**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](#_Toc380414077)

[2. Przetwarzanie języka naturalnego – wprowadzenie 6](#_Toc380414078)

[2.1. Odkrywanie wiedzy w bazie danych (KDD) 6](#_Toc380414079)

[2.2. Statystyczna analiza języka 8](#_Toc380414080)

[2.2.1. Uczenie maszynowe 8](#_Toc380414081)

[2.2.2. Eksploracja danych 9](#_Toc380414082)

[2.3. Zastosowania NLP 10](#_Toc380414083)

[3. Implementacja rozwiązania 11](#_Toc380414084)

[3.1. Środowisko programistyczne 11](#_Toc380414085)

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, pary 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 tekstu będącego zapisem rozmów telefonicznych 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 – utworzenie 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 znaczenie. Abstrakt powstaje poprzez utworzenie semantycznej reprezentacji danych i wygenerowanie na jej podstawie tekstu w języku naturalnym.
* tłumaczenie
* podział morfologiczny – wyodrębnienie z tekstu i klasyfikacja morfemów (najmniejszej jednostki gramatycznej). Kluczowy wpływ ma tu złożoność morfologiczna (struktura wyrazów) danego języka.
* generowanie języka naturalnego – przekształcanie danych zapisanych w bazach danych do formy czytelnej dla człowieka. Ważnym zadaniem jest dobór odpowiednich słów, poprawność gramatyczna i naturalność.
* rozumienie języka naturalnego –
* rozróżnianie części mowy –
* wykrywanie kolokacji

Kolejnym szerokim zastosowaniem jest wykrywanie kolokacji, czyli utartych zwrotów składających się ze słów występujących razem. Można w ten sposób śledzić trendy w języku (związki frazeologiczne ulegają modzie i

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 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