Sprawozdanie

Metody obliczeniowe w nauce I technice

Temat: Singular Value Decomposition

Autor: Ryszard Sikora

Kod źródłowy wszystkich zadań dostępny jest pod adresem: https://github.com/rychuhardy/mownit-lab6.git

Zadanie 1: Wyszukiwarka

W zadaniu wykorzystałem korpus dokumentów tekstowych Reuters21578 oraz pakiet tm do zarządzania tym korpusem. Przed wyznaczeniem tzw. Bag-of-words teksty zostały odpowiednio przetworzone:

- zmiana wielkości liter na małe,
- usunięcie wyrazów nieistotnych (stop words),
- stemming,
- usunięcie znaków interpunkcyjnych,
- usunięcie liczb,
- usunięcie nadmiarowych białych znaków.

Przykładowy artykuł (nr 120) przed i po transformacji:

```
"<High Point Financial Corp> said it filed a registration statement with the
Securities and Exchange Commission covering six mln dlrs principal amount of
redeemable subordinated debentures due March one and cancellable mandatory stock
purchase contracts requiring the purchase of 6.66 mln dlrs in common no later than
March one.    It said the offering will be underwritten by Ryan, Beck and Co, West
Orange, N.J. Reuter "
```

```
"high point financi corp said file registr statement secur exchang commiss cover
six mln dlrs princip amount redeem subordin debentur due march one cancel
mandatori stock purchas contract requir purchas mln dlrs common later march one
said offer will underwritten ryan beck co west orange nj reuter "
```

Następnie wyznaczyłem TermDocumentMatrix jako macierz składająca się ze słów w wierszach oraz z artykułów w kolumnach. Poniżej fragment tej macierzy:

	Docs																			
Terms	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
haven	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
havew	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hawaii	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hawk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hawley	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hawthorne	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hay	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hayashi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hayes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hayesalbion	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hazard	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hazleton	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hcc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
head	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
headach	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
headed	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
headlines	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
headlund	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
headquart	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
headquarters	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
heal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
heali	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
health	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
healthcar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
healthcorp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
healthi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
healthvest	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
heap	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hear	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
heard	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hearing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Można zauważyć, że w losowo wybranym fragmencie macierzy na 600 elementów tylko jeden jest różny od zera.

Inverse Document Frequency dla tego samego zestawu słów:

haven	7.600902
havew	7.600902
hawaii	7.600902
hawk	6.214608
hawley	7.600902
hawthorne	7.600902
hay	7.600902
hayashi	7.600902
hayes	7.600902
hayesalbion	7.600902

```
hazard       5.809143
hazleton     6.214608
hcc          6.907755
head         3.772261
headach      7.600902
headed       7.600902
headlines    7.600902
headlund     7.600902
headquart    5.298317
headquarters 7.600902
heal         7.600902
heali        5.654992
health       3.989985
healthcar    6.214608
healthcorp   7.600902
healthi      5.654992
healthvest   6.907755
heap         7.600902
hear         4.828314
heard        6.907755
hearing      5.521461
```

Po zastosowaniu IDF często występujące słowa takie jak head oraz health mają ponad dwukrotnie niższą wagę niż raczej specyficzne słowo jak hawaii.

W celu przetestowania działania wyszukiwarki przygotowałem dwa zestawy słów z dwóch wybranych artykułów:

- Artykuł nr 2: “oil”, “committee”, “money”, “management”, “investment”, “market”
- Artykuł nr 15: “national”, “intergroup”, “inc”

Dla pierwszego zestawu oraz k=5, bez stosowania LRA ani IDF, zapytanie zwróciło wynik (numer artykułu oraz liczba wystąpień poszczególnych słów w nawiasach w kolejności malejącej):

- 248(9x oil, 10x market)
- 2(3x oil, committee, money, management, investment, market)
- 352(5x oil, 2x market)
- 1867(committee, management, 17x market)
- 247(oil, 4x money, 2x investment, 15x market)

Widać, że pojedyncze, często występujące, słowa miały wpływ na wyniki wyszukiwania bardziej niż ilość dopasowanych słów.

Dla porównania wyniki z zastosowaniem IDF:

- 2(3x oil, committee, money, management, investment, market)
- 1921(5x money, 3x market)
- 1973(5x money, 3x market) – ten sam artykuł co 1921
- 943(money, market)
- 336 (money, market) – ten sam artykuł co 336

Widać, że pożądany artykuł (nr 2) wyszedł na prowadzenie przy wyszukiwaniu z użyciem IDF. Natomiast pozostałe wyniki wydają się być gorzej dopasowane od wcześniej uzyskanych wyników.

Po zastosowaniu low rank approximation oraz IDF uzyskałem następujące wyniki:

- 2(3x oil, committee, money, management, investment, market)
- 247(oil, 4x money, 2x investment, 15x market)
- 350(money, investment)
- 241()
- 1207()

W tym przypadku największe dopasowania były lepsze lub porównywalne niż bez stosowania LRA, natomiast kolejne wyniki były zdecydowanie gorzej dopasowane.

Przy wyszukiwaniu z drugim zestawem artykuł 15 nie wystąpił wcale dla k = 5. Natomiast po zastosowaniu IDF artykuł 15 został wyszukany jako pierwszy. Po zastosowaniu LRA wyniki zostały zdominowane przez słowo “inc” nawet przy stosowaniu IDF.