kaminski-stemmer

Autor

Mateusz Kamiński, student 5-tego roku Wydziału Elektroniki i Technik Informacyjnych, Politechnika Warszawska.

Projekt na przedmiot *Inteligentne Systemy Informacyjne*.

Styczeń 2018.

Opis

Kaminski-stemmer to aplikacja - biblioteka pełniąca rolę procesora tekstowego. Jej zadaniem jest stemming dowolnego tekstu dla języka obsługiwanego przez specjalnie przygotowany słownik zawierający reguły stemujące. Stemmer przeprowadza normalizację pewnego tekstu do tekstu wynikowego składającego się ze słów - nazywanych termami - sprowadzonych do prostej formy (np. dla czasownika będzie to forma bezokolicznika).

Domyślnie, kaminski-stemmer posługuje się odpowiednio przygotowanym słownikiem morfologicznym języka polskiego o nazwie PoliMorf http://zil.ipipan.waw.pl/PoliMorf (połączenie SGJP + Morfologik).

Kaminski-stemmer może być używany jako niezależna aplikacja lub wchodzić w skład istniejących aplikacji, które powstaływ językach bazujących na wirtualnej maszynie javy. Dla maksymalnej wydajności przetwarzania zalecana jest druga opcja, gdzie aplikacja utrzymuje w pamięci strukturę danych wymaganą do działania procesu stemmingu i może być wykorzystywana wielokrotnie.

Aplikacja kaminski-stemmer udostępnia mechanizm konwersji słownika w dowolnym języku na wewnętrzną reprezentację używaną przez stemmera. Kaminski-stemmer może być z powodzeniem wykorzystywany w systemach wykorzystujących Natural Language Processing (NLP), czyli w systemach przetwarzających tekst naturalny.

Modularność stemmera pozwala na dostosowanie procesu stemmingu do własnych potrzeb.

Charakterystyka

W aktualnej implementacji, słownik jest reprezentowany w postaci struktury danych o nazwie Patricia Tree, po polsku - w skompresowanym drzewie trie. Taka struktura jest kompromisem wydajności wyszukiwania ciągów znaków do wymagań pamięciowych. W tej strukturze każdy kolejny znak może być oddzielnym węzłem, gdzie węzeł odpowiada formie prostej danego termu. Więcej informacji jest dostępnych na: https://pl.wikipedia.org/wiki/Skompresowane_drzewo_trie

Kaminski-stemmer jest stemmerem, który opiera normalizację termów na uprzednio zdefiniowanym słowniku reguł. W odróżnieniu od języka angielskiego, język polski jest fleksyjny. Praktycznie niemożliwe jest opracowanie algorytmu stemującego tekst w języku fleksyjnym bez wsparcia w postaci słownika morfologicznego. Dlatego kaminski-stemmer jest dobrym wyborem dla tego typu języków.

Należy zdawać sobie sprawę z tego, że stemmer normalizuje słowa bezkontekstowo, bez rozróżnienia ich znaczenia w zdaniu. Z tego powodu znaczna ilość informacji może zostać utracona.

Kaminski-stemmer przetwarza tekst z pominięciem znaków interpunkcyjnych i znaków białych - pierwszym etapem działania jest tokenizacja tekstu na pojedyncze słowa. To zachowanie może być odpowiednio dostosowane w przypadku wykorzystania stemmera w roli biblioteki. Rezultatem tokenizacji jest podział tekstu na pojedyncze termy, gotowe do poddania się procesowi stemmingu. Każdy term jest wyszukiwany w słowniku. Jeśli dla pewnego termu forma prosta nie zostanie znaleziona, to dodatkowo stosowana jest funkcja "lowercase". W takiej formie term jest ponownie wyszukiwany w słowniku.

Wydajność

Wydajność aplikacji kaminski-stemmer została zmierzona dla różnorodnych źródeł tekstowych. Dla porównania, testy zostały przeprowadzone w identycznych przypadkach dla istniejącego i rozwijanego przez wiele lat stemmera o nazwie morfologik, który jest dostępny na: http://github.com/morfologik/morfologik-stemming/wiki .

Nazwa źródła	kaminski-stemmer czas wykonywania	kaminski-stemmer ilość różnych słów	morfologik czas wykonywania	morfologik ilość różnych słów
zbikowski-doktorat	122 ms	4932	292 ms	4708
pan_tadeusz	145 ms	10811	374 ms	13520
tadeusz-micinski-nauczycielka	25 ms	3608	42 ms	4377
sztuka-zdobywania-pieniedzy	7 ms	1306	17 ms	1428
ogniem-i-mieczem	346 ms	17490	540 ms	20394

Krótki opis źródeł:

1. zbikowski-doktorat - rozprawa doktorska Pana Kamila Żbikowskiego na temat "Konstrukcja funkcji kary dla klasyfikatorów SVM w automatycznych strategiach inwestycyjnych"; charakteryzuje się wyspecjalizowanym słownictwem naukowym z dziedziny informatyki. 127 stron tekstu może być sporym wyzwaniem dla stemmera.
2. pan_tadeusz - znana wszystkim epopeja narodowa, cechuje się staromodnym i trudnym językiem poetyckim.
3. tadeusz-micinski-nauczycielka - proza jednego z czołowych pisarzy polskiego ekspresjonizmu, prekursora surrealizmu, z trudnym ale współczesnym językiem
4. sztuka-zdobywania-pieniedzy - poradnik pisany przystępnym, współczesnym językiem. Objętościowo kilkanaście stron.
5. ogniem-i-mieczem - utwór pisany prozą, język staromodny, ok. 430 stron.

Kaminski-stemmer działa wyraźnie szybciej od morfologika i zazwyczaj produkuje tekst wynikowy składający się z mniejszej ilości różnych słów.

Przykłady

- *Nauczycielka - Tadeusz Miciński*

Oryginał:

> Otóż jestem na pensji. Zaczynam od zapisania swych wrażeń, aby siłę ich rozproszyć przez refleksję... Uczuciom swoim chcę dać reprezentację, aby nagłym wybuchem rewolucji nie zaskoczyły mnie znienacka. Chodźcie, chodźcie moje smutki, tęsknoty, nieokreślone marzenia — do tego niewodu, w którym pośniecie. Taki los ryb i wszystkiego, co żyje. Więc i ta śmieszna potrzeba, która wychodzi z serca blada, tęskna jak Goplana i chwyta się pierwszego sznura żurawi —

Tekst poddany stemmingowi:

> otóż być na pensja zaczynać od zapisać swój wrażenie aby siła on rozproszyć przez refleksja uczucie swój chcieć dać reprezentacja aby nagły wybuch rewolucja on zaskoczyć miąć znienacka chodzić chodzić mój smutki tęsknota określić marzyć — do to niewód w który posnąć taki los ryba i wszystko co żyć więc i ten śmieszny potrzeba który wychodzić z serce blady tęskny jaka goplana i chwytać się pierwszy sznura żurawi — zginąć muszy

- *Ogniem i mieczem*

Oryginał:

> Rok 1647 był to dziwny rok, w którym rozmaite znaki na niebie i ziemi zwiastowały jakoweś klęski i nadzwyczajne zdarzenia. Współcześni kronikarze wspominają, iż z wiosny szarańcza w niesłychanej ilości wyroiła się z Dzikich Pól i zniszczyła zasiewy i trawy, co było przepowiednią napadów tatarskich. Latem zdarzyło się wielkie zaćmienie słońca, a wkrótce potem kometa pojawiła się na niebie. W Warszawie widywano też nad miastem mogiłę i krzyż ognisty w obłokach; odprawiano więc posty i dawano jałmużny, gdyż niektórzy twierdzili, że zaraza spadnie na kraj i wygubi rodzaj ludzki. Nareszcie zima nastała tak lekka, że najstarsi ludzie nie pamiętali podobnej.

Tekst poddany stemmingowi:

> rok 1647 być to dziwny rok w który rozmaity znak na niebo i ziemia zwiastować jakowyś klęska i nadzwyczajny zdarzyć współczesny kronikarz wspominać iż z wiosna szarańcza w słychać ilość wyroić się z dziki pole i zniszczyć zasiew i trawa co były przepowiednia napad tatarski lato zdarzyć się wielki zaćmienie słońce a wkrótce potem kometa pojawić się na niebo W Warszawa widywać też nad miasto mogiła i krzyż ognisty w obłoki odprawiać więc posta i dawać jałmużna gdyż niektóry twierdzić że zaraza spaść na krajać i wygubić rodzaj ludzki nareszcie zimać nastały taka lekki że stary ludzie on pamiętać podobny

- *Rozprawa doktorska Pana Kamila Żbikowskiego*

Oryginał:

> Handel algorytmiczny jest relatywnie młodą dziedziną. Pod wzgl˛edem różnorodności źródeł wiedzy, na których bazują metody wykorzystywane przy tworzeniu automatycznych strategii inwestycyjnych dziedzina ta nie ust˛epuje tak zaawansowanym zagadnieniom jak loty kosmiczne. W ramach szerokiego spektrum prac badawczych wykorzystywana jest nie tylko ekonomia i finanse ale również fizyka, statystyka, teoria optymalizacji, uczenie maszynowe, sztuczna inteligencja, filozofia czy socjologia. Badanie istotności wyników musi odbywać si˛e z niezwykłą precyzją i dbałością o zachowanie właściwego metodologicznie podejścia w zakresie testowania strategii. Nie jest bowiem sztuką dopasowanie algorytmu do danych z przeszłości.

Tekst poddany stemmingowi:

> handel algorytmiczny być relatywnie młody dziedzina pod w r z wiedza na który bazować metoda wykorzystywać przy tworzyć automatyczny strategia inwestycyjny dziedzina ten on usta taka zaawansowany zagadnienie jaka lot kosmiczny W ramy szeroki spektrum prace badawczy wykorzystywać być on tylko ekonomia i finanse ale równie fizyka statystyka teoria optymalizacja uczyć maszynowy sztuczny inteligencja filozofia czy socjologia badanie istotny wynik muszy odbywać si z niezwykły precyzja i dbały o zachowanie w metodologicznie pod w zakres testować strategia on być bowiem sztuka dopasować algorytm do dany z przeszły

Do konwersji formatu pdf do formatu tekstowego wykorzystano oprogramowanie `pdftotext` https://linux.die.net/man/1/pdftotext . Na powyższych przykładach można zauważyć, że konwersja formatu może być skomplikowanym procesem i rzutować na właściwe i oczekiwane zachowanie stemmera.

Uruchomienie

Poniższa instrukcja zakłada uruchomienie aplikacji na dowolnym systemie operacyjnym UNIX-pochodnym. Dla systemów Windows kroki wyglądają identycznie, należy jednak zwrócić uwagę na różnice w obsłudze dostępnego terminala.

Aplikacja zakłada format tekstowy źródła jak i zapisuje wynik działania w tym formacie.

Przygotowanie do uruchomienia

Przed uruchomieniem aplikacji należy upewnić się, że posiadamy zainstalowaną wirtualną maszynę javy co najmniej w wersji 8. W systemie powinna być zdefiniowana zmienna JAVA_HOME wskazująca na miejsce instalacji Javy.

```
$ echo $JAVA_HOME
```

Aplikacja bazuje na systemie budowania o nazwie gradle. Nie jest jednak potrzebna jego instalacja ze względu na zastosowanie tzw. wrappera, który autoamtycznie pobierze z Internetu wszystkie potrzebne zasoby do zbudowania stemmera.

Szybkie uruchomienie

Pierwsze uruchomienie aplikacji może trwać dłużej ze względu na automatyczne pobranie niezbędnych bibliotek do zbudowania kodu aplikacji i jej działania.

Stemmer możemy uruchomić następująco (zakładajac plik źródłowy `example` zawierający tekst w języku polskim):

```
$ ./runner.sh -stem -source ./example -dest ./stemmed-example
```

Problemy

Jeżeli w wyniku uruchomienia otrzymamy następujący komunikat:

```
Exception in thread "main" java.lang.UnsupportedClassVersionError: com/mkaminski/stemmer/KaminskiStemmerRun
```

W takim przypadku zmienna środowiskowa `JAVA_HOME` nie została poprawnie skonfigurowana i nie wskazuje na oprogramowanie `Java` w wersji 8.

Konwersja słownika

Aplikacja kaminski-stemmer umożliwia konwersję dowolnego słownika w formie pliku TSV do wewnętrznej postaci, wykorzystywanej do procesu stemmingu. Plik TSV to plik CSV gdzie kolejne wartości oddzielone są znakiem tabularycznym. Pierwsza kolumna pliku TSV powinna zawierać termy w formie dowolnej, druga natomiast w formie prostej.

```
$ ./runner.sh -convert -source ./dict.tsv -dest ./dict.ser
$ ./runner.sh -stem -dict ./dict.ser -source ./example -dest ./stemmed-example
```

Dalszy rozwój stemmera

Wydajność wczytywania słownika

Mimo usilnych starań wyeliminowania problemu, wczytanie ogromnego słownika z formy poddanej serializacji bajtowej (ok. 122MB dla języka polskiego) do obiektu w pamięci maszyny wirtualnej jest czasochłonnym zadaniem dla komputerów klasy PC. Stąd dla maksymalnej wydajności rekomenduje się wykorzystanie kaminski-stemmer jako biblioteki, gdzie obiekt słownika może zostać wczytany do pamięci jednokrotnie i wykorzystywany wielokrotnie przez wielu "klientów".

Pluginy do systemów Information Retrieval - silniki Apache Solr i Elasticsearch

Silniki Apache Solr jak i Elasticsearch są otwarte na zewnętrzne narzędzia. Ich API wymaga implementacji ustalonego interfejsu w celu wykorzystania nowego stemmera.