**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](#_Toc379446204)

[2. Środowisko programistyczne 5](#_Toc379446205)

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.

1. Analiza statystyczna 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.

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:

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.

Eksploracja danych (data mining) to ogół metod wykrywających wzorce w dużych zbiorach danych z wykorzystaniem sztucznej inteligencji, statystyki, hurtowni danych. 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 techniki eksploracji danych:

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. 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. Interfejs aplikacji