**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.

Contents

[1. Cel pracy 4](#_Toc380586065)

[2. Przetwarzanie języka naturalnego – wprowadzenie 5](#_Toc380586066)

[2.1. Odkrywanie wiedzy w bazie danych (KDD) 5](#_Toc380586067)

[2.2. Statystyczna analiza języka 7](#_Toc380586068)

[2.2.1. Uczenie maszynowe 7](#_Toc380586069)

[2.2.2. Eksploracja danych 8](#_Toc380586070)

[2.3. Główne zastosowania NLP 9](#_Toc380586071)

[2.4. Problemy NLP 11](#_Toc380586072)

[3. Język włoski – krótka charakterystyka pod kątem analizy statystycznej 12](#_Toc380586073)

[4. Implementacja rozwiązania 13](#_Toc380586074)

[4.1. Środowisko programistyczne 13](#_Toc380586075)

[5. Przykłady użycia 14](#_Toc380586076)

1. Cel pracy

Celem pracy jest opracowanie narzędzi informatycznych wspomagających analizę statystyczną języka włoskiego.

Jest to aplikacja napisana w języku Python, przy użyciu biblioteki graficznej PyQt. Analizy dokonuje się wczytując plik zawierający próbkę tekstu w języku włoskim.

2-3 strony

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 skupiam się na języku 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 1 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, itd. 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 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:

* wieloznaczna składnia zdania – zwłaszcza długie, wielokrotnie złożone zdania dopuszczają mnogość interpretacji. W wielu 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.
* 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*). W języku polskim takie wieloznaczności rodzi np. rzeczownik w funkcji przydawki lub rzeczownik mogący spełniać zarówno funkcję podmiotu jak i dopełnienia, np. *Samochód wyprzedził motocykl*. Problemy tego typu świadczą o błędach językowych tekstu i nie tylko komputer miałby trudności z ich interpretacją.
* niejednoznaczność słów – wiele wyrazów ma więcej niż jedno znaczenie, więc tylko kontekst, w jakim dane słowo występuje, może pomóc w określeniu jego znaczenia. „*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 może być położony na

Podsumowując,

Therefore, a practical NLP system must be good at making disambiguation decisions of word sense, word category, syntactic structure, and semantic scope.

1. Język włoski – krótka charakterystyka pod kątem analizy statystycznej
2. 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 QtCreatora (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. Przykłady użycia