżnić od siebie części mowy. Przykładem jest angielskie zdanie *Time flies like an arrow.* Oprócz oczywistego porównania upływu czasu do lotu strzały można wyróżnić co najmniej 6 innych znaczeń, w zależności od zaklasyfikowania poszczególnych wyrazów do odpowiednich części mowy. Również w językach syntaktycznych, np. w języku polskim, nie zawsze da się jednoznacznie określić część mowy – przykładem może być słowo *śledź*, które można interpretować jako rzeczownik lub czasownik w trybie rozkazującym (może to prowadzić do zabawnych konsekwencji – pozycję w menu pewnej restauracji brzmiącą po polsku *‘śledź w oleju’* przetłumaczono na angielski jako *‘follow in oil’*).
* interpretacja idiomów – ponieważ nie można ich traktować dosłownie, wymagają one wcześniej zdefiniowanej bazy znaczeń. W przypadku dokonywania tłumaczenia, można także spróbować znaleźć analogiczny idiom w języku docelowym.
* błędne dane wejściowe – literówki, brak znaków interpunkcyjnych mogą niekiedy całkowicie uniemożliwiać ich poprawne zrozumienie przez komputer (a także przez człowieka). Drobne błędy pisowni mogą być łatwo wykryte i naprawione automatycznie, o ile nie zachodzi wieloznaczność.
* wpływ akcentu na znaczenie zdania – w interpretacji języka mówionego dużą rolę odgrywa akcent, który wskazując na wybrane słowo zmienia zupełnie znaczenie całego zdania.
* wieloznaczność mowy, homofonów – kolejny rodzaj wieloznaczności, dotyczący tylko języka mówionego, np. zlepek słów *mamamanastroje* można rozbić na co najmniej 8 różnych zdań, każde o innym znaczeniu.

Podsumowując, dobry system NLP mający znaleźć praktyczne zastosowanie, musi być dobry w podejmowaniu decyzji ujednoznaczniających znaczenie wyrazu oraz część mowy, do której należy, strukturę składniową zdania i zakres semantyczny. Powinien być także odporny na błędne dane wejściowe (w miarę możliwości poprawiać błędy pisowni). Dodatkowo, jeśli system ma rozumieć mowę, musi radzić sobie także z jej niuansami – wieloznacznością homonimów i zmianą znaczenia zdania w zależności od wyrazu, na który położony jest akcent.

1. Zagadnienia statystyczne – omówienie
   1. Kwantytatywne prawa językowe
      1. Prawa Zipfa

Prawa Zipfa uważa się za najstarsze i najlepiej zbadane ilościowe prawa językowe. Prawa te opisują szereg prawidłowości językowych o charakterze ilościowym, w szczególności wymienia się związki pomiędzy:

* częstością wyrazów a ich pozycją na liście rangowej,
* częstością wyrazów a ich długością,
* częstością wyrazów a liczbą ich znaczeń,
* częstością wyrazów a ich wiekiem i pochodzeniem.

W literaturze lingwistycznej pojęcie prawa Zipfa kojarzone jest najczęściej tylko z pierwszą zależnością, opartą na powszechnie znanej prawidłowości, zaobserwowanej m. in. przez J.B. Estoupa, zgodnie z którą iloczyn rang i częstotliwości słów z listy frekwencyjnej jest wartością stałą. W celu opisu tej zależności Zipf zaproponował model:

gdzie:

- prawdopodobieństwo wystąpienia wyrazu o randze r

k – stała

Pierwsze prawo Zipfa było wielokrotnie modyfikowane. Jedną z ważniejszych modyfikacji jest propozycja B. Mandelbrota, który wprowadził do równania poprawki uwzględniające nieregularny kształ krzywej modelu w obszarze najniższych rang. Prawo Zipfa – Mandelbrota opisane jest wzorem:

gdzie:

- prawdopodobieństwo wystąpienia wyrazu o randze r

k – stała

B – współczynnik modelu (stały dla konkretnego korpusu)

– współczynnik spełniający warunek dla i dla

Kolejna zależność odkryta przez Zipfa mówi, że długość wyrazu maleje w miare wzrostu jego częstości (przy czym jednostką długości jest fonem lub sylaba). Zipf zaproponował następujący wzór opisujący ową zależność:

gdzie:

k – długość wyrazu

C – stała

r - ranga wyrazu

Zipf sformułował też prawo mówiące, że liczba znaczeń wyrazu jest wprost proporcjonalna do pierwiastka jego częstości. Zależność tę wyraził wzorem:

gdzie:

k – liczba znaczeń wyrazu

C – stała

f - częstość wyrazu

Sposób, w jaki należy odrębniać znaczenia wyrazu, nie został przez Zipfa sprecyzowany.

* 1. Metody wyszukiwania kolokacji

Związek frazeologiczny to utrwalone w użyciu połączenie dwóch lub więcej wyrazów, którego znaczenie jest inne niż prosta suma jego części składowych. Pod względem gramatycznym wyróżniamy trzy typy:

* Wyrażenie, w którym ośrodkiem frazeologizmu jest rzeczownik, przymiotnik bądź przysłówek;
* Zwrot – ośrodkiem frazeologizmu jest czasownik;
* Fraza – związek frazeologiczny będący całym zdaniem bądź równoważnikiem zdania.

Zaś pod względem znaczenia, możemy wyodrębnić 3 typy:

* Terminy specjalistyczne, np. *pamięć podręczna*, *dysk twardy*,
* Nazwy własne, nazwiska, np. *Nowy Jork*, *Albert Einstein*,
* Wyrażenia numeryczne, np. *sto dwadzieścia osiem*.

Kolokacje mają ograniczoną kompozycyjność, co oznacza, że nie można jednoznacznie stwierdzić, jakie mają znaczenie na podstawie znaczenia samych ich części składowych. Wyrażeniami charakteryzującymi się najmniejszą kompozycyjnością są idiomy, które mają znaczenie całkowicie przenośne. Kolokacje mówią nam o tym, jakie połączenia wyrazowe brzmią naturalnie (np. mówi się *mocna herbata*, a nie *silna herbata*, *podjąć decyzję*, a nie *zrobić decyzję*, itd.). Wyszukiwanie kolokacji danego słowa pozwala nam poznać częstotliwość jego występowania w różnych kontekstach. Znajomość takich kontekstów jest bardzo przydatna osobom uczącym się języków obcych.

Istnieje wiele sposobów wyodrębniania z tekstu kolokacji. W następnych podrozdziałach krótko omówię kilka z nich.

* + 1. Zliczanie wystąpień

Najprostszą metodą znajdywania kolokacji jest zliczanie wystąpień bigramów (par wyrazów). Oczywistym wynikiem będą nic nieznaczące pary wyrazów – w przypadku języka włoskiego, pary składające się z rodzajników, zaimków, przyimków, spójników. Należy więc przefiltrować listę wynikową pod kątem części mowy. Pomysł takiego filtra został opracowany przez Justesona i Katza. Metoda ta jest dobra również do wyszukiwania najczęstszych kolokacji danego wyrazu, w celu sprawdzenia kontekstów występowania.

* + 1. Testy statystyczne

Istotą testów statystycznych jest ocena przypadkowości współwystępowania dwóch słów. Definiujemy zdarzenie P – wystąpienie kolokacji składającej się ze słów w­1 i w2. Następnie formułujemy hiptezę zerową H0 twierdzącą, że zdarzenie P jest przypadowe. prawdopodobieństwo wystąpienia zdarzenia P zakładając, że hipoteza jest prawdziwa. Odrzucamy hipotezę H0 na pewnym poziomie ufności, np. p < 0,01. W przeciwnym przypadku, hipotezę uznajemy za prawdziwą. Przyjmijmy, że jeśli współwystąpienie słów w­1 i w2 jest przypadkowe, to:

* + 1. Test t Studenta

Test t pozwala na podstawie średniej i wariancji próby określić prawdopodobieństwo pochodzenia próby z populacji o rozkładzie ze średnią µ:

gdzie:

- średnia próby

- wariancja próby

N – wielkością próby

µ - średnią rozkładu

Jeśli t jest odpowiednio duże, możemy odrzucić hipotezę H0. Wartość krytyczną dla zadanego poziomu ufności można uzyskać z tablic statystycznych.

* + 1. Test Pearsona

Test Pearsona jest kolejnym testem badającym zależność występowania dwóch zdarzeń. W przeciwieństwie do testu t, nie zakłada on normalnego rozkładu prawdopodobieństw.

Istotą testu jest porównanie częstotliwości zaobserwowanych z częstotliwościami oczekiwanymi dla niezależności zdarzeń. Jeśli różnica między tymi częstotliwościami jest duża, można odrzucić hipotezę zerową (zdarzenie nie jest przypadkowe – bigram jest kolokacją).

Dla każdego bigramu w­1, w2 tworzona jest tabela przedstawiająca rozkład częstotliwości dla kombinacji bigramów zawierających i nie zawierających wyrazy tworzące dany bigram.

|  |  |  |
| --- | --- | --- |
|  | w­1 | ~w­1 |
| w2 | nii | noi |
| ~w2 | nio | noo |

Miara sumuje różnice między obserwowaną i oczekiwaną częstotliwością dla wszystkich pól tabeli, zgodnie ze wzorem:

gdzie *i* odnosi się do numeru wiersza, a *j* do numeru kolumny tabeli, oznacza wartość obserwowaną dla komórki (*i,j*), a oznacza wartość oczekiwaną.

Wyliczoną wartość należy porównać z wartością krytyczną dla danego poziomu ufności. Jeśli jest od niej większa, możemy odrzucić hipotezę zerową.

* + 1. Test Pearsona

Jest to wariant testu – różnica polega na uwzględnieniu we wzorze ilości wszystkich bigramów.

* + 1. Wskaźnik prawdopodobieństwa

Innym podejściem do testowania hipotez jest wskaźnik prawdopodobieństwa (*likelihood ratio*). Jest to liczba określająca jak bardzo jedna hipoteza jest prawdopodobniejsza od drugiej.

Zastosowanie wskaźnika prawodpodobieństwa zakłada sformułowanie dwóch hipotez, wyjaśniających występowanie bigramu :

Hipoteza pierwsza oznacza niezależność występowania dwóch słów obok siebie, a hipoteza druga sugeruje występowanie zależności pomiędzy tymi słowami.

Wartość prawdopodobieństw , i zostaje oszacowana za pomocą ilości wystąpień osobnych wyrazów oraz bigramów.

gdzie:

- liczba wystąpień słowa ,

- liczba wystąpień słowa ,

- liczba wystąpień bigramu

Przyjmując rozkład dwumianowy:

prawdopodobieństwo otrzymania zaobserwowanych ilości wystąpień , oraz wynosi dla pierwszej hipotezy i dla hipotezy drugiej.

* + 1. Punktowa wzajemna informacja

Kolejną miarą określania zależności między wyrazami jest wzajemna informacja, opisująca związek między losowymi zmiennymi – w tym przypadku, liczbą wystąpień poszczególnych słów.

Jest ona opisana wzorem:

Miara ta jest symetryczna, zatem: = . Jej wartość określa ilość informacji w bitach, jaką dostarcza wystąpienie wydarzenia o wystąpieniu wydarzenia . Wartość zerowa oznacza, że wystąpienia wyrazów nie są ze sobą w żaden sposób powiązane.

* + 1. Miara Poisssona-Sterlinga

Miara Poissona-Sterlinga jest kolejną miarą oceniającą prawdziwość hipotezy zerowej. Im wyższa jej wartość, tym większa zależność pomiędzy dwoma wyrazami – większa szansa na odrzucenie hipotezy. Miara ocenia bigramy za pomocą następującego wzoru:

gdzie jest zaobserwowaną częstotliwością występowania bigramu, a częstotliwością oczekiwaną dla zdarzenia niezależnego.

* + 1. Współczynnik Dice

Współczynnik ten określa podobieństwo dwóch próbek rozkładu. Wzór na współczynnik dla bigramu ma postać:

gdzie:

jest liczbą wystąpień bigramu

jest sumą wystąpień bigramów ( jest wyrazem różnym od )

jest sumą wystąpień bigramów

* + 1. Indeks Jaccarda

Indeks ten służy do oceniania podobieństwa i różnicy między zbiorem próbek. Ogólny wzór na indeks Jaccarda wygląda następująco:

gdzie A i B są zbiorami próbek.

W szczególności może on posłużyć jako miara prawdopodobieństwa wystąpienia danego bigramu – dla takiego zastosowania można użyć wzoru w następującej formie:

gdzie:

jest liczbą wystąpień bigramu

jest sumą wystąpień wszystkich bigramów

1. Język włoski –charakterystyka pod kątem analizy statystycznej

Język włoski ma wiele interesujących cech, które ułatwiają jego analizę statystyczną. Dokładniejszy opis języka znajduje się w – w tym rozdziale powtórzę tylko najważniejsze informacje.

* 1. Wyrazy pochodzenia obcego

Ponieważ każdy wyraz w języku włoskim (poza kilkoma wyjątkami) kończy się na samogłoskę, łatwo znaleźć w tekście wyrazy obcego pochodzenia. Co więcej, w alfabecie języka włoskiego nie występują litery *k, j ,w, x, y,* toteż wyrazy zawierające choć jedną z nich również można zakwalifikować jako zapożyczone z innych języków. Wyjątkiem są tu słowa, które na stałe zagościły w słowniku języka włoskiego, np. *extracommunitari, xenofobia.*

Innym, o wiele skuteczniejszym sposobem wykrywania wyrazów obcych jest sprawdzanie słów pod kątem występowania w nich bigramów nietypowych dla języka włoskiego. Metoda ta działa dla każdego języka – w istocie jest to popularny sposób automatycznego rozpoznawania języków. Posiada jednak dużą wadę – przy alfabecie zawierającym 21 liter podstawowych i 10 liter ze znakami diakrytycznymi (5 liter obcych dla języka włoskiego także może być tu brane pod uwagę, jeśli chcemy wykryć wspomniane wcześniej wyrazy z obcą pisownią) istnieje aż 961 bigramów, z czego większość stanowią te wymagające sprawdzenia, nietypowe dla języka włoskiego. Metoda ta jest zatem bardzo złożona obliczeniowo. W mojej pracy poprzestanę więc na wspomnianych wcześniej metodach wykorzystujących specyficzne cechy języka włoskiego.

* 1. Oznaczanie części mowy (POS tagging)

Język włoski charakteryzuje się dużą regularnością, bardzo pomocną w określaniu części mowy. Tę właściwość można wykorzystać programując taggery części mowy.

* + 1. Przyrostki

Istnieje wiele przyrostków wyrazów typowych tylko dla czasowników, rzeczowników i przymiotników (trzeba jednak wziąć pod uwagę nieliczne wyjątki). Za pomocą taggera wykorzystującego wyrażenia regularne można zatem określić pewien odsetek słów. A oto najczęściej występujące wzorce:

* Przysłówki:
  + –mente;
* Rzeczowniki:
  + -ina, -ino (z wyjątkiem przymiotników divino, vicino) – końcówka zdrabniająca wyraz
  + -etto – (z wyjątkiem detto, letto – czasowniki w formie biernej) – jw.
  + -one, -ona (z wyjątkiem przymiotnika *buono*) – końcówka oznaczająca zgrubienie wyrazu
  + -ore z wyjątkiem peggiore, maggiore;
  + -ezza – rzeczowniki utworzone z czasowników (np. *bellezza - piękno)*
  + –tà – rzeczowniki rodzaju żeńskiego nie zmieniające formy w liczbie mnogiej
  + -aggio
  + -ismo
* Czasowniki:
  + -are, -ere, -ire – w formie bezokolicznika;
  + -armi, -arti, -arla, -arle, -arli, -arci, arvi, argli, arsi – bezokolicznik zakończony zaimkiem (forma skrócona). Należy tu wykluczyć czasownik *parlare*, posiadający w odmianie formy *parla*, *parli.*
  + -amo – czasownik w drugiej osobie liczby mnogiej;
  + -rà – końcówka czasownika w trzeciej osobie liczby pojedynczej w czasie Futuro Semplice
  + -ando, -endo - imiesłów
  + -andola, -andole, -andone, -andosi – imiesłów zlożony (zakończony zaimkiem)
* Przymiotniki:
  + -issimo, -issima, -issimi, -issime –stopień najwyższy bezwzględny;
  + -ile;
  + -oso, -osa.
    1. Składnia

Informacji o przynależności wyrazu do części mowy mogą dostarczyć części mowy w jego sąsiedztwie (jeśli zostały już określone).

* Zaimki dzierżawcze zwykle poprzedzają rzeczownik, którego dotyczą (w niektórych przypadkach kolejność jest odwrotna, wtedy po zaimku następuje jakaś określona wcześniej część mowy, np. spójnik);
* Zaimki dopełnienia bliższego w przeważającej liczbie przypadków poprzedzają swój czasownik zwrotny (w zdaniach z czasownikami modalnymi mogą dołączyć do bezokolicznika – sytuacja omówiona w 3.2.1 lub wystąpić bezpośrednio przed czasownikiem modalnym).
* Przymiotniki w stopniu najwyższym są tworzone przez strukturę il/la più [przymiotnik], np. *Marco è il ragazzo più alto.* Zatem, gdy tagger napotka parę *il/la più* , lub *il/la* [rzeczownik] *più* , kolejnym wyrazem musi być przymiotnik.
* Rodzajniki i przyimki, które uległy zabiegowi elizji w większości przypadków stoją przed rzeczownikiem (zaczynającym się samogłoską).

Istnieją też utrudnienia w oznaczaniu części mowy:

* wyrazy, które mogą być zarówno rzeczownikiem jak i przymiotnikiem – np. w języku polskim wyrazy *chory, zakochany.* W języku włoskim może to być zwrot *il mio [ragazzo],* w którym zaimek dzierżawczy w domyśle pełni funkcję rzeczownika *chłopak.*
* przyimki złożone, pełniące jednocześnie funkcję rodzajnika – della, nello;
* zaimki nieokreślone liczby mnogiej (*dei, degli, delle*) – brzmią tak samo jak wspomniane w poprzednim podpunkcie przymki złożone;
* czasowniki złożone, z doklejonym na końcu zaimkiem osobowym;
* homonimy – wyrazy o tej samej pisowni, lecz innym znaczeniu i części mowy, np. *letto* może być rzeczownikiem, czasownikiem (imiesłów czasu przeszłego od czasownika *leggere)* orazprzymiotnikiem (pochodzącym od tego samego czasownika).
* słówko *ci* ma wiele zastosowań – jest to zaimek dopełnienia bliższego i dalszego dla pierwszej osoby liczby mnogiej, a także partykuła zastępująca dłuższe wyrażenia, m. in.. określenia czasu i miejsca, dopełnienia wprowadzanie przez przyimki *a, da, con, su*. Ponadto wchodzi w skład wielu czasowników złożonych, np. *esserci, volerci, avercela* (w odmianie tych czasowników składnik *ci* jest odrębnym słowem).

1. Implementacja rozwiązania
   1. Środowisko programistyczne

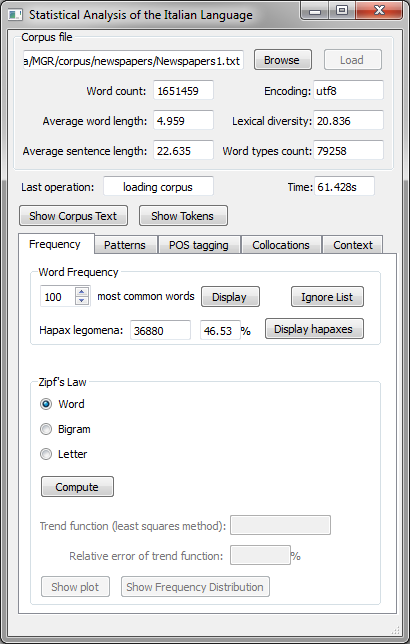
Aplikacja została napisana w języku Python (wersja 2.7). Interfejs graficzny powstał przy użyciu PyQt (nakładka na popularną bibliotekę Qt dla języka Python). Cały projekt powstał w środowisku Eclipse z nakładką PyDev, umożliwiającą integrację Pythona i QtDesignera (edytor GUI). W celu kompilacji projektu w systemie operacyjnym Windows potrzebny był także zestaw narzędzi deweloperskich MinGW.

Oprócz tego, pomocne były moduły dla Pythona: wspomagający obliczenia NumPy, wyświetlający wykresy PyQtGraph oraz biblioteka NLTK dostarczająca wiele narzędzi wspomagających obliczenia statystyczne.

* 1. Interfejs graficzny

Interfejs został zaprojektowany przy użyciu dostarczonego wraz z PyQt QtDesignera. Główne okno aplikacji składa się z trzech części – interfejsu właściwego oraz dwóch okien podglądu (do wyświetlania nieprzetworzonego tekstu korpusu oraz tekstu podzielonego na tokeny). Interfejs podzielony jest na obszary poświęcone charakterystyce statystycznej tekstu, informacjom o ostatniej wykonanej operacji i czasie jej trwania oraz pięć zakładek odpowiadających różnym zagadnieniom z zakresu analizy statystycznej języka:

* *Frequency* – są tu informacje związane z częstotliwością występowania i prawem Zipfa dla wyrazów, bigramów i liter oraz wartości średnich długości słowa i zdania dla korpusu;
* *Patterns* – umożliwia wyszukiwania oparte na wzorcach i wyrażeniach regularnych, m. in. słowa pochodzenia obcego;
* *POS tagging* – pozwala na taggowanie wyrazów korpusu częściami mowy przy użyciu 4 rodzajów taggerów. Możliwe jest załadowanie korpusu już otagowanego, w celu sprawdzenia efektywności tagowania.
* *Collocations* – umożliwia wyszukiwanie związków frazeologicznych za pomocą 10 różnych testów statystycznych;
* *Context* – zakładka ta zawiera wyszukiwarkę kontekstu dla podanego słowa.



Rysunek Główne okno aplikacji

* 1. Struktura aplikacji

Aplikacja została zaprojektowana zgodnie ze wzorcem Model-View-Presenter, będącym pochodną popularnego wzorca Model-View-Controller.



Rysunek Zależności pomiędzy elementami wzorców MVC i MVP

Model to klasa zawierająca całą logikę aplikacji. Przechowuje wczytany korpus, wyodrębnione z niego tokeny, rozmaite dane potrzebne do przetwarzania tekstu (np. reguły określania części mowy) oraz wyniki bieżących wyszukiwań. Udostępnia szereg metod przeprowadzających obliczenia i pozwalających na pobranie ich wyników. Jest to model pasywny – zmiany zachodzą jedynie na żądanie Prezentera. Istnieje niezależnie od Prezentera i Widoku („nie wie” o ich istnieniu).

Widok to zbiór widżetów biblioteki PyQt umożliwiający prezentację danych. Jest wygenerowany przez QtDesignera na podstawie projektu interfejsu opisanego w poprzednim podrozdziale. Podobnie jak Model, Widok jest niezależny od pozostałych komponentów.

Prezenter posiada informacje o Widoku i Modelu oraz zarządza przepływem danych. W konstruktorze klasy Prezentera następuje wczytanie widoku z pliku wygenerowanego przez QtDesignera oraz połączenie sygnałów tego widoku ze slotami – metodami Prezentera. W metodach tych Prezenter sprawdza, jeśli jest to konieczne, zaznaczone opcje w interfejsie i wywołuje odpowiednie metody Modelu przeprowadzające pożądane obliczenia i dostarczające ich wyniki. Pobrane wyniki są wyświetlane w widżetach Widoku. Prezenter jest zatem ściśle zależny od interfejsu Modelu i Widoku.

* 1. Rozpoczęcie pracy z programem

Pracę z aplikacją rozpoczyna się poprzez wczytanie pliku z korpusem języka za pomocą przycisków Browse i Load (slotem dla sygnału – wciśnięcia przycisku Browse jest standardowa funkcja biblioteki PyQt otwierająca okienko dialogowe umożliwiające wskazanie pliku). Przyciski i ścieżka do pliku z korpusem są widoczne z każdego miejsca programu i w dowolnej chwili można zmienić plik korpusu, z którym się pracuje. Podczas każdego wywołania metody ładowania pliku są przeprowadzane procedury opisane w podpunktach 5.4.1 - 5.4.4, które mają na celu przygotowanie wczytywanego tekstu z korpusem i obsłużenie ewentualnych wyjątków, dokonanie prostych obliczeń dostarczających podstawowe informacje statystyczne o tekście oraz podział na tokeny umożliwiający dalsze analizy pod kątem części mowy i wyszukiwania związków frazeologicznych.

* + 1. Rozpoznanie kodowania znaków

W trakcie wczytywania tekstu konieczne jest określenie jego kodowania – znaki specjalne języka włoskiego są najczęściej kodowane za pomocą standardu Unicode lub ISO. Domyślnie plik jest odczytywany za pomocą popularniejszego kodowania UTF-8, a w razie wystąpienia wyjątku stosowany jest ISO-8859-1 (zwany też Latin-1). Gdy i tym razem dojdzie do wyjątku, plik nie zostaje wczytany, a w polu zatytułowanym ‘Encoding’ pojawia się komunikat o błędzie. Natomiast po udanym załadowaniu pliku surowy tekst jest zapisywany w pamięci programu.

* + 1. Obliczenie średniej długości wyrazów i zdań

W celu znalezienia średniej długości zdania należy wyodrębnić je z tekstu, zliczyć i wyrazić ich długość w liczbie składających się na nie słów. W tym celu użyłam klasy PunktSentenceTokenizer bliblioteki NLTK. Domyślnie tokenizer ten traktuje kropki, wykrzykniki i pytajniki jako koniec zdania. Można go skonfigurować, dostarczając listę skrótów (ciągów znaków, po których kropka nie oznacza końca zdania). Wynikiem działania tokenizera jest lista stringów wraz ze wszystkimi znakami specjalnymi (białe znaki i inne znaki specjalne).

Do wydzielenia słów w zdaniach użyłam kolejnego tokenizera, bazującego na wyrażeniach regularnych. Jako słowa traktuje on ciągi liter alfabetu włoskiego (ignoruje wszelkie inne znaki). Na podstawie wyznaczonych list zdań i wyrazów obliczone zostają średnia długość zdania i wyrazu w korpusie.

* + 1. Sporządzenie rozkładu częstotliwości wyrazów

Wykonany wcześniej podział tekstu na słowa można wykorzystać do utworzenia rozkładu częstotliwości występowania wyrazów, uprzednio zamieniając wszystkie litery na małe, by uniknąć duplikatów. Dobrym narzędziem jest klasa FreqDist z biblioteki NLTK. Ma strukturę opartą na bazie słownika i szereguje wyrazy od najczęściej występującego. Wyznaczony rozkład zostaje zapisany w pamięci programu jako pole składowe Modelu.

Dzięki rozkładowi częstotliwości można też obliczyć różnorodność leksykalną. Jest to iloraz liczby typów słów występujących w tekście oraz liczby wszystkich słów, może zatem osiągać wartości w przedziale , przy czym różnorodność równa 1 oznacza, że żadne słowo się nie powtarza.

* + 1. Tokenizacja

Korzystając z listy wyodrębnionych zdań (zawierających także m. in. znaki przestankowe) można utworzyć listę tokenów. Tokenem może być wyraz (ciąg liter), ciąg liter i cyfr, a także dowolny inny znak mogący pojawić się w tekście. Ponieważ znaki przestankowe (kropki, przecinki, pytajniki, wykrzykniki, nawiasy, itd.) mają znaczenie dla interpretacji składni zdania, należy wydzielić je jako osobne tokeny. W tym celu użyłam kolejnego taggera opartego na wyrażeniach regularnych. Rezultatem działania taggera jest lista tokenów, zapisana do pamięci aplikacji.

Po wykonaniu tych procedur wszystkie zakładki programu stają się aktywne, a obliczone charakterystyki tekstu: całkowita liczba wyrazów, kodowanie, średnia długość wyrazu, zdania, współczynnik różnorodności leksykalnej zostają wyświetlone w odpowiednich polach. W lewym oknie podglądu pojawia się nieprzetworzony tekst wczytanego korpusu, a w prawym – lista tokenów całego tekstu. Program jest gotowy do pracy.

* 1. Analiza zjawisk związanych z częstotliwością występowania wyrazów.

Pierwsza zakładka, zatytułowana *„Frequency”* (częstotliwość), zawiera narzędzia pozwalające na badanie częstotliwości wyrazów i innych powiązanych zjawisk. Jej zawartość została podzielona na grupy omówione w kolejnych podrozdziałach.

* + 1. Ranking częstotliwości wyrazów

Grupa przycisków zatytułowana *„Word Frequency”*, przedstawiona na rysunku nr 4 umożliwia poznanie wybranej liczby najczęściej występujących wyrazów i listy hapaksów.



Rysunek Narzędzia badające ranking częstotliwości wyrazów

Przycisk *„Ignore List”* wyświetla okno dialogowe z edytowalną listą wyrazów ignorowanych, które nie będą pojawiać się w zestawieniu, można ją również dostarczyć poprzez plik o nazwie „*commonIgnoredWords.txt”* znajdujący się w katalogu programu (domyślnie lista ta jest pusta). Po kliknięciu przycisku *„Display”* pożądana lista wyrazów wraz z liczbą ich wystąpień zostaje pobrana ze sporządzonego wcześniej rozkładu częstotliwości i wyświetlona w osobnym okienku dialogowym.

Innym zjawiskiem związanym z częstotliwością występowania wyrazów są tzw. hapax legomena – wyrazy pojawiające się w tekście tylko raz. Można poznać ilość takich wyrazów, ich procentowy udział w tekście, a także wyświetlić je wszystkie w osobnym oknie dialogowym za pomocą przycisku *„Display Hapaxes”* (obiekt klasy FreqDist z biblioteki NLTK posiada metodę hapaxes()).

* + 1. Prawo Zipfa

W kolejnej grupie zatytułowanej *„Zipf’s Law”* (Prawo Zipfa) znajdują się narzędzia do analizy prawa Zipfa dla wyrazów, bigramów literowych i pojedynczych liter.



Rysunek Opcje związane z wykresem ilustrującym Prawo Zipfa

Po wybraniu interesującej nas jednostki tekstu i kliknięciu przycisku *„Compute”*, następuje przygotowanie danych do sporządzenia wykresu ilustrującego prawo Zipfa – wykresu zależności liczby wystąpień danej jednostki tekstu od jej pozycji w rankingu częstotliwości występowania. Dane te - współrzędne punktów na wykresie - są uzyskiwane w różny sposób, w zależności od wybranej jednostki tekstu:

* w przypadku wyrazu, punktów na wykresie musi być tyle, ile w tekście jest typów wyrazów. Liczby wystąpień są pobrane z gotowego rozkładu częstotliwości wyrazów.
* bigramy literowe – liczba wszystkich możliwych bigramów jest niezależna od tekstu i wynosi 1296 - jest to moc zbioru znaków alfabetu włoskiego (36) podniesiona do potęgi drugiej (zostały uwzględnione także znaki specjalne). Ilość wystąpień każdego bigramu w tekście jest obliczana poprzez iterację po zbiorze wyodrębnionych wcześniej tokenów i zliczenie bigramów występujących w każdym z nich.
* litery – analogicznie do bigramów, liczba liter jest naturalnie ograniczona przez moc zbioru znaków alfabetu włoskiego – 36, uwzględniając znaki specjalne; podobnie jest też określana liczba ich wystąpień w tekście (poprzez iterację po zbiorze tokenów i zliczenie liter).

W przypadku bigramów i liter znaki nienależące do z góry zdefiniowanego zbioru znaków alfabetu włoskiego są ignorowane, w przeciwieństwie do wyrazów.

Dla dobrego zilustrowania prawa Zipfa, przy pomocy metody najmniejszych kwadratów, zostaje wyliczona linia trendu. W tym celu użyłam funkcji *polyfit* biblioteki Numpy, która zwraca wektor współczynników wielomianu aproksymującego o zadanym stopniu – w tym przypadku równym 1. Współczynniki zostają zapisane w pamięci programu w celu dalszego użycia. Znalazłszy współczynniki wielomianu drugiego stopnia (linii trendu) można również obliczyć średni względny błąd przybliżenia, wg wzoru:

gdzie:

n – ilość punktów pomiarowych,

x­­­­­­i – i-ta wartość mierzona,

xia - i-ta wartość aproksymowana.

Obliczone wartości (wzór funkcji – linii trendu i średni błąd względny) pojawiają się w odpowiednio opisanych polach. Przycisk *„Show Plot”* staje się aktywny. Po jego naciśnięciu pojawia się okienko dialogowe z wykresem punktowym oraz wykresem funkcji aproksymującej (w skali logarytmicznej).

W przypadku bigramów i liter, można poznać ich rozkład częstotliwości za pomocą przycisku *„Show Frequency Distribution”* (aktywnego po dokonaniu obliczeń dla prawa Zipfa). Podobnie jak w przypadku wyrazów, zostaje wyświetlone okienko dialogowe z listą bigramów lub liter, posortowanych wg liczby wystąpień. Możliwe jest jednoczesne wyświetlenie okna z wykresem Zipfa oraz rozkładu częstotliwości, w celu przenalizowania liczby wystąpień poszczególnych jednostek.

* 1. Wyszukiwanie wzorców w tekście

Kolejnym zagadnieniem dostarczającym wielu informacji o tekście jest wyszukiwanie wzorców. W kolejnych podrozdziałach omówię zawartość zakładki *„Patterns”* (Wzorce), zawierającej narzędzia do wyszukiwania wyrazów spełniających określone wzorce.

* + 1. Wyrazy pochodzenia obcego

Pierwsza grupa widżetów jest zatytułowana *„Foreign Words”* (Wyrazy pochodzenia obcego) i zawiera narzędzia do wyszukiwania takich wyrazów na podstawie dwóch reguł omówionych w podrozdziale 4.1. Reguły te można stosować oddzielnie – do ich wybierania służą przyciski typu CheckBox. Oprócz tego, istnieje możliwość zdefiniowania listy wyjątków – wyrazów, które nie zostaną uznane za pochodzenia obcego mimo spełnienia reguły. W tym celu należy kliknąć przycisk „*Exceptions”* (Wyjątki). Pojawi się wtedy okienko dialogowe, w którym można wpisywać rzeczone wyjątki.



Rysunek Wyszukiwanie słów pochodzenia obcego

Znalezienie wyrazów następuje po wciśnięciu przycisku „*Find”*. Z listy tokenów zostaną wybrane te spełniające wybrane reguły i nienależące do zbioru wyjątków. W pustych polach pojawi się całkowita liczba znalezionych wyrazów oraz ich procentowy udział w tekście. Przycisk „*Preview”* wyświetla okienko dialogowe z listą - rozkładem częstotliwości znalezionych słów.

Dodatkową funkcjonalnością jest możliwość dodawania wyjątków bezpośrednio w okienku wyników wyszukiwania wyrazów – możemy je usuwać za pomocą klawisza *„Delete”*. Zostaną one dodane do listy wyjątków, a wyniki wyszukiwania zostaną automatycznie zaktualizowane.

* + 1. Wyrazy spełniające wyrażenia regularne

Kolejną funkcjonalnością jest wyszukiwarka wyrazów spełniających dowolne wyrażenie regularne.



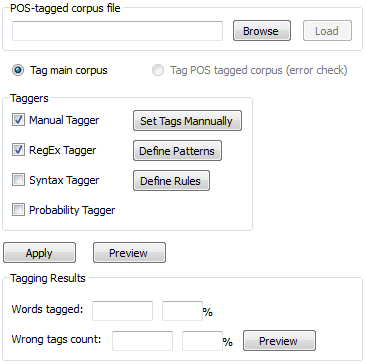
Rysunek Parametry wyszukiwania słów spełniających wyrażenie regularne

W polu „*Pattern”* wpisujemy interesujące nas wyrażenie regularne. Po wciśnięciu przycisku „*Find”*, wszystkie tokeny korpusu zostają sprawdzone pod kątem spełniania wyrażenia. Liczba wszystkich wystąpień i odsetek procentowy wyrazów są wyświetlane odpowiednich polach. Po kliknięciu przycisku „*Preview”*, w osobnym okienku zostaje wyświetlona lista znalezionych wyrazów.

* 1. Oznaczanie części mowy

Automatyczne określanie części mowy można zaimplementować na wiele sposobów, wykorzystując różne zjawiska językowe i statystyczne. W mojej aplikacji zaimplementowałam 4 rodzaje taggerów. W celu sprawdzenia ich skuteczności i poprawności istnieje możliwość wczytania wzorcowo otagowanego korpusu i porównania trafności przypisanych tagów.

Na potrzeby przypisywania tagów wyrazom użyłam kolejnej struktury danych przechowującej tokeny. Ponieważ jeden z taggerów wykorzystuje informacje o składni zdania (położenie przecinków i kropek), znaki przestankowe powinny być traktowane jako tokeny. Początkowo wszystkie tokeny mają przypisany tag domyślny – pusty ciąg znaków. Zestaw tagów używanych w aplikacji wraz z krótkim opisem został zamieszczony w załączniku A.



Rysunek Zakładka z narzędziami do określania części mowy

W kolejnych podrozdziałach opisane zostały zasady działania poszczególnych taggerów, analiza wyników, a także sposób, w jaki jest oceniana skuteczność taggerów.

* + 1. Tagger ręczny

Najprostszy tagger, polegający na ręcznym określeniu przynależności do części mowy danego wyrazu. Biorąc pod uwagę zasadę Pareto (będącą uogólnieniem prawa Zipfa), ręczne określenie części mowy dla stosunkowo niewielkiej ilości najczęściej występujących wyrazów może przynieść zaskakująco duże efekty. Nie zawsze jednak można jednoznacznie określić część mowy.

Ręczne określenie tagów odbywa się za pomocą przycisku „*Set Tags Manually”*. W okienku dialogowym pojawia się edytowalna lista słów, czytana z pliku o nazwie „*manualTaggingRules.txt”*. Po edycji listy w okienku następuje zapis do pliku o tej samej nazwie. W pliku tym każda linia zaczyna się tagiem, po którym następuje lista odpowiadających mu wyrazów, np:

ART l il i un una

ART jest symbolem części mowy – rodzajnika, a kolejne wyrazy to tokeny, które zostaną oznaczone jako rodzajniki.

Działanie taggera sprowadza się do przeparsowania wspomnianego pliku i stworzenia struktury – słownika, w której kluczem jest tag (część mowy), a wartością – zbiór tokenów. Następnie tagger iteruje po wszystkich tokenach i sprawdza ich przynależność do zbioru tokenów każdego taga. Jeśli token należy do zbioru, zostanie mu przypisany odpowiedni tag.

Zawartość pliku nie jest walidowana pod kątem opisania danego tokena różnymi tagami. Słownik tokenów jest tworzony w kolejności występowania w pliku, zatem token zostanie oznaczony takim tagiem, jaki pierwszy został przeczytany.

* + 1. Tagger wykorzystujący wyrażenia regularne

Podobnie jak w taggerze manualnym, reguły dla tagowania można edytować poprzez okno dialogowe, dostępne pod przyciskiem „*Define Patterns”*. Reguły są przechowywane w pliku o nazwie „*RegexpTaggingRules.txt*”. Składnia pliku również jest podobna – każda linia zaczyna się symbolem taga, po czym następuje wyrażenie regularne, którego spełnienie kwalifikuje token do otrzymania danego taga. Po wyrażeniu opcjonalnie mogą wystąpić wyjątki od reguły, np:

NOUN .+on[ei]$ buone buoni

NOUN jest symbolem rzeczownika. Wyrażenie regularne oznacza, że każdy token kończący się ciąg iem *–one* lub *–oni* zostanie uznany za rzeczownik, z wyjątkiem wyrazów *buone*, *buoni* (są to przymiotniki).

* + 1. Tagger składniowy

Do określenia taga danego wyrazu tagger składniowy wykorzystuje znane już tagi sąsiednich wyrazów lub wyrazy same w sobie. Reguły dla tego taggera znajdują się w pliku „*syntaxTaggingRules.txt”.* Edycja reguł jest taka sama jak w przypadku taggerów opisanych w 5.7.1 i 5.7.2. Pojedyncza reguła zawiera ciąg tagów lub wyrazów, z czego dokładnie jeden tag jest poprzedzony znakiem dolara – jest to szukany tag. Przykładowo, reguła:

ADJ $NOUN PUNCT

oznacza, iż wyraz następujący po przymiotniku i poprzedzający znak interpunkcyjny powinien otrzymać tag rzeczownika.

Analogicznie można również określić wyrazy, mające występować w sąsiedztwie, np. reguła:

si $VERB

oznacza, że każdy (jeszcze nieoznaczony tagiem) wyraz następujący po słowe *si* jest czasownikiem.

Każda reguła może zawierać ciąg tagów lub wyrazów o dowolnej długości. Podczas parsowania linii odpowiadającej danej regule, w przypadku nienapotkania na tag rozpoczynający się od znaku dolara, reguła ta jest pomijana.

Ponieważ działanie tego taggera opiera się między innymi na już przypisanych tagach, zastosowanie taggera z regułami wykorzystującymi tylko i wyłącznie sąsiednie tagi (w przeciwieństwie do wyrazów) nie przyniesie żadnych rezultatów. Dlatego też, jeśli do tagowania korpusu nie zostaną wybrane inne taggery (manualny lub oparty na wyrażeniach regularnych), reguły taggera składniowego muszą być sprawdzone pod kątem zawierania wyrazów bądź tagów. Jeśli nie zostanie dostarczona taka reguła, która opiera się tylko i wyłącznie na sąsiadujących wyrazach, tagowanie korpusu nie zostanie wykonane (pojawi się stosowny komunikat). Jeżeli taka reguła istnieje, musi mieć największy priorytet (być pierwszą sprawdzaną regułą) podczas działania taggera.

* + 1. Tagger prawdopodobieństwa

Ostatnim zaimplementowanym przeze mnie taggerem jest tagger wykorzystujący inny, już otagowany korpus. Warunkiem zastosowania tego taggera jest więc uprzednie załadowanie wzorcowo otagowanego korpusu – służą do tego specjalne przyciski, opisane w podrozdziale 5.7.6.

Na jego podstawie zostaje sporządzona mapa kojarząca każdy występujący w nim wyraz z listą tagów (i liczbą wystąpień), którymi został opisany. Następnie dla każdego wyrazu zostaje znaleziony tag, którym wyraz ten był tagowany najczęściej. Tak znaleziony zbiór par wyraz – tag zostaje użyty do otagowania właściwego korpusu. Oczywiście, nie wszystkie występujące w nim wyrazy muszą znajdować się w zbiorze, jednak zgodnie z wnioskami płynącymi z prawa Zipfa, znaczna większość wyrazów zostanie w ten sposób otagowana.

* + 1. Ocena rezultatów tagowania

Jakość rezultatów tagowania korpusu częściami mowy przez taggery – ich skuteczność oraz poprawność można poznać po każdym zastosowaniu dowolnego taggera.

Skuteczność rozumiana jako ogólny odsetek otagowanych słów jest prosta do określenia – wystarczy liczbę przypisanych tagów podzielić przez liczbę wszystkich tokenów i pomnożyć razy 100%. Wartość ta jest obliczana i wyświetlana po każdym zastosowaniu taggera.

Ocena poprawności, czyli informacja o ilości źle otagowanych wyrazów, wymaga naturalnie wiedzy o poprawnych tagach wyrazów w celu porównania ich z tymi przydzielonymi przez tagger. Dlatego, by poznać trafność taggerów, trzeba dokonywać tagowania na wzorcowo otagowanym korpusie, który można osobno załadować w aplikacji – proces ten jest opisany w podrozdziale 5.7.6. Po załadowaniu takiego korpusu należy wybrać opcję *„Tag POS tagged corpus (error check)”* (w przeciwieństwie do domyślnej opcji „*Tag main corpus*”). W ten sposób, oprócz informacji o skuteczności taggera, pojawiają się także statystyki dotyczące jego poprawności – liczba i odsetek źle przydzielonych tagów, oraz uaktywnia się przycisk „*Wrong Tags”* umożliwiający wyświetlenie okienka z listą wyrazów źle otagowanych (wraz z tagiem przydzielonym i wzorcowym).

Tak przygotowane narzędzia pozwalają na samodzielne opracowywanie i testowanie zasad tagowania – znajdowanie wyrażeń regularnych typowych dla części mowy, czy też zależności składniowych pomiędzy konkretnymi częściami mowy.

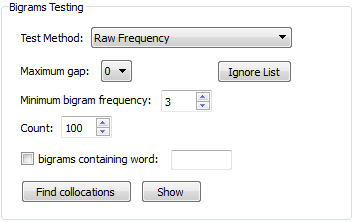
* + 1. Korpus wzorcowo otagowany częściami mowy

Dla celów opisanych w podrozdziałach 5.7.4 i 5.7.5 – użycie taggera prawdopodobieństwa i sprawdzanie poprawności działania taggerów - w programie można załadować korpus wzorcowo otagowany częściami mowy. Plik z takim korpusem musi posiadać określoną strukturę: w każdej linii znajduje się osobny wyraz, po którym następuje jeden ze ściśle określonej listy tagów (lista ta jest dostępna w załączniku A). Linia nie spełniająca tego wzorca zostanie pominięta.

Wczytany korpus jest przechowywany jako lista par token – tag. Przed określaniem poprawności taggerów lista tokenów jest kopiowana, a tagi ustawiane na domyślne (tagiem „domyślnym” jest pusty string – brak taga).

* 1. Wyszukiwanie kolokacji

W kolejnej zakładce zatytułowanej *„Collocations”* znajdują się narzędzia pozwalające na analizę bigramów występujących w korpusie i ich ocenę pod kątem nieprzypadkowości występowania. Można tego dokonać za pomocą szeregu metod i testów statystycznych, których opis znajduje się w rozdziale 3.2. Na rysunku niżej przedstawiona została zawartość zakładki dotyczącej wyszukiwania kolokacji. Poszczególne opcje zostaną omówione w następnych podrozdziałach.



Rysunek Zakładka poświęcona wyszukiwaniu kolokacji w tekście

* + 1. Przygotowanie listy bigramów

Przydatnym narzędziem do przygotowania bigramów do dalszej analizy jest klasa BigramCollocationFinder, z modułu *collocations* biblioteki NLTK. Najprostszym sposobem stworzenia obiektu tej klasy jest wywołanie funkcji *from\_words*, przyjmującej jako parametr listę słów korpusu. Funkcja ta sporządza dwa rozkłady częstotliwości: wyrazów i bigramów, potrzebne do zastosowania metod oceniających bigramy. Ja zdecydowałam się na ręczne przygotowanie tych rozkładów, ponieważ mogę przy tym wykorzystać dokonany uprzednio podział korpusu na zdania (w ten sposób uwzględnione zostaną tylko bigramy należące do jednego zdania) oraz filtrować bigramy pod kątem zawierania słów ignorowanych.

Użytkownik ma wpływ na tworzenie rozkładów poprzez podanie maksymalnej odległości między wyrazami tworzącymi bigram, minimalnej wymaganej częstotliwości występowania bigramów i opcjonalnie listy ignorowanych słów. Może również zaznaczyć opcję wyszukiwania tylko bigramów zawierających konkretny wyraz.

Wybranie dużej maksymalnej odległości między wyrazami może wiązać się z bardzo dużą złożonością obliczeniową, ale pozwala na wykrycie kolokacji, które są oddzielone zmienną liczbą wyrazów. Przykładowo, kolokacja *risolvere problema* (rozwiązać problem) może występować w następujących wariantach:

***risolvere*** *il* ***problema***

***risolvere*** *in qualche modo il* ***problema***

***risolvere*** *insieme un* ***problema***

***risolvere*** *con un suo amico un* ***problema***

Jak widać, słowa te mogą występować w dowolnej odległości od siebie, stanowią one jednak utartą kolokację wyrazową.

Ustawienie minimalnej częstotliwości występowania bigramów na wartość równą 1 powoduje, że niektóre testy stają się niemiarodajne – przypisują najwyższe oceny bigramom występującym tylko raz, przez co wyniki są zaburzone. Domyślnie wartość jest ustawiona na 3.

Lista ignorowanych wyrazów jest pobierana z pliku o nazwie „*words\_ignored\_in\_collocations.txt”*. Można ją zmienić w programie za pomocą przycisku „*Ignore list”*, który otwiera okienko dialogowe z listą wyrazów.

* + 1. Ocena bigramów korpusu

Użytkownik może wybrać jedną z 10 metod oceniania bigramów oraz liczbę kolokacji do wyświetlenia. Po kliknięciu przycisku „*Find Collocations”* następuje przygotowanie bigramów i zapis podanej liczby kolokacji do pola składowego modelu. Przy kolejnych wywołaniach, w przypadku, gdy zmienił się jeden z parametrów: wielkość okna, minimalna częstotliwość występowania, lista ignorowanych słów lub opcja wyszukiwania bigramów z podanym wyrazem, rozkład bigramów jest sporządzany od nowa. W przeciwnym razie rozkład jest aktualny i można zastosować metodę oceniania.

By poznać wynik wyszukiwań, należy kliknąć przycisk *„Show” –* pojawia się okienko dialogowe z tabelą zawierającą bigramy posortowane wg uzyskanego wyniku. W drugiej kolumnie znajduje się punktacja przyznana przez wybraną metodę, a w trzeciej – liczba wystąpień bigramu w korpusie.

* 1. Kontekst słów

Ostatnią zaimplementowaną przeze mnie funkcjonalnością jest wyszukiwarka kontekstu słowa w korpusie. Znajduje się ona w ostatniej zakładce zatytułowanej *„Context”*.



Rysunek Wyszukiwarka kontekstu wyrazów

Oprócz poszukiwanego słowa można także wybrać liczbę wyników oraz długość kontekstu (w słowach, w obie strony – szukany wyraz jest w środku kontekstu, zatem całkowita długość kontekstu będzie wynosić , gdzie *n* – zadana długość kontekstu).

Lista wyników jest budowana w oparciu o klasę ConcordanceIndex biblioteki NLTK. Podczas pierwszego użycia wyszukiwarki kontekstu dla bieżącego korpusu sporządzany jest indeks wszystkich wyrazów - wywoływany jest konstruktor wspomnianej klasy, biorący jako parametr listę wyodrębnionych tokenów korpusu i funkcję zmieniającą wszystkie litery na małe. Tak skonstruowany obiekt jest zapisywany dla dalszych wyszukiwań, jako pole składowe Modelu. Na utworzonym indeksie wywoływana jest metoda *offsets,* przyjmująca jako parametr poszukiwany wyraz i zwracająca listę pozycji, na których się on znajduje. Bazując na liście wyników, do poszukiwanego wyrazu doklejana jest odpowiednia ilość wyrazów sąsiednich. W ten sposób powstaje lista kontekstów, która zostaje wyświetlona w osobnym oknie. Dla czytelności, szukany wyraz jest pisany dużymi literami.

1. Przykład użycia

Niniejszy rozdział poświęcony jest szczegółowej analizie przypadku użycia aplikacji oraz wnioskom wynikającym z uzyskanych rezultatów. Do testów wykorzystane zostały korpusy tekstów pochodzących z blogów internetowych, prasy i serwisu społecznościowego Twitter (są to próbki zawierające ok. 2mln słów). Podane czasy obliczeń dotyczą komputera z procesorem Intel Core i3-3110M 2x2.40GHz i 4GB RAM.

* 1. Podstawowe informacje o korpusie

W tabeli nr 1 znajduje się zestawienie charakterystyk dla tekstów różnego pochodzenia.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Twitter | Prasa | Blogi |
| Liczba słów | 2008363 | 2009185 | 2003226 |
| Śr. długość słowa | 4.514 | 4.957 | 4.848 |
| Śr. długość zdania | 14.063 | 22.556 | 21.568 |
| Różnorodność leksykalna | 0.05106 | 0.04383 | 0.05297 |
| Ilość typów wyrazów | 102554 | 88066 | 106110 |
| Odsetek hapaksów [%] | 55.09 | 46.31 | 50.01 |
| Czas wczytywania [s] | 106.56s | 73.119s | 112.456s |

Tabela 1 Zestawienie charakterystyk dla tekstów różnego pochodzenia

Na uwagę zasługuje znacznie mniejsza średnia długość zdania dla tekstów z Twittera (jest to zrozumiałe zjawisko, biorąc pod uwagę styl wypowiedzi w Internecie oraz ograniczenie długości wiadomości na Twitterze do 140 znaków.

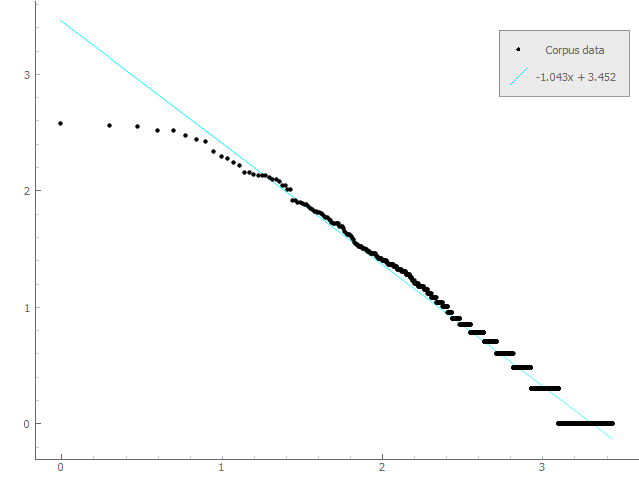
Kolejnym ciekawym spostrzeżeniem jest mniejszy udział hapaksów, mniejsze zróżnicowanie leksykalne tekstów prasowych i, co się z tym wiąże, mniejsza liczba typów wyrazów. Może to wynikać z większej dokładności przy redagowaniu tekstu (błędy literowe powodują zwiększenie ogólnej liczby typów wyrazów) oraz z mniejszej liczby zapożyczeń czy nazw własnych – te kwestie mogą być wyjaśnione przez sprawdzenie odsetku wyrazów pochodzenia obcego.

Można zauważyć, że większe zróżnicowanie leksykalne przekłada się na dłuższy czas wczytywania korpusu. Ma to z pewnością związek z operacjami na przechowywanym rozkładzie częstotliwości, którego rozmiar jest ściśle związany z różnorodnością leksykalną.

Dla tekstów prasowych średnia długość słowa jest nieznacznie większa. Może to wynikać z bardziej specjalistycznego, oficjalnego słownictwa używanego w prasie, w przeciwieństwie do prostych wypowiedzi pojawiających się na blogach i Twitterze.

* 1. Prawo Zipfa

Obliczenia zajmują ok. godzinę!



Rysunek Wykres punktowy rozkładu częstotliwości i wykres linii trendu

* 1. Wyszukiwanie wzorców

W następnych podpunktach omówię wyniki dokonanych wyszukiwań wzorców, z podziałem na wyrazy pochodzenia obcego oraz inne dowolne wyrażenia regularne.

* + 1. Wyrazy pochodzenia obcego

W Załączniku C zostały przedstawione słowa ignorowane przy poszukiwaniach wyrazów pochodzenia obcego. Są to głównie spójniki, przyimki, wyrazy ulegające zabiegowi elizji (np. bezokoliczniki pozbawione ostatniej litery ‘e’ – *aver*, *poter*) i zabiegowi apokopy (np. *mezz*, *anch*, *sant*) oraz zaasymilowane wyrazy obce, np. *film*).

W tabeli 2 znajduje się porównanie 20 najczęściej występujących słów uznanych za wyrazy pochodzenia obcego dla trzech różnych korpusów.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Blogi | | Prasa | | Twitter | |
| Wyraz | Ilość | Wyraz | Ilość | Wyraz | Ilość |
| 1 | blog | 1917 | inter | 1793 | twitter | 3120 |
| 2 | post | 990 | milan | 1592 | cmq | 1098 |
| 3 | com | 314 | gol | 1408 | com | 1063 |
| 4 | mail | 264 | club | 714 | tweet | 653 |
| 5 | ecc | 256 | champions | 575 | lol | 622 |
| 6 | internet | 218 | juventus | 547 | blog | 593 |
| 7 | art | 213 | sport | 415 | internet | 499 |
| 8 | super | 167 | premier | 412 | post | 474 |
| 9 | web | 156 | match | 399 | mail | 449 |
| 10 | etc | 150 | real | 397 | web | 444 |
| 11 | cmq | 141 | pdl | 379 | uff | 434 |
| 12 | and | 139 | leader | 342 | fan | 391 |
| 13 | leader | 123 | madrid | 317 | ipad | 390 |
| 14 | contest | 112 | fiat | 283 | social | 381 |
| 15 | blogger | 110 | benitez | 258 | mah | 377 |
| 16 | gas | 101 | manchester | 244 | justin | 363 |
| 17 | computer | 98 | team | 229 | mac | 343 |
| 18 | premier | 96 | van | 200 | inter | 325 |
| 19 | forum | 92 | assist | 185 | prof | 323 |
| 20 | prof | 86 | record | 179 | app | 320 |
| Ogółem | 55542 (2.77%) | | 80903 (4.03%) | | 122599 (6.1%) | |

Tabela Porównanie występowania wyrazów pochodzenia obcego

Można zauważyć spore różnice w ilości wwyrazów zaklasyfikowanych jako obce. Przedstawione listy znacznie się od siebie różnią – ma to związek z odmienną tematyką i stylistyką tekstów. Korpus pochodzący z prasy wyróżnia się dużą ilością nazwisk, nazw własnych i słownictwa sportowego zaczerpniętego z języka angielskiego, podczas gdy wiadomości z Twittera obfitują w żargon internetowy. Notki umieszczane na blogach charakteryzują się najmniejszym odsetkiem wyrazów pochodzenia obcego i są ściśle powiązane z tematyką blogową (wyrazy *blog*, *post*, *internet*, *mail*)

* + 1. Wyrażenia regularne

Wykorzystując zasady koniugacji czasowników włoskich i wyrażenia regularne można przeprowadzać proste wyszukiwania mające na celu np. porównanie zastosowania czasu przyszłego w tekście.

W tabeli nr 3 znajdują się wyniki wyszukiwań wyrażenia regularnego: *.\*rò$*, które spełniają wyrazy kończące się na –*rò.* Końcówka ta jest typowa dla czasowników pierwszej osoby w czasie przyszłym. Wyjątek stanowi wyraz *però*, który nie został uwzględniony w wynikach.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Blogi | | Prasa | | Twitter | |
| Wyraz | Ilość | Wyraz | Ilość | Wyraz | Ilość |
| 1 | farò | 165 | sarò | 48 | farò | 165 |
| 2 | sarò | 124 | farò | 34 | sarò | 124 |
| 3 | proverò | 107 | avrò | 18 | proverò | 107 |
| 4 | avrò | 82 | dovrò | 18 | avrò | 82 |
| 5 | dovrò | 60 | cercherò | 17 | dovrò | 60 |
| 6 | potrò | 51 | andrò | 14 | potrò | 51 |
| 7 | dirò | 46 | potrò | 13 | dirò | 46 |
| 8 | cercherò | 43 | dirò | 11 | cercherò | 43 |
| 9 | riuscirò | 42 | tornerò | 11 | riuscirò | 42 |
| 10 | andrò | 39 | entrò | 10 | andrò | 39 |
| 11 | tornerò | 34 | continuerò | 9 | tornerò | 34 |
| 12 | seguirò | 31 | deciderò | 9 | seguirò | 31 |
| 13 | continuerò | 28 | parlerò | 7 | continuerò | 28 |
| 14 | metterò | 27 | vedrò | 7 | metterò | 27 |
| 15 | vedrò | 25 | comprò | 6 | vedrò | 25 |
| 16 | pubblicherò | 23 | incontrerò | 6 | pubblicherò | 23 |
| 17 | posterò | 21 | metterò | 6 | posterò | 21 |
| 18 | parlerò | 20 | porterò | 6 | parlerò | 20 |
| 19 | racconterò | 19 | riuscirò | 6 | racconterò | 19 |
| 20 | entrò | 18 | tiferò | 6 | entrò | 18 |
| Ogółem | 1989 (0,1%) | | 546 (0,027%) | | 3217 (0,16%) | |

Tabela Porównanie występowania czasowników czasu przyszłego w pierwszej osobie

Można zaobserwować znaczną różnicę w występowaniu czasowników czasu przyszłego w pierwszej osobie. Teksty prasowe zawierają ich najmniej, poza tym badając kontekst słów (rozdział 6.6) można również stwierdzić, że są to głównie cytowane wypowiedzi. Największy odsetek tej formy czasownika występuje w wiadomościach z Twittera, co jest zrozumiałe ze względu na luźny styl wypowiedzi.

* 1. Określanie części mowy
  2. Wyszukiwanie kolokacji
  3. Podgląd kontekstu

1. Podsumowanie

Cele pracy określone w Rozdziale 1. zostały osiągnięte. Została opracowana aplikacja umożliwiająca analizę statystyczną języka naturalnego pod różnym kątami.

W szczególności zostały zrealizowane funkcjonalności: sporządzenie charakterystyki statystycznej tekstu, określanie części mowy poszczególnych słów za pomocą 4 konfigurowalnych metod z możliwością porównania wyników ze wzorcem (otagowanym poprawnie korpusem) oraz wyszukiwanie kolokacji w tekście korzystając z jednego z dziesięciu testów statystycznych. Dodatkowo, użytkownik ma możliwość wyszukiwania wyrazów spełniających podane wyrażenie regularne oraz wyszukiwania kontekstu słów. Zważywszy na specyfikę języka włoskiego, istnieje także możliwość oszacowania udziału w tekście słów pochodzenia obcego.

By efektywnie korzystać z aplikacji, nie jest wymagana wiedza techniczna, przydatna jest jednak znajomość tworzenia wyrażeń regularnych oraz specyfiki używanych testów statystycznych, w celu osiągnięcia najlepszych rezultatów i wyciągania poprawnych wniosków.

Przetwarzanie języka naturalnego jest bardzo szeroką dziedziną, zatem istnieje oczywiście wiele możliwości dalszego rozwoju takiej aplikacji, a oto niektóre z nich:

* wykorzystując zasady deklinacji czasowników włoskich i wprowadzając bardziej szczegółowe tagi części mowy można zaimplementować bardziej niezawodne metody określania części mowy;
* dużym udogodnieniem byłoby wprowadzenie parametru części mowy w wyszukiwarce kolokacji, tak, że użytkownik mógłby wyszukać tylko kolokacje będące parą: czasownik – rzeczownik, przymiotnik – rzeczownik, itd;
* stosując istniejące wykrywanie części mowy i dodając obsługę zewnętrznego słownika można zaimplementować automatycznego tłumacza z języka włoskiego.

1. Załącznik A

Lista tagów używanych w programie, symbolizujących części mowy

* NOUN - rzeczownik
* CONJ - spójnik
* ADJ - przymiotnik
* VERB - czasownik
* ADV - przysłówek
* PREP - przyimek
* ART - rodzajnik
* NUM - liczebnik
* PRON - zaimek

Następne dwa tagi zostały utworzone z myślą o wyrazach pełniących funkcję dwóch części mowy:

* DPREP – zaimek z rodzajnikiem określonym, np. *al, delle, sullo*
* PRONVERB – czasownik z zaimkiem osobowym, np. *dammi, svegliati*

Pozostałe tagi służą do oznaczania innych niż części mowy elementów tekstu przydatnych do określania części mowy.

* PUNCT – znaki interpunkcyjne, tj. kropki, nawiasy, cudzysłowy, myślniki
* SPECIAL – inne znaki mogące pojawić się w tekście (znak procentu, plus, itd.)

1. Załącznik B
   1. Reguły taggera manualnego

PUNCT . ? ! , ; : ... ( ) [ ] { } " -

SPECIAL / % = +

ART l il i un

CONJ e ed o se ma mentre oppure sebbene quantunque allorché perciò anziché allorquando

NUM due tre quattro cinque sette otto ott nove dieci venti vent trenta trent

PREP a ad di d su in per con senza contro dietro dentro fra tra fuori accanto dopo oltre attorno attraverso durante

PRON io mi me tu ti te lei lui si s noi voi li c esso chi niente qualcosa qualcuno qualcuna cui ciò ognuno ognuna quelli quelle

DPREP al alla ai agli alle all allo del della dei degli delle dell dello nel nella nei nelle negli nell nello dal dalla dalle dall dallo dagli sul sulla sui sulle sull sugli sullo

ADJ ex mio mia mie miei tuo tua tue tuoi suo sua sue suoi nostro nostra nostri nostre vostro vostra vostri vostre alcuno alcuna alcune qualche qualunque

ADV poi qui già là sempre mai adesso ormai allora più piu quasi meno non molto tanto forse invece ancora anche anch pur così pure ieri oggi domani poi infine soprattutto tuttavia purtroppo nemmeno affatto abbastanza almeno pertanto soltanto intanto comunque circa inoltre addirittura

VERB sono è siamo siete ho hai ha abbiamo avete hanno

* 1. Reguły taggera opartego na wyrażeniach regularnych

NUM .\*\d.\*

ADV .+mente$

NOUN .{2,}[ae][nz]za$

NOUN .{2,}tà$

NOUN .+aggio?$

NOUN .+ism[io]$

NOUN .+on[ei]$ buone buoni

NOUN .+or[ei]$

VERB .+[ae]ndo$ quando fernando comando contrabbando

VERB .+amo$

VERB .+rà$

PRONVERB .+[ae]r[mcsvl]i$ diversi merci parla parli

PRONVERB .+[ae]r[mtscv]e[ln][aeio]$ diversi

PRONVERB .+[ae]rglie[ln][aeio]$ diversi

PRONVERB .{3,}[ae]ndo[mtscv]e[ln][aeio]$

PRONVERB .{3,}[ae]ndo[mtscv]i$

ADJ .{3,}issim[oaie]$

ADJ .{2,}ile$

ADJ .{2,}os[oa]$

* 1. Reguły taggera składniowego

ART $NOUN PREP

ART $NOUN DPREP

ADJ $NOUN PUNCT

DPREP $NOUN PUNCT

DPREP $NOUN DPREP

si $VERB

ART $NOUN VERB

1. Załącznik C

Lista słów włoskich spełniających reguły wyszukiwań wyrazów pochodzenia obcego:

|  |
| --- |
| 1. ad |
| 1. ah |
| 1. al |
| 1. alcun |
| 1. all |
| 1. anch |
| 1. aver |
| 1. beh |
| 1. bel |
| 1. ben |
| 1. boh |
| 1. buon |
| 1. c |
| 1. cin |
| 1. col |
| 1. con |
| 1. cos |
| 1. d |
| 1. dal |
| 1. dall |
| 1. del |
| 1. dell |
| 1. dev |
| 1. don |
| 1. dov |
| 1. ed |
| 1. eh |
| 1. est |
| 1. far |
| 1. film |
| 1. fin |
| 1. gliel |
| 1. il |
| 1. in |
| 1. l |
| 1. lasciar |
| 1. m |
| 1. maggior |
| 1. mal |
| 1. mezz |
| 1. miglior |
| 1. n |
| 1. neanch |
| 1. neh |
| 1. nel |
| 1. nell |
| 1. non |
| 1. nord |
| 1. oh |
| 1. ovest |
| 1. peggior |
| 1. per |
| 1. press |
| 1. pur |
| 1. qual |
| 1. quant |
| 1. quatt |
| 1. quel |
| 1. quell |
| 1. quest |
| 1. s |
| 1. sant |
| 1. saper |
| 1. signor |
| 1. sud |
| 1. sul |
| 1. sull |
| 1. t |
| 1. tal |
| 1. tutt |
| 1. un |
| 1. vent |
| 1. vien |
| 1. voler |
| 1. vuol |

1. Bibliografia

Bird, S., Klein, E. i Loper, E. (2009). *Natural Language Processing with Python.* Sebastopol: O’Reilly Media, Inc.

ChengXiang, Z. (2007, kwiecień). Statistical Language Models for Information Retrieval. Urbana-Champaign, Illinois, United States.

*Dokumentacja biblioteki NLTK*. (2001-2011). Pobrano z lokalizacji http://nltk.googlecode.com/svn/trunk/doc/api/frames.html

*Dokumentacja biblioteki PyQt4*. (2014). Pobrano z lokalizacji http://pyqt.sourceforge.net/Docs/PyQt4/

*Dokumentacja biblioteki pyqtgraph*. (2011). Pobrano z lokalizacji http://www.pyqtgraph.org/documentation/apireference.html

Fayyad, U., Piatetsky-Shapiro, G. i Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases.

Manning, C. D. i Schuetze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, Massachusetts: The MIT Press.

Pawłowski, A. (2001). *Metody kwantytatywne w sekwencyjnej analizie tekstu.* Warszawa: Uniwersytet Warszawski.

Quasthoff, U. i Wolff, C. (2002). The Poisson Collocation Measure and its Applications. *Second International Workshop on Computational Approaches to Collocations*.

Wittgenstein, L. (1958). *Dociekania filozoficzne.* Oksford: Blackwell Publishers Ltd.

Zawadzka, D. (1993). *Gramatyka języka włoskiego.* Warszawa: PWN.