Karol Łyskawiński 137335
Jakub Słota 137405

Inteligentne Wyszukiwanie Informacji – sprawozdanie z projektu

1. Temat projektu

Tematem realizowanego projektu jest **ekstrakcja fraz kluczowych z korpusów tekstowych**. Wydobywanie wyrazów bądź fraz o kluczowym znaczeniu dla danego korpusu tekstowego pozwala na zwięzłe opisanie jego zawartości. Frazy kluczowe znajdują wiele zastosowań w przetwarzaniu języka naturalnego, służą m. in. do kategoryzacji dokumentów, ich klasteryzacji oraz streszczania. Przyjęto założenie, że korpus tekstowy został napisany w języku angielskim.

2. Zastosowany algorytm

W celu realizacji projektu zaimplementowano algorytm **TextRank**. Jest to przykład algorytmu niewymagającego nadzoru, więc działającego bez zbioru uczącego. TextRank wyznacza ranking najlepszych fraz kluczowych poprzez zastosowanie algorytmu PageRank na odpowiednio przygotowanym grafie. Wierzchołkami grafu są pewne jednostki badanego korpusu tekstowego, a krawędzie stanowi miara podobieństwa pomiędzy tymi jednostkami. Za jednostki tekstowe uznaje się unigramy, które można połączyć w wielowyrazowe frazy już po uzyskaniu rankingu. Natomiast miarę podobieństwa określa wspólne występowanie wyrazów w oknie mieszczącym N wyrazów, typowo 2-10.

W pierwszym kroku algorytmu tworzony jest zbiór kandydatów na wyrazy kluczowe. Wybierane są tokeny wyrazów wyszukane w tekście. Następnie odfiltrowywane są tokeny należące do zbioru stop-words w j. angielskim, będące znakami interpunkcyjnymi lub stanowiące inne części mowy, niż rzeczownik i przymiotnik.

W drugim kroku powstaje graf kandydatów na wyrazy kluczowe. Wierzchołki są tworzone ze zbioru unikatowych tokenów uzyskanych w pierwszym kroku. Krawędzie występujące między wierzchołkami są nieskierowane i nieważone. Krawędź jest tworzona, jeżeli w zbiorze kolejnych kandydatów wyrazy sąsiadowały bezpośrednio ze sobą, zatem okno sąsiedztwa mieści N=2 wyrazów.

Na powstałym w kroku drugim grafie uruchamiany jest algorytm PageRank celem utworzenia rankingu kandydatów na wyrazy kluczowe. Część najlepszych kandydatów w rankingu, wyznaczona na podstawie empirycznego progu sumy wag, przechodzi dalej.

Na podstawie zbioru najlepszych kandydatów oraz zbioru wszystkich wyrazów w tekście tworzony jest zbiór ostatecznych fraz kluczowych. Występujące blisko siebie wyrazy kluczowe (do 4 wyrazów w przód) w zbiorze wszystkich wyrazów łączy się we frazy kluczowe, których waga jest normalizowana względem liczby tworzących je wyrazów.

Wszystkie wyrazy wchodzące w skład inicjalnego zbioru kandydatów na wyrazy kluczowe, jak i zbioru wszystkich wyrazów, są ujednolicane poprzez zamianę wielkich liter na małe oraz lematyzację. Dzięki temu powtórzenie wyrazu w liczbie mnogiej lub zapis wielką literą nie zaburza wyników uzyskiwanych przez algorytm.

Po wyznaczeniu wszystkich fraz kluczowych następuje selekcja najlepszych pozycji do wypisania na wyjściu programu. W tym kroku wybierane są takie najwyżej punktowane frazy, których suma wag nie przekracza określonego przez użytkownika progu.

Na koniec możliwe jest wykonanie klasteryzacji znalezionych fraz, czyli zgrupowania ich wokół powtarzających się słów. Warunkiem zaklasyfikowania frazy do klastra jest występowanie w niej danego wyrazu. Pod uwagę nie jest brana długość frazy ani kolejność jej słów, a każda fraza może należeć jednocześnie do kilku klastrów.

Przewidziano możliwość porównywania dokumentów powiązanych z głównym dokumentem. Porównywanie zaczyna się od ekstrakcji fraz kluczowych z poszczególnych dokumentów pobocznych. Następnie frazy kluczowe dokumentu głównego i pobocznych są sprowadzane do zbiorów wyrazów. Podobieństwo dokumentu głównego i pobocznych określa się przez cosinus kąta pomiędzy wektorami złożonymi z uzyskanych zbiorów wyrazów.

3. Zastosowane narzędzia

Projekt został zrealizowany w języku programowania Python w wersji 2.7. Podjęto taką decyzję ze względu na ekspresyjność samego języka, jak i spektrum dostępnych bibliotek pomocnych w implementacji algorytmu TextRank. Oprócz standardowego API Pythona wykorzystano następujące biblioteki:

- networkx: biblioteka służąca do tworzenia oraz przeprowadzania operacji na sieciach i grafach, zawiera implementację algorytmu PageRank;
- wikipedia: biblioteka ułatwiająca korzystanie z API Wikipedii celem wyszukiwania artykułów oraz pobierania ich treści;
- nltk: biblioteka dostarczająca zestaw narzędzi związanych z przetwarzaniem języka naturalnego, m. in. do lematyzacji, tokenizacji, tagowania części mowy.

4. Funkcjonalność

Zaimplementowany program pozwala na wydobywanie najlepszych fraz kluczowych z tekstu pochodzącego z wybranego przez użytkownika źródła. Do obsługiwanych rodzajów wejścia należą:

a) Wikipedia – korpus tekstu stanowi artykuł lub zbiór artykułów o tytułach podanych przez użytkownika, oddzielonych przecinkiem, np. „Linux, Python (programming language)”;
b) Pojedynczy plik – korpusem tekstu jest plik tekstowy znajdujący się pod wskazaną ścieżką;
c) Katalog – program tworzy korpus tekstu odczytując wszystkie pliki umieszczone w katalogu znajdującym się pod wskazaną ścieżką.

Na wyjściu programu podawane są najlepsze frazy kluczowe wraz z przypisaną do nich wagą. Mierzony jest czas wykonania algorytmu ekstrakcji fraz kluczowych.

Program jest też w stanie wyznaczyć słowa będące klastrami i przypisać do nich znalezione frazy. Zwrócone klastry posortowane są po liczbie należących do nich elementów.

Możliwe jest wybranie opcji szukania dokumentów podobnych do głównego, określanego mianem „master”. W przypadku wejścia będącego artykułem z Wikipedii analizowane są artykuły, do których prowadzą linki z artykułu głównego. Ze względu na liczność linków – typowo kilkaset – jest to długotrwały proces. W przypadku pojedynczego pliku wejściowego, program przeanalizuje pozostałe pliki zawarte w tym samym katalogu, co „master”. Zależnie od wybranego rodzaju wejścia, na wyjściu zostaną wypisane tytuły artykułów na Wikipedii lub ścieżki do plików na dysku, mających podobną zawartość do głównego dokumentu wraz z ich miarą podobieństwa.

5. Rezultaty

Ocena, które wyrazy i frazy najlepiej podsumowują korpus tekstowy, jest trudna. Poniżej wypisano ranking fraz kluczowych uzyskanych dla artykułu na Wikipedii o tytule „Linux”. Wydają się być intuicyjnie zgodne ze spodziewanymi w artykule zagadnieniami.

```
linux                          : 0.0094071668
linux system                   : 0.0076900915
linux distribution             : 0.0067519241
system                         : 0.0059730162
desktop linux                  : 0.0056895815
linux desktop                  : 0.0056895815
linux kernel                   : 0.0056502344
linux server                   : 0.0055545326
support linux                  : 0.0054426424
linux support                  : 0.0054426424
linux application              : 0.0054268801
linux user                     : 0.0053541777
linux component                : 0.0053293873
linux version                  : 0.0053103274
ubuntu linux                   : 0.0052454574
linux market                   : 0.0052265695
use linux                      : 0.0052251001
linux community                : 0.0050665312
many linux distribution        : 0.0050452610
linux name                     : 0.0050065084
name linux                     : 0.0050065084
linux distribution support     : 0.0049939887
linux foundation               : 0.0049552020
free linux distribution        : 0.0049536272
trademark linux                : 0.0049157827
fedora linux                   : 0.0048988209
hat linux                      : 0.0048946157
linux derivative               : 0.0048837674
linux gaming                   : 0.0048799490
```

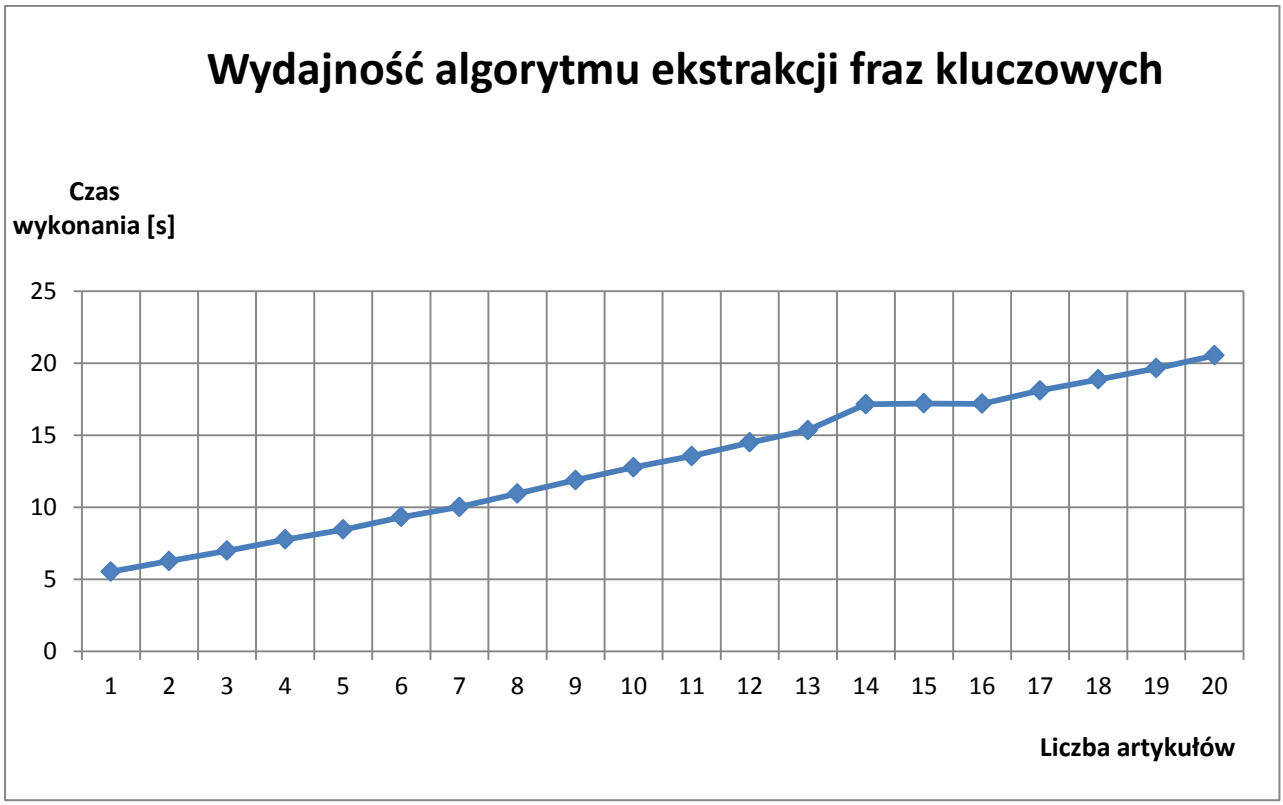
```
suse linux                          : 0.0048799055
linux installed                     : 0.0048753873
year linux                          : 0.0048736831
linux vendor                        : 0.0048682914
linux focus                         : 0.0048506184
enterprise linux                    : 0.0048383517
several linux distribution          : 0.0048372351
customized linux                    : 0.0048201513
```

6. Wydajność

Przeprowadzono testy wydajnościowe zaimplementowanego programu. Jako zbiór danych wejściowych posłużyły manualnie wybrane artykuły na Wikipedii dotyczące informatyki. Wybrane artykuły są rozbudowane. Charakteryzują się podobną długością i posiadaniem przynajmniej kilku wieloakapitowych sekcji.

Pomiary wykonano na laptopie z zainstalowanym systemem operacyjnym Windows 7 64-bit oraz 4-rdzeniowym procesorem Intel i7-3612 QM.

Czasy pracy algorytmu od 1 do 20 artykułów zmierzono 5 razy dla zmiennej kolejności artykułów oraz uśredniono. Wyniki pomiarów prezentuje wykres na rys. 1.



Rys. 1. Wykres wydajności algorytmu ekstrakcji fraz kluczowych.