SPRAWOZDANIE Z ZADANIA PROJEKTOWEGO Z PRZEDMIOTU METODY EKSPLORACJI W ODKRYWANIU DANYCH

Adam Nowik

Nr albumu: 210602

SPIS TREŚCI

1 DEFINICJA PROBLEMU

1.1 TEMAT

Grupowanie wsparte ontologią.

Opracowanie systemu tworzącego ontologię służącą do grupowania tekstów.

1.2 ZAKRES ROZWIĄZANIA PROBLEMU

Zadanie projektowe polegało na opracowaniu metod budowania ontologii opartej na analizie słów tekstów w języku angielskim, która mogłaby być później wykorzystana do grupowania teksów.

Stworzenie gotowego systemu grupującego teksty w zbiory o podobnej tematyce znajduje się poza zakresem projektu.

2 OPIS ROZWIĄZANIA

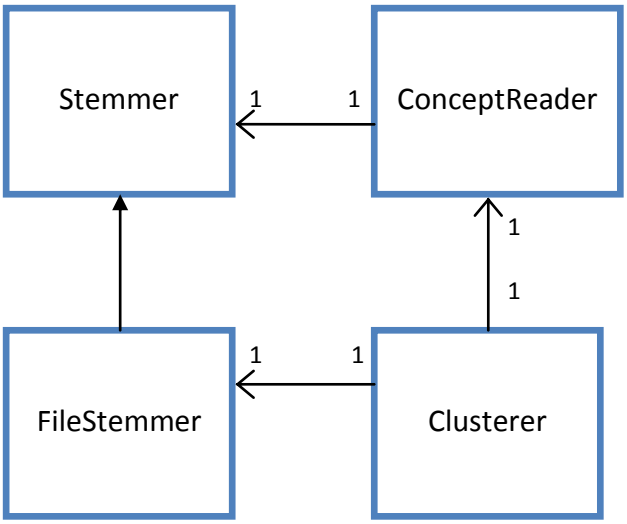
2.1 GŁÓWNE ELEMENTY ROZWIĄZANIA

Rozwiązanie zostało podzielone na następujące elementy:

- Stemmer tekstów – służący do tworzenia form podstawowych wyrazów w analizowanych dokumentach, w projekcie wykorzystano gotowy stemmer tekstów dostępny pod adresem http://www.tartarus.org/~martin/PorterStemmer wykorzystujący Porter Stemming Algorithm
- Czytnik konceptów – służący do stworzenia zbioru konceptów oraz wprowadzania hierarchii pomiędzy poszczególnymi konceptami. Dane na podstawie których czytnik tworzy hierarchię i zbiór konceptów zostały zaczerpnięte z bazy danych WordNet dla windows (część dotycząca rzeczowników w języku angielskim) dostępnej pod adresem: http://wordnet.princeton.edu/
  Czytnik zawiera zaimplementowany parser plików WordNet.
- Element wyznaczający słowa kluczowe tekstów – który z form podstawowych dostarczonych przez stemmer oraz hierarchii konceptów dostarczonej przez czytnik konceptów wyznacza słowa kluczowe najlepiej opisujące analizowany dokument.

3 IMPLEMENTACJA

3.1 DIAGRAM KLAS



3.2 STEMMER

Klasa implementująca stemmer portera – tworząca z wyrazów w języku angielskim ich formy podstawowe co pozwala na grupowanie różnych form tego samego wyrazu.

3.3 FILESTEMMER

Klasa dziedzicząca po klasie Stemmer – umożliwiająca łatwe zastosowanie stemmera do pracy z plikami tekstowymi. W trakcie analizy wskazanego pliku obiekt tej klasy tworzy tablicę hashującą zawierającą formy podstawowe słów znajdujących się w analizowanym pliku wraz z ich licznością.

3.4 CONCEPTREADER

Klasa implementująca parser plików WordNet – przystosowany do analizy plików zawierających spis rzeczowników. Obiekty tej klasy analizują wskazany plik glosariusza WordNet i na jego podstawie tworzą zbiór tablic hashujących zawierający:

- Tablicę konceptów w pełnej formie wraz z ich numerami referencyjnymi
- Tablicę konceptów w formie podstawowej (na których zastosowano algorytm stemmera) wraz z numerami referencyjnymi (takimi samymi jak w zbiorze konceptów „pełnych”)
- Tablicę synonimów zawierającą listę synonimów każdego konceptu
- Posortowany zbiór zawierający wsparcie konceptów – określone jako liczbę występowania danego konceptu w listach synonimów innych konceptów
- Tablicę hierarchii konceptów określonej jako zbiór dwójek <koncept, rodzic>.

3.4.1 BUDOWANIE HIERARCHII

Budowanie hierarchii odbywa się w sposób następujący:

- Z posortowej listy konceptów wybierane są kolejno koncepty o największym wsparciu:
    - Jeśli wybrany koncept C nie znajduje się w tablicy hierarchii jest tam umieszczany jako korzeń hierarchii to znaczy nadawany jest mu kod rodzica „0” - <nr ref. C, 0>
    - Dla każdego z synonimów konceptu wykonywane jest:
        - Jeśli wybrany synonim S nie znajduje się w tablicy hierarchii to jest tam umieszczany z numerem referencyjnym konceptu C jako numerem rodzica - <nr ref. S, nr ref. C>

Jeśli w pliku WordNet ten sam wyraz występuje w kilku różnych znaczeniach (związanych np. z kontekstem) każde ze znaczeń otrzymuje inny numer referencyjny, w związku z czym hierarchia zbudowana na numerach referencyjnych konceptów uwzględnia informację o znaczeniu danego wyrazu.

3.5 CLUSTERER

3.5.1 TWORZENIE LEKSYKONU TERMINÓW

Klasa implementująca główny obiekt budujący ontologię, obiekt tej klasy wykorzystuje obiekt klasy FileStemmer w celu analizy plików tekstowych z dokumentami i dostarczenia tablic zawierających posortowaną pod względem częstości występowania listę form podstawowych zgrupowanych według nazwy dokumentu, z którego terminy te pochodzą. Zbiór ten traktowany jest jako leksykon terminów.

3.5.2 OBLICZANIE WARTOŚCI FUNKCJI TFIDF

Obiekt klasy Clusterer przechowuje tablicę hashującą zawierającą częstość występowania danych form podstawowych w poszczególnych dokumentach. Na jej podstawie za pomocą algorytmu TES[1][2] wyznaczana jest wartość funkcji tfidf dla dokumentu mówiąca o istotności formy podstawowej liczonej jako iloczyn częstości występowania terminu w dokumencie i logarytmu z ilorazu liczby dokumentów oraz ilości dokumentów w jakiej występuje dany termin. Dzięki własnościom tej funkcji usuwane są terminy występujące zbyt często (we wszystkich dokumentach) oraz występujące bardzo rzadko, wychwytywane są natomiast terminy często występujące w dużym podzbiorze właściwym grupy analizowanych dokumentów.

3.5.3 OBLICZANIE W

Dodatkowo dla całej grupy tekstów obliczana jest wartość funkcji W będąca sumą wartości funkcji tfidf dla wszystkich dokumentów, funkcja ta wskazuje na istotność danego terminu w całej grupie dokumentów. Następnie terminy są sortowane malejąco pod względem wartości funkcji W.

3.5.4 WYZNACZANIE SŁÓW KLUCZOWYCH

Wyznaczanie słów kluczowych polega na mapowaniu n najbardziej istotnych terminów (pod względem wartości funkcji W dla grupy dokumentów, lub funkcji tfidf dla pojedynczego dokumentu) na koncepty znajdujące się w zbiorze konceptów dostarczonych przez obiekt klasy ConceptReader. Po odnalezieniu konceptu odpowiadającego terminowi następuje wędrówka do konceptu będącego korzeniem w hierarchii konceptów – w ten sposób znajdowany jest najbardziej powszechny synonim terminu – staje się on słowem kluczowym dokumentu bądź grupy dokumentów.

4 INSTRUKCJA OBSŁUGI

W katalogu z plikiem .jar powinien znajdować się plik data.noun zawierający glosariusz – aplikacja nie będzie działać prawidłowo bez pliku glosariusza.

Program powinien być uruchamiany za pomocą wiersza poleceń poleceniem:

```
java -jar clusterer.jar  -c licznik_slow_kluczowych
lista_plikow_lub_ciezka_do_katalogu
```

Parametr `-c licznik_slow_kluczowych` jest parametrem opcjonalnym i pozwala na określenie liczby form podstawowych terminów znaczących jakie system będzie określał dla każdego z analizowanych tekstów. Jeśli parametr –c nie zostanie podany, system będzie określał po 5 słów kluczowych dla tekstów.

Parametr `lista_plikow_lub_ciezka_do_katalogu` może zawierać dowolną liczbę ścieżek względnych lub bezwzględnych do katalogów lub plików tekstowych rozdzielonych spacjami. W przypadku, gdy parametr jest ścieżką do katalogu, system wykona analizę wszystkich plików z rozszerzeniem txt znajdujących się wewnątrz wskazanego katalogu. Możliwe jest mieszanie ścieżek względnych i bezwzględnych do katalogów oraz plików w dowolnej kolejności np.:

```
java -jar clusterer.jar –c 10 . ..\text8.txt C:\text9.txt
```

Spowoduje przeanalizowanie wszystkich plików znajdujących się w bieżącym katalogu (parametr ‘.’), pliku text8.txt znajdującym się w katalogu wyżej oraz pliku o ścieżce C:\text9.txt, dla każdego analizowanego tekstu zostanie określonych do 10 form podstawowych terminów znaczących – do których zostaną dopasowane słowa kluczowe.

Jeśli w ścieżce do pliku znajdują się spacje należy podać ją w cudzysłowie.

Jeśli program zostanie uruchomiony bez parametrów nastąpi jego natychmiastowe zamknięcie z informacją o braku plików do przeanalizowania.

Interpretacja wyników działania programu:

`Document .\text7.txt` – nazwa analizowanego dokumentu (może zawierać ścieżkę)
`Keywords:` - słowa kluczowe

```
differenti->differential->quality, differenti->differentiation->basic cognitive
process, differenti->differentiator->person, differenti->differentiation-
>mathematical process, argentin->argentine->soft-finned fish, tax->tax->tax,
option->option->option, option->option->decision making, export->exporter-
>industrialist, export->export->fabric, export->exporting->commercial enterprise,
```

Słowa kluczowe wyświetlane wg wzoru <forma podstawowa -> koncept odpowiadający formie podstawowej -> słowo kluczowe> dzięki takiemu sposobowi wyświetlania łatwo można zorientować się ile konceptów odpowiada wersji podstawowej słowa (po zastosowaniu stemowania algorytmem portera – zobacz rozdział 6) i w związku z tym, ile słów kluczowych zostało wybranych do wyznaczonego z tekstu terminu.

5 TESTY/EKSPERYMENTY

Testy polegające na wyznaczeniu słów kluczowych dla poszczególnych tekstów wykonano z wykorzystaniem 7 różnych tekstów wybranych z korpusu Reutersa 21578, na koniec wykonano próbę wyznaczenia słów kluczowych dla grupy wszystkich siedmiu tekstów. Wszystkie testy wykonano z parametrem wyznaczenia 5 terminów najbardziej istotnych dla analizowanego tekstu (bądź grupy tekstów).

Wyznaczone terminy istotne oraz odpowiadające im słowa kluczowe będą prezentowane wg wzoru

Termin -> słowo kluczowe

5.1 TEST 1

Tekst poddany analizie dotyczył kontrataku Stanów zjednoczonych na platformę wiertniczą w zatoce perskiej w Iranie.

Terminy znaczące wyznaczone przez system:

```
diplomat, gulf, respons, attack, platform
```

Terminy oraz odpowiadające im słowa kluczowe wyznaczone przez system:

```
diplomat->diplomat->mediator, diplomat->diplomat->person, gulf->gulf->disparity,
gulf->gulf->lake, gulf->gulf->angle, respons->response->consequence, respons-
>response->term, respons->response->property, respons->responsiveness->quality,
respons->responsiveness->sensitivity, respons->responsibility->responsibility,
attack->attack->motion, attack->attack->organic process, attack->attack->activity,
attack->attack->crime, attack->attack->attack, attack->attack->criticism, attack-
>attack->fire, attack->attacker->wrongdoer, platform->platform->platform,
platform->platform->system, platform->platform->structure, platform->platform-
>legal document
```

5.2 TEST 2

Tekst poddany analizie dotyczył strajku kanadyjskiego związku pracowników fabryk firmy General Motors

Terminy znaczące wyznaczone przez system:

```
white, worker, union, negoti, issu
```

Terminy oraz odpowiadające im słowa kluczowe wyznaczone przez system:

```
white->whiteness->chromatic color, white->whiting->saltwater fish, white->white-
>white, white->whiting->soft-finned fish, white->white->ball, white->whiting-
>saltwater fish, white->whiting->percoid fish, white->whiting->percoid fish,
worker->worker->person, worker->worker->insect, union->union->organization, union-
>union->collection, union->union->sound, union->unionization->motion, union-
>union->fabric, union->union->organization, union->unionism->artistic movement,
union->union->organic process, union->union->condition, union->union->motion,
negoti->negotiation->activity, negoti->negotiator->person, negoti->negotiation-
```

>saying, issu->issue->store, issu->issue->commercial enterprise, issu->issue->doctrine, issu->issue->activity, issu->issue->book

5.3 TEST 3

Tekst poddany analizie dotyczył konsekwencji kontrataku stanów zjednoczonych na platformę wiertniczą Iranu w odwecie za atak na amerykański tankowiec.

Terminy znaczące wyznaczone przez system:

expert, flag, gulf, attack, east

Terminy oraz odpowiadające im słowa kluczowe wyznaczone przez system:

expert->expert->person, expert->expertness->condition, flag->flag->structure, flag->flagging->collection, flag->flagging->walk, flag->flag->fabric, flag->flag->signal, flag->flag->flag, gulf->gulf->disparity, gulf->gulf->lake, gulf->gulf->angle, attack->attack->motion, attack->attack->organic process, attack->attack->activity, attack->attack->crime, attack->attack->attack, attack->attack->criticism, attack->attack->fire, attack->attacker->wrongdoer, east->east->compass point, east->east->compass point, east->east->region

5.4 TEST 4

Tekst poddany analizie dotyczył komputeryzacji jednego z banków narodowych.

Terminy znaczące wyznaczone przez system:

comput, servic, financi, person, base

Terminy oraz odpowiadające im słowa kluczowe wyznaczone przez system:

comput->computer->device, servic->serviceability->quality, servic->service->activity, servic->services->position, servic->service->position, servic->service->transportation, servic->service->service, servic->service->religious ceremony, servic->service->tableware, servic->service->activity, servic->service->company, servic->servicing->organic process, servic->service->activity, financi->financier->financier, person->personableness->quality, person->personality->quality, person->person->collection, person->personality->person, person->person->person, person->person->human body, person->personal->article, base->base->number, base->base->device, base->base->fabric, base->base->region, base->base->quality, base->base->device, base->base->apparatus, base->base->region, base->base->structure, base->baseness->quality, base->base->boundary, base->base->object, base->base->acid

5.5 TEST 5

Tekst poddany analizie dotyczył ustaleń zarządu spółki Nertheast Savings odnośnie sprzedaży akcji firmy.

Terminy znaczące wyznaczone przez system:

share, right, northeast, compani, exercis

Terminy oraz odpowiadające im słowa kluczowe wyznaczone przez system:

```
share->share->legal document, share->share->assets, share->sharing->activity,
share->sharing->relation, share->sharing->relation, right->right->assets, right-
>right->clique, right->rightness->quality, right->right->point, right->right-
>structure, right->right->honesty, right->right->turn, right->right->right,
northeast->northeast->compass point, northeast->northeaster->physical phenomenon,
northeast->northeast->compass point, northeast->northeast->region, compani-
>company->company, compani->company->condition, compani->company->organization,
compani->company->gathering, compani->company->military unit, compani->company-
>organization, exercis->exercise->exercise, exercis->exercise->school assignment,
exercis->exercise->activity, exercis->exercise->ceremony
```

5.6 TEST 6

Tekst poddany analizie dotyczył spotkania ministrów finansów w sprawie spadku wartości dolara opisywanym przez prezesa West German Savings bank association.

Terminy znaczące wyznaczone przez system:

```
dollar, damag, west, german, financ
```

Terminy oraz odpowiadające im słowa kluczowe wyznaczone przez system:

```
dollar->dollar->signal, dollar->dollar->coin, dollar->dollar->bill, dollar-
>dollar->monetary unit, damag->damages->compensation, damag->damage->sound, damag-
>damage->motion, damag->damage->sound, west->west->compass point, west->west-
>region, west->west->compass point, german->germaneness->relation, financ-
>finance->discipline, financ->finance->group action, financ->financing->commercial
enterprise, financ->finance->commercial enterprise
```

5.7 TEST 7

Tekst poddany analizie dotyczył sporów spowodowanych przez podwyższenie przez stany zjednoczone podatku na eksport ziarna soi do Argentyny.

Terminy znaczące wyznaczone przez system:

```
different, argentin, tax, option, export
```

Terminy oraz odpowiadające im słowa kluczowe wyznaczone przez system:

```
differenti->differential->quality, differenti->differentiation->basic cognitive
process, differenti->differentiator->person, differenti->differentiation-
>mathematical process, argentin->argentine->soft-finned fish, tax->tax->tax,
option->option->option, option->option->decision making, export->exporter-
>industrialist, export->export->fabric, export->exporting->commercial enterprise,
```

5.8 TEST 8

Wyznaczenie słów kluczowych dla wszystkich tekstów.

Terminy znaczące wyznaczone przez system:

Share, right, northeast, diplomat, compani

Terminy oraz odpowiadające im słowa kluczowe wyznaczone przez system:

share->share->legal document, share->share->assets, share->sharing->activity, share->sharing->relation, share->sharing->relation, right->right->assets, right->right->clique, right->rightness->quality, right->right->point, right->right->structure, right->right->honesty, right->right->turn, right->right->right, northeast->northeast->compass point, northeast->northeaster->physical phenomenon, northeast->northeast->compass point, northeast->northeast->region, diplomat->diplomat->mediator, diplomat->diplomat->person, compani->company->company, compani->company->condition, compani->company->organization, compani->company->gathering, compani->company->military unit, compani->company->organization

6 WNIOSKI

Z wyniku przeprowadzonych testów wynika, że algorytm wyboru terminów znaczących oparty na obliczaniu funkcji tfidf[1] w połączeniu ze stemmerem portera dobrze radzi sobie w wyborze słów dobrze określających charakter analizowanego tekstu.

Dobrym przykładem są tutaj teksty dotyczące ataków wojskowych na obiekty rafinerii naftowej w zatoce perskiej. Algorytm dla tych tekstów za znaczące uznał terminy takie jak: `diplomat, gulf, response, attack, platform` oraz `expert, flag, gulf, attack, east.` Wszystkie terminy, poza terminem `flag` odnoszą się do analizowanego tekstu i pozwalają stworzyć dobre wrażenie jego streszczenia. Po przeczytaniu listy terminów znaczących łatwo można zorientować się w tematyce tekstu.

Również dla innych tekstów terminy wybierane są w większości trafnie wyjątkami są wyrazy takie jak `base` w teście czwartym lub `right` w teście piątym słowa zostały zakwalifikowane jako istotne z powodu ich częstego występowania – lecz pojęcia, których dotyczą są zbyt abstrakcyjne aby dobrze określić analizowany tekst.

Należy pamiętać, że terminy wyznaczane przez stemmer sprowadzane są do form podstawowych w początkowym etapie analizy – w związku z tym czasem trudno jest odgadnąć faktyczne znaczenie wyrazu. Na przykład termin `white` w teście drugim może być formą podstawową stworzoną przez stemmer z wyrazów takich jak `whiteness, whiting, white.`

Z tego samego powodu listy słów kluczowych wydłużają się – synonimy wyznaczane są na podstawie formy podstawowej słowa – może się zdarzyć, że mapowany jest na nie więcej niż jeden koncept. Drugim powodem nadmiarowych i często niecelnych słów kluczowych wyznaczanych jako najbardziej ogólne synonimy jest brak badania kontekstu wybieranych terminów – w plikach WordNet, będących glosariuszem konceptów występują te same wyrazy w różnym znaczeniu opisane innymi synonimami, w efekcie w hierarchii konceptów występują różni rodzice dla tego samego wyrazu. Ponieważ podczas wyboru synonimów do form podstawowych pozbawionych kontekstu nie sposób określić ich znaczenia, wybierane są synonimy dla każdego znanego znaczenia wyrazu.

Warto zaznaczyć, że takie działanie może powodować istotne błędy, co pokazuje test 7, w którym słowo Argentina zostało zmienione na argentin co błędnie zostało zmapowane na termin soft-finned-fish.

Być może sposobem na poprawienie zaistniałej sytuacji byłoby wybieranie trójek bądź par terminów sąsiednich i próba wyznaczenia dla nich znaczenia kontekstowego, a następnie zmapowania ich na odpowiedni koncept jednakże powyższe rozwiązanie wykracza poza zakres określony w projekcie.

Podsumowując – udało się stworzyć system, który dla grupy wskazanych tekstów oraz pliku zawierającego glosariusz tworzy:

- Leksykon terminów – za pomocą stemmera portera
- Zbiór konceptów – za pomocą parsera plików WordNet
- Funkcję referencyjną – mapowanie terminów na formy podstawowe konceptów
- Hierarchię konceptów – stworzoną za pomocą analizy list synonimów między konceptami oraz określania wsparcia każdego z konceptów w glosariuszu.

7 LITERATURA

1. Ontology-based Text Clustering, A. Hotho and S. Staab and A. Maedche
2. Ontology-based Text Document Clustering, Andreas Hotho and Alexander Maedche and Steffen Staab