# WEDT - DOKUMENTACJA WSTĘPNA

Piotr Doniec

Grudzień 2011

## 1 Cel projektu, zadanie

Dany jest zbiór dokumentów tekstowych (czysty tekst oraz HTML). Należy dokonać podziału zbioru na grupy, tak aby dokumenty należące do pojedynczej grupy były jak najbardziej zbliżone tematycznie do siebie, a jednocześnie odmienne od dokumentów w pozostałych grupach. Grupy mogą (ale nie muszą) - tworzyć strukturę hierarchiczną.

## 2 Algorytm

1. Wczytanie dokumentu do analizy
2. Utworzenie w pamięci 2 kopii dokumentu: bez i z znacznikami HTML. Funkcja usuwająca HTML jest zaimplementowana w NLTK.
3. Usunięcie wszystkich znaków specjalnych, interpunkcji z przetwarzanego dokumentu przy pomocy prostego wyrażenia regularnego.
4. Usunięcie spójników i innych słów nie mających wpływu na treść dokumentu (ang. stopwords). Biblioteka NLTK zawiera listę takich wyrazów dla kilku języków. Dla języka angielskiego twórcy biblioteki przewidzieli 127 wyrazów.
5. Stemming tekstu, zliczenie wystepujących wyrażeń (terms). Wykorzystany jest stemmer Portera. Do prowadzenia statystyk wyrazów wykorzystany jest typ FreqDist umożliwiający łatwe uaktualnianie statystyki a także dostęp do liczby i częstotliwości wystąpienia wyrażenia bez konieczności przeprowadzania dodatkowych obliczeń.
6. Wykorzystanie kopii zawiarającej znaczniki HTML do lepszej oceny treści. Z treści zadania wynika, że dokumenty mogą zawierać znaczniki HTML. Można wykorzystać tę własność w celu poznanania potencjalnie istotnych słów dla dokumentu. Znaczniki HTML można w łatwy sposób poszeregować względem ważności. Im ważniejszy znacznik, tym istotniejsze są słowa które zawiera. W chwili obecnej program uwzględnia tylko znacznik <TITLE></TITLE> i <H1></H1>. Wartościowanie

znaczników odbywa się poprzez zwiększenie liczby wystąpień słowa o pewną stałą zadaną jako parametr.

7. Obliczenie dla każdego dokumentu współczynnika tf-idf zgodnie ze wzorem:
   $(tf - idf)_{i,j} = tf_{i,j} * \log(\frac{|D|}{|d:t_i \in d|})$
   gdzie,
   $tf_{i,j}$ - częstotliwość występowania wyrażenia 'i' w dokumencie 'j'
   $|D|$ - liczba przetwarzanych dokumentów
   $|d : t_i \in d|$ - liczba dokumentów w których występuje wyrażenie 'i'

8. Grupowanie dokumentów na podstawie współczynnika tf-idf. Aktualnie wykorzystywany jest algorytm KMeans oraz odległość Euklidesowa. Próby wykorzystania podobieństwa cosinusowego zakończyły się błędami - wysoce prawdopodobne, że to błąd w bibliotece.

# 3 Narzędzia

Projekt został zaimplementowany w języku Python. Wynika to z łatwości języka i ilością dostępnych bibliotek zapewniających stopień abstrakcji umożliwiający skupienie się na rozwiązaniu problemu bez konieczności implementacji algorytmów składowych.

Głównym ogniwem jest przedstawiona na wykładzie biblioteka nltk zapewniająca większość wymaganej funkcjonalności wykorzystanej w projeckie. Z poziomu nltk możliwa jest między innymi tokenizacja tekstu, dostęp do korpusów oraz stemming wyrazów. Zaimplementowane są także najpopularniejsze funkcje obliczające podobieństwo między dokumentami oraz algorytmy grupowania. Dodatkowo wykoszystano bibliotekę lxml która umożliwia w łatwe i szybkie parsowanie dokumentów HTML.

Kod projekt jest objęty systemem kontroli wersji i dostępny pod adresem: http://github.com/pejotr/doc-clustering

# Literatura


[1] A. Huang, *Similarity Measures for Text Document Clustering*, Department of Computer Science, The University of Waikato, Hamilton, New Zealand

[2] M. Steinbach, G. Karypis, V. Kumar, *A Comparison of Document Clustering Techniques*, Department of Computer Science and Egineering, University of Minnesota

[3] C. Manning, H. Schütze, *Foundations of Statistical Natural Language Processing*,MIT Press, 1999