Politechnika Warszawska

WYDZIAŁ ELEKTRYCZNY

Instytut Elektrotechniki Teoretycznej
i Systemów Informacyjno-Pomiarowych
Zakład Elektrotechniki Teoretycznej
i Informatyki Stosowanej

Praca dyplomowa inżynierska

na kierunku Informatyka
w specjalności Inżynieria oprogramowania

Badanie algorytmów do porównywania stylów tekstów w języku polskim

Szymon Masłowski
nr albumu 271070

promotor
dr inż. Grzegorz Sarwas

WARSZAWA 2023

Badanie algorytmów
do porównywania stylów tekstów w języku polski

Streszczenie

Praca składa się z krótkiego wstępu jasno i wyczerpująco opisującego oraz uzasadniającego cel pracy, trzech rozdziałów (2-4) zawierających opis istniejących podobnych rozwiązań, komponentów rozpatrywanychjako kandydaci do tworzonego systemu i wreszcie zagadnień wydajności wirtualnych rozwiązań. Piąty rozdział to opis środowiska obejmujący opis konfiguracji środowiska oraz przykładowe ćwiczenia laboratoryjne. Ostatni rozdział pracy to opis możliwości dalszego rozwoju projektu.

Opisać co się zawiera, proces myślowy. Jakie metody i co wyszło(ale nie mocno rozpisywać) **TODO**Napisac to = **Słowa kluczowe:** NLP, Klasyfikacja, Analiza stylu, Przetwarzanie języka naturalnego, eksploracja tekstu

TODO Przetłumaczyć to

THESIS TITLE

Abstract

This thesis presents a novel way of using a novel algorithm to solve complex problems of filter design. In the first chapter the fundamentals of filter design are presented. The second chapter describes an original algorithm invented by the authors. Is is based on evolution strategy, but uses an original method of filter description similar to artificial neural network. In the third chapter the implementation of the algorithm in C programming language is presented. The fifth chapter contains results of tests which prove high efficiency and enormous accuracy of the program. Finally some posibilities of further development of the invented algoriths are proposed.

Keywords: thesis, LaTeX, quality

WARSZAWA, 12 Grudnia 2012

POLITECHNIKA WARSZAWSKA
WYDZIAŁ ELEKTRYCZNY

OŚWIADCZENIE

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa inżynierska pt. Badanie algorytmów do porównywania stylów tekstów w języku polskim:

- została napisana przeze mnie samodzielnie,
- nie narusza niczyich praw autorskich,
- nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam, że przedłożona do obrony praca dyplomowa nie była wcześniej podstawą postępowania związanego z uzyskaniem dyplomu lub tytułu zawodowego w uczelni wyższej. Jestem świadom, że praca zawiera również rezultaty stanowiące własności intelektualne Politechniki Warszawskiej, które nie mogą być udostępniane innym osobom i instytucjom bez zgody Władz Wydziału Elektrycznego.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Szymon Masłowski.....

Spis treści

Podziękowania

Dziękujemy bardzo serdecznie wszystkim, a w szczególności Rodzinom i Unii Europejskiej...

Zdolny Student i Pracowity Kolega

Rozdział 1

Wstęp

Od kiedy człowiek zapisuje swoje myśli na świecie utrwalane jest coraz więcej danych. Pisane są listy, książki i krótkie wiadomości. Każdy z tych dzieł ma swojego autora, choć nie zawsze jest on znany. Niemniej dzięki uważnej obserwacji można ustalić kto jest autorem tekstu, jeżeli posiadamy inne do porównania. Pisanie, podobnie jak mowa czy każde działanie, które człowiek podejmuje, jest pod wpływem osobistego zachowania–stylu. Na podstawie sprawdzania stylu można przypisać autorstwo starożytnych tekstów, których autorzy mogli być zmieniani, bądź pominięci, albo odnaleźć autora anonimowego terrorystycznego manifestu. Celem tej pracy jest poruszenie tematu takiego automatycznego rozpoznawania tekstów, które napisane są w języku polskim. Tworzy to specyficzne warunki, w odróżnieniu od np. języka angielskiego. Chciałbym sprawdzić czy i z jaką pewnością można określić autora tekstu dodając go do zbioru kilku znanych autorów.

Opisać organizację pracy **TODO**powołać się na jakąś bibliografię

Rozdział 2

Wprowadzenie do problematyki porównywania tekstów

2.1 NLP

Pojęciem NLP (ang. Natural Language Processing) określa się zbiór technik komputerowych służących do analizy i reprezentacji tekstów występujących na poziomie analizy lingwistycznej w celu uzyskania sposobu przetwarzania języka przypominającego ludzki w określonym zakresie zadań i zastosowań. - Soldacki 'https://pl.wikipedia.org/wiki/Przetwarzanie_języka_naturalnego'

2.2 Budowa zdania

Tekst składa się ze znaków, tworzących słowa, składających się na zdania. Za Słownikiem Języka Polskiego, zdanie to

1. «myśl wyrażona słowami»
2. "«zespół wyrazów powiązanych zależnościami gramatycznymi i zawierający orzeczenie»"

'https://encenc.pl/budowa-zdania/' 'https://pl.wikipedia.org/wiki/Szyk_wyrazów'

2.3 Klasyfikacja języków naturalnych

2.4 Styl tekstu

Każdy człowiek jest inny, ze względu na swoje DNA, środowisko i temperament. To wszystko rzutuje na jego działania, zarówno rodzaj jak i sposób ich wykonania. Każda czynność, którą wykonuje człowiek obarczona podpisem wykonawcy, który jest mniej lub bardziej wyraźny. Niemniej czynności przez nas wykonywane są robione na nasz własny sposób. Tekst, który zapisujemy jest odbiciem naszego sposobu myślenia, więc rzutuje to na przykład na użyte słownictwo, czy sposób budowania zdań. Cechy, które chcę użyć w pracy to:

- kolejność części mowy

  Dzięki temu, że język polski jest językiem fleksyjnym, kolejność słów może ulegać znacznej zmianie, bez zmiany znaczenia zdania. Więc stała kolejność części zdania może być cechą charakterystyczną
- zbiór słów - wektor binarny?

  Ile różnych słów występuje, jak bogate słownictwo jest wykorzystywane. IDF do porównania jak bardzo różne od codziennego słownictwa jest to użyte w tekście.
- złożoność zdań - ile takich zdań występuje?

  Czy zdania są proste, czy wielokrotnie złożone. Czy dzieje się to często, czy każde ze zdań jest takie.
- Statystyka zdania/słów
  1. Ile słów/zdanie
  2. Jak długie słowa(znaki/sylaby)
  3. Ile słów krótkich $<= 3$ znaki
  4. Ile słów długich $>= 7$ znaków
  5. Monosylabowe słowa (1 sylaba)
  6. Wielosylabowe słowa ($>= 3$ sylaby)
  7. Najdłuższe zdanie (ilość słów)
  8. Najdłuższe słowo (ilość znaków, ilość sylab)
- N-gramy literowe i słowne N-gram jest ciągiem n elementów. Tworząc 2-gramy ze zdania *Niezmiennik pętli jest techniką dowodzenia poprawności algorytmów.* Otrzymamy następujące 2-gramy:
  - Niezmiennik – pętli
  - pętli – jest

  - jest – techniką
  - etc. W przypadku n-gramów literowych, zamiast słów używamy następujących po sobie liter. Ta cecha może mieć znaczenie w przypadku, kiedy autor ma jakieś problemy z wymową i naturalnie przez to nie będzie używał pewnego zakresu słów.

- (Odwrotna) Częstotliwość słów
  Mówi nam o tym, jak szerokiego zasobu słów używa autor tekstu i w oparciu o korpus wzorcowy, jak ma się to do zwyczajów piśmienniczych danego czasu. Odwrotna częstotliwość słów (zwana dalej z ang. IDF –*inversed document frequency*)

- Kwestia podstawowych form i przyrostów (archaizmy mają inne żostki")

- Indeksy czytelności tekstu
  Te cechy są pośrednimi wskaźnikami, które będą ilościowym wyznacznikiem czy tekst jest "łatwyćzy "trudny"w odbiorze i czytaniu.

  1. SMOG
     `'https://en.wikipedia.org/wiki/SMOG'`
  2. Indeks czytelności Flescha-Kincaida
     `'https://en.wikipedia.org/wiki/Flesch-Kincaid_readability_tests'`
  3. Linsear Write
     `'https://en.wikipedia.org/wiki/Linsear_Write'`
  4. Lix
     `'https://en.wikipedia.org/wiki/Lix_(readability_test)'`

2.5 Analiza morfologiczna

https://core.ac.uk/download/pdf/11337572.pdf - Bień

2.6 Analiza syntaktyczna

`'http://bazhum.muzhp.pl/media//files/Prace_Jezykoznawcze/Prace_Jezykoznawcze-r2008-t10/Prace_Jezykoznawcze-r2008-t10-s187-200/Prace_Jezykoznawcze-r2008-t10-s187-200.pdf'` ??

2.7 Klasyfikacja

In classification, inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one or more (multi-label classification) of these classes. This is typically tackled in a supervised way. Spam filtering is an example of classification, where the inputs are email (or other) messages and the classes are śpamąnd ńot spam". - Richard O. Duda, Peter E. Hart, David G. Stork (2001) Pattern classification (2nd edition), Wiley, New York, ISBN 0-471-05669-3. Z wiki M

Rozdział 3

Przegląd dostępnych rozwiązań

3.1 Rozprawa doktorska z EITI

3.2 Jednolity System Antyplagiatowy

Jednolity system antyplagiatowy jest w posiadaniu Ministra właściwego do spraw szkolnictwa wyższego, a jego budowa i administracja realizowana jest przez Ośrodek Przetwarzania Informacji - Państwowy Instytut Badawczy pod nadzorem Ministra właściwego do spraw szkolnictwa wyższego. System jest wymaganym elementem sprawdzającym czy praca dyplomowa jest samodzielna: jest plagiatem czy nie. Co ciekawe, autorzy przypominają na każdym kroku, że "Wynik badania antyplagiatowego nie stanowi ostatecznego rozstrzygnięcia czy praca dyplomowa jest plagiatem czy też nie.".

Rysunek 3.1: Infografika pokazująca proces użycia JSA

System generuje wynik na podstawie płaszczyzn i cech prezentowanych

w grupach wg obszarów, których dotyczą:

- **Analiza tekstu** jest pierwszym obszarem sprawdzenia pracy. Polega głównie na sprawdzeniu czy w tekście nie pojawiają się białe znaki i nie zostały wykorzystane inne techniki, mające na celu zaburzyć automatyczne przetworzenie tekstu i jednoczesnej neutralności dla ludzkiego czytelnika.
  Na te techniki składają się znaki białe, bądź pochodzące z innego języka a graficznie będące podobne do języka pracy. Te techniki zmieniają długości słów, przez co porównania do innych ćzystych"tekstów są nieefektywne. Na podstawie porównania rozkładu długości i częstości słów z danymi z Ogólnopolskiego Repozytorium Pisemnych Prac Dyplomowych można znajdować przesłanki, że nastąpiły takie manipulacje. Dodatkowo System bada spójność ogólnopojętego stylu tekstu i oczekuje, że co najmniej 70% tekstu będzie napisana w takim samym stylu. Niestety nie jest udostępnione na jakiej podstawie badany jest styl.
- **Procentowy Rozmiar Podobieństwa** jest wskaźnikiem, który pokazuje jak duża część pracy składa się z fragmentów pochodzących z innych tekstów. Proponowane są 4 rozmiary zbiorów porównawczych: dla ciągów słownych nie dłuższych niż 5, 10, 20 i 40 słów. Dodatkowo tworzona jest lista źródeł (w przypadku, gdy fragment zostanie gdzieś odnaleziony).

3.3 Otwarty System Antyplagiatowy

3.4 Bank ING i czytelność dokumentóW

3.5 Użyte technologie

3.5.1 Python

3.5.2 Biblioteki

Rozdział 4

Model klasyfikacji

4.1 Warunki

4.1.1 Ile użytych tekstów/autorów

4.2 Metody porównywania obiektów

4.2.1 SVM

4.3 Budowa i opis algorytmu

4.3.1 Przygotowanie danych

4.3.2 Struktura danych

4.3.3 Cechy

4.3.4 Algorytmy użyte do ekstrakcji cech

4.3.5 Klasyfikacja

Rozdział 5

Testy i wyniki

5.1 Przypadki testowe

5.2 Cel testów

5.3 Wyniki

Rozdział 6

Wyniki

6.1 Czy było warto?

6.2 Podjęte decyzje

Klasyfikator

Do rozwiązania zagadnienia klasyfikacji można użyć wielu rozwiązań, takich jak: sieci neuronowe, klasyfikator Bayesowski, drzewo decyzyjne, metoda najbliższych sąsiadów czy w końcu SVM. Ze względu na rozmiar pracy, postanowiłem nie sprawdzać wpływu różnych klasyfikatorów na dokładność wyniku. Jak pokazują inne prace, użycie tego klasyfikatora zapewniało najdokładniejsze wyniki.

TODObibliografia wyliczenia

TODObibliografia, że SVM jest ok do tekstów

Kolejna decyzja

6.3 Co należało by poprawić?

Dodatek A

Pierwszy dodatek

Bibliografia

[1] W. R. Stevens, G. R. Wright, „Biblia TCP/IP tom 1”, RM, 1998.

Opinia

o pracy dyplomowej magisterskiej wykonanej przez dyplomanta
Zdolnego Studenta i Pracowitego Kolegę
Wydział Elektryczny, kierunek Informatyka, Politechnika Warszawska
Temat pracy
TYTUŁ PRACY DYPLOMOWEJ

Promotor: **dr inż. Miły Opiekun**
Ocena pracy dyplomowej: **bardzo dobry**

Treść opinii

Celem pracy dyplomowej panów dolnego Studenta i Pracowitego Kolegi było opracowanie systemu pozwalającego symulować i opartego o oprogramowanie o otwartych źródłach (ang. Open Source). Jak piszą Dyplomanci, starali się opracować system, który łatwo będzie dostosować do zmieniających się dynamicznie wymagań, będzie miał niewielkie wymagania sprzętowe i umożliwiał dalszą łatwą rozbudowę oraz dostosowanie go do potrzeb. Przedstawiona do recenzji praca składa się z krótkiego wstępu jasno i wyczerpująco opisującego oraz uzasadniającego cel pracy, trzech rozdziałów (2-4) zawierających opis istniejących podobnych rozwiązań, komponentów rozpatrywanychjako kandydaci do tworzonego systemu i wreszcie zagadnień wydajności wirtualnych rozwiązań. Piąty rozdział to opis przygotowanego przez Dyplomantów środowiska obejmujący opis konfiguracji środowiska oraz przykładowe ćwiczenia laboratoryjne. Ostatni rozdział pracy to opis możliwości dalszego rozwoju projektu. W ramach przygotowania pracy Dyplomanci zebrali i przedstawili w bardzo przejrzysty sposób duży zasób informacji, co świadczy o dobrej orientacji w nowoczesnej i ciągle intensywnie rozwijanej tematyce stanowiącej zakres pracy i o umiejętności przejrzystego przedstawienia tych wyników. Praca zawiera dwa dodatki, z których pierwszy obejmuje wyniki eksperymentów i badań nad wydajnością, a drugi to źródła skryptów budujących środowisko.

Dyplomanci dość dobrze zrealizowali postawione przed nimi zadanie, wykazali się więc umiejętnością zastosowania w praktyce wiedzy przedstawionej w rozdziałach 2-4. Uważam, że cele postawione w założeniach pracy zostały pomyślnie zrealizowane. Proponuję ocenę bardzo dobrą (5).

(data, podpis)

Recenzja

pracy dyplomowej magisterskiej wykonanej przez dyplomanta
Szymona Masłowskiego
Wydział Elektryczny, kierunek Informatyka, Politechnika Warszawska
Temat pracy
Badanie algorytmów
do porównywania stylów tekstów w języku polski

Recenzent: **prof. nzw. dr hab. inż. Jan Surowy**
Ocena pracy dyplomowej: **bardzo dobry**

Treść recenzji

Celem pracy dyplomowej panów dolnego Studenta i Pracowitego Kolegi było opracowanie systemu pozwalającego symulować i opartego o oprogramowanie o otwartych źródłach (ang. Open Source). Jak piszą Dyplomanci, starali się opracować system, który łatwo będzie dostosować do zmieniających się dynamicznie wymagań, będzie miał niewielkie wymagania sprzętowe i umożliwiał dalszą łatwą rozbudowę oraz dostosowanie go do potrzeb. Przedstawiona do recenzji praca składa się z krótkiego wstępu jasno i wyczerpująco opisującego oraz uzasadniającego cel pracy, trzech rozdziałów (2-4) zawierających bardzo solidny i przejrzysty opis: istniejących podobnych rozwiązań (rozdz. 2), komponentów rozpatrywanych jako kandydaci do tworzonego systemu (rozdz. 3) i wreszcie zagadnień wydajności wirtualnych rozwiązań, zwłaszcza w kontekście współpracy kilku elementów sieci (rozdział 4). Piąty rozdział to opis przygotowanego przez Dyplomantów środowiska obejmujący opis konfiguracji środowiska oraz przykładowe ćwiczenia laboratoryjne (5 ćwiczeń). Ostatni, szósty rozdział pracy to krótkie zakończenie, które wylicza także możliwości dalszego rozwoju projektu. W ramach przygotowania pracy Dyplomanci zebrali i przedstawili w bardzo przejrzysty sposób duży zasób informacji o narzędziach, Rozdziały 2, 3 i 4 świadczą o dobrej orientacji w nowoczesnej i ciągle intensywnie rozwijanej tematyce stanowiącej zakres pracy i o umiejętności syntetycznego, przejrzystego przedstawienia tych wyników. Drobne mankamenty tej części pracy to zbyt skrótowe omawianie niektórych zagadnień technicznych, zakładające dużą początkową wiedzę czytelnika i dość niestaranne podejście do powołań na źródła. Utrudnia to w pewnym stopniu czytanie pracy i zmniejsza jej wartość dydaktyczną (a ta zdaje się być jednym z celów Autorów), ale jest zrekompensowane zawartością merytoryczną. Praca zawiera dwa dodatki, z których pierwszy obejmuje wyniki eksperymentów i badań nad wydajnością, a drugi to źródła skryptów budujących środowisko. Praca zawiera niestety dość dużą liczbę drobnych błędów redakcyjnych, ale nie wpływają one w sposób istotny na na jej czytelność

i wartość. W całej pracy przewijają się samodzielne, zdecydowane wnioski Autorów, które są wynikiem własnych i oryginalnych badań. Rozdział 5 i dodatki pracy przekonują mnie, że Dyplomanci dość dobrze zrealizowali postawione przed nimi zadanie. Pozwala to stwierdzić, że wykazali się więc także umiejętnością zastosowania w praktyce wiedzy przedstawionej w rozdziałach 2-4. Kończący pracę rozdział szósty świadczy o dużym (ale moim zdaniem uzasadnionym) poczuciu własnej wartości i jest świadectwem własnego, oryginalnego spojrzenia na tematykę przedstawioną w pracy dyplomowej. Uważam, że cele postawione w założeniach pracy zostały pomyślnie zrealizowane. Proponuję ocenę bardzo dobrą (5).

(data, podpis)