Document Search using low rank approximation of TFIDF document representation

Ajay DSouza

May 24, 2017

1 Method

$$A \in \mathbf{R}^{m \times n}\text{,is the term-document matrix}$$
$$A = U\Sigma V^T$$
$$\text{Where, } \Sigma \text{ is the diagonal matrix with singular values } \sigma_1..\sigma_r,$$
$$\text{where } r \text{ is the rank of the matrix}$$
$$\mathbf{Rank}(A) = r$$
$$\text{Get a low rank approximation } A_k \text{ with } \mathbf{Rank}(A_k) = k \text{ for } A$$
$$\text{Set } \sigma_i = 0 \; for \; i > k \text{ such that it minimizes } ||A - A_k||_F$$
$$\text{Now, we can express the SVD with the lower rank } \Sigma_k\text{, as}$$
$$A_k = U^{m \times k} \Sigma_k^{k \times k} V^{T^{k \times n}}$$
$$\text{Thus we have the low rank matrix as}$$
$$A_k \in \mathbf{R}^{m \times n}$$

If $A_{k(i,j)} = 0 \;\; \forall i > l \; and \, \forall j \in 1 \rightarrow n$, we can remove those rows from A_k and represent it as a $l \times n$ matrix

The query vector Q can be mapped to the new space using the following transformation

$$q_k = {\Sigma_k}^{-1} U_k^T Q$$

Then the proximity of the query to each of the document can be got from the $\cos\theta$ of the angle between them as

$$\cos\theta = \frac{q_k^T . A_{k(j)}}{||q_k^T|| ||A_{k(j)}||} \; where \; A_{k(j)} \text{ is each column vector in } A_k$$

2 Project File Organization

2.1 Source code files

Document search project is implemented in the following files in matlab

```
$ ls -l *.m
-rwx------ 1 ajdsouza mkpasswd 2867 Apr 24 23:14 genlowranktd.m
-rwx------ 1 ajdsouza mkpasswd 3106 Apr 24 00:58 jacobi.m
-rwx------ 1 ajdsouza mkpasswd 2022 Apr 24 19:38 qrpivotingg.m
-rwx------ 1 ajdsouza mkpasswd 2068 Apr 24 23:46 vecquery.m
```

2.2 Data Files Used

Files containing the generated tfidf matrix, the document Index and the term index generated from the documents taken from news articles on the CNN web site.

matlab, matfiles used to store the SVD, and 5 low rank approximations of the term document matrix and used for the subsequent search

```
$ ls -l
total 13236
-rwx------ 1 ajdsouza mkpasswd 3923102 Apr 24 23:15 docsearchmat.mat
-rwx------ 1 ajdsouza mkpasswd    5924 Apr 24 22:07 documentIndex.txt
-rwx------ 1 ajdsouza mkpasswd  334982 Apr 24 23:15 rankApprox_1.mat
-rwx------ 1 ajdsouza mkpasswd  418681 Apr 24 23:15 rankApprox_2.mat
-rwx------ 1 ajdsouza mkpasswd  837873 Apr 24 23:15 rankApprox_3.mat
-rwx------ 1 ajdsouza mkpasswd 1721307 Apr 24 23:15 rankApprox_4.mat
-rwx------ 1 ajdsouza mkpasswd 3441952 Apr 24 23:15 rankApprox_5.mat
-rwx------ 1 ajdsouza mkpasswd 2788352 Apr 24 20:05 termDocumentMatrix.txt
-rwx------ 1 ajdsouza mkpasswd   64516 Apr 24 21:54 termIndex.txt
```

3 Results

The following are the results of some of the queries we executed

```
EDU>> vecquery
Enter the string to search (e.g. tripoli peopl hide gadhafi ) :
ghadafi
Searching documents fro query string ghadafi
===================================================
Searching through term doc matrix with rank 8
Closest matching DocNo=28 _document11088.txt
===================================================
Searching through term doc matrix with rank 10
Closest matching DocNo=28 _document11088.txt
===================================================
Searching through term doc matrix with rank 20
Closest matching DocNo=84 _document5350.txt
===================================================
Searching through term doc matrix with rank 41
Closest matching DocNo=64 _document5015.txt
===================================================
Searching through term doc matrix with rank 82
Closest matching DocNo=64 _document5015.txt
EDU>>
EDU>>
EDU>>
EDU>>
EDU>> vecquery
Enter the string to search (e.g. tripoli peopl hide gadhafi ) :
georg zimmerman claim fire defenc
Searching documents fro query string georg zimmerman claim fire defenc
===================================================
Searching through term doc matrix with rank 8
Closest matching DocNo=93 _document509.txt
===================================================
Searching through term doc matrix with rank 10
Closest matching DocNo=93 _document509.txt
===================================================
Searching through term doc matrix with rank 20
Closest matching DocNo=93 _document509.txt
===================================================
Searching through term doc matrix with rank 41
Closest matching DocNo=93 _document509.txt
===================================================
Searching through term doc matrix with rank 82
Closest matching DocNo=93 _document509.txt
EDU>>
EDU>>
EDU>>
EDU>>
EDU>>
EDU>> vecquery
Enter the string to search (e.g. tripoli peopl hide gadhafi ) :
georg zimmerman claim fire defenc
Searching documents fro query string georg zimmerman claim fire defenc
===================================================
Searching through term doc matrix with rank 8
Closest matching DocNo=93 _document509.txt
===================================================
Searching through term doc matrix with rank 10
Closest matching DocNo=93 _document509.txt
===================================================
Searching through term doc matrix with rank 20
```

```
Closest matching DocNo=93 _document509.txt
===================================================
Searching through term doc matrix with rank 41
Closest matching DocNo=93 _document509.txt
===================================================
Searching through term doc matrix with rank 82
Closest matching DocNo=93 _document509.txt
EDU>>
EDU>>
EDU>>
EDU>>
EDU>> vecquery
Enter the string to search (e.g. tripoli peopl hide gadhafi ) :
tripoli peopl hide moammar
Searching documents fro query string tripoli peopl hide moammar
===================================================
Searching through term doc matrix with rank 8
Closest matching DocNo=31 _document5053.txt
===================================================
Searching through term doc matrix with rank 10
Closest matching DocNo=31 _document5053.txt
===================================================
Searching through term doc matrix with rank 20
Closest matching DocNo=31 _document5053.txt
===================================================
Searching through term doc matrix with rank 41
Closest matching DocNo=12 _document5950.txt
===================================================
Searching through term doc matrix with rank 82
Closest matching DocNo=12 _document5950.txt
EDU>>
EDU>>
EDU>>
```