MSc Electronic and Computer Engineering

Data Mining and Machine Learning (2019)

Lab 1 – Text Retrieval

Group 12

Miaoyu Niu (1893824)

Kai Chen (1948361)

2019-2-18

UNIVERSITY OF BIRMINGHAM

PART 1: TF-IDF BASED TEXT RETRIEVAL

Task 1: Ready to build a simple text-IR system.

Task 2: Compile stop.c and porter-stemmer.c

cl stop.c : Create stop.exe and stop.obj in the folder lab1-2019.

cl porter-stemmer.c : Create porter-stemmer.exe and porter-stemmer.obj in the folder lab1-2019.

Task 3: Stop word removal: This operation could remove common ‘noise words’ from the texts, including punctuation, paragraphs and words in the Stop List. And stopScript.bat is a batch file used to process all files in the ‘beng’ folder. All of the ‘stopped’ documents are placed in a new folder called ‘stop’, each with a name of the form ‘filename.stp’.

Question 1: For AbassiM.txt, Agricole.txt and AliR.txt, after running the stop-word removal, the percentage reduction in the number of words are about 36.78%, 35.29% and 36.22% respectively. All stop words, such as ‘a’ and ‘the’, are removed from multiple word queries to increase search performance. At the same time, all uppercase formats, paragraph formats and punctuation are also removed.

Task 4: Stemming: This operation is used to extract the stem or root form of the words. It could remove irrelevant differences from different ‘versions’ of the same word by using ‘porter-stemmer stop\AbassiM.stp’ command. And stemScript.bat is another batch file used to process all files in the ‘stop’ folder. All of the ‘stemmed’ documents are placed in a new folder called ‘stem’, each with a name of the form ‘filename.stm’.

Question 2: ‘Stemming’ could remove surface markings from words to reveal their basic form: communications common,→sophisticated sophist,→
transmissions → transmiss.

Task 5: This task is to create the document index files: This task is to create three index files: textIndex, stopIndex and stemIndex. It contains the lengths and the number, IDF and weight of each word, as well as the number of files containing this word and the file name. IDF is used to measure the significance of a term for discriminating between documents. The weight of a term for a document is IDF multiplies term frequency.

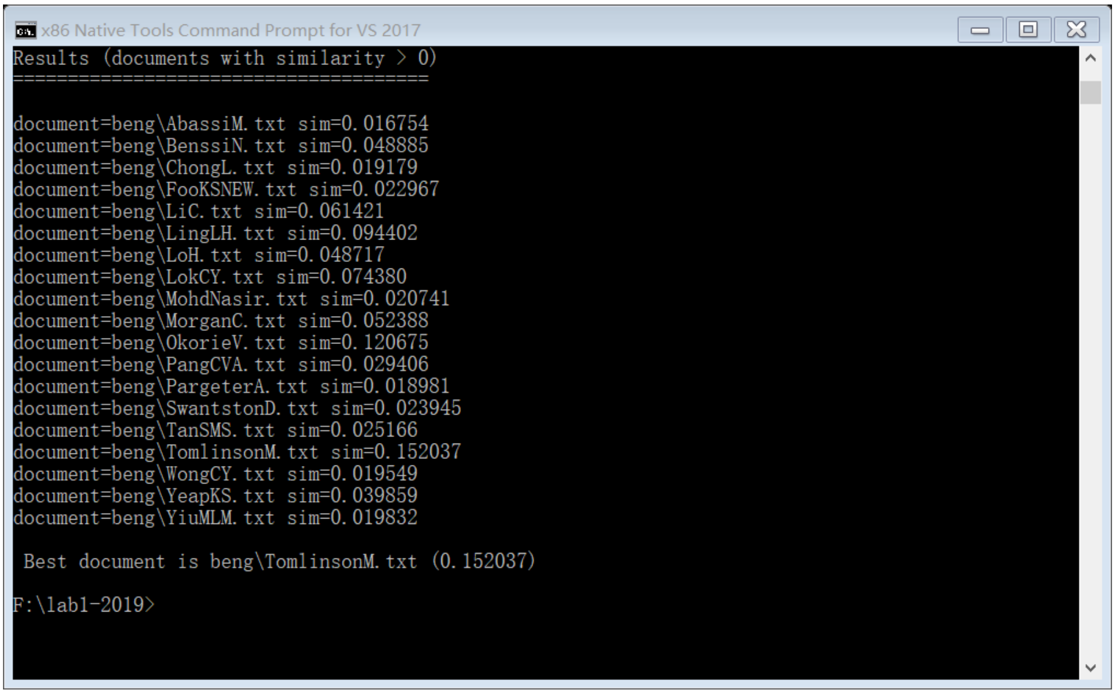
Question 3: The document lengths of beng\DongP.txt is 42.396210. The document lengths of stop\DongP.stp is 42.392876. The document lengths of stem\DongP.stm is 40.547955. Because the stop file reduces the stop words compared to the original file, and the stem file extracts the stem or root form of these words based on the stop file. Therefore, the difference between the document lengths of stem\DongP.stm and beng\DongP.txt is greater than the difference between the document lengths of beng\DongP.stp and beng\DongP.txt

Question 4: This word appears in almost every article, so its IDF is close to 0. The more the number of documents containing the current word, the smaller the value of IDF, indicating that the less important the word is.

Question 5: There are three forms of word 'algorithm' in the text files is 'algorithm', 'algorithmic' and 'algorithms', as same as in the stop files. Because the operation 'stop word removal' does not change the form of the word. Their wordCount are 14, 3 and 19 respectively. However, since the operation 'stemming' would extract the stem of the word, the word about 'algorithm' in the stem files is only 'algorithm' itself. And the wordCount of this word is the sum of the number of three words, which is 36.

Task 6: Firstly, create a query contains the text: communication and networks. Then apply stop word removal and stemming to the query. Secondly, the command 'retrieve textIndex query' will return a list of all the documents for which the similarity with the query is greater than 0. Finally, repeat this for the stopped documents and stopped query, and stemmed documents and stemmed query. The commands are 'retrieve stopIndex query.stp' and 'retrieve stemIndex query.stm'.

The content relevance between the query and the document is Sim(q, d). It defines the similarity between them. The above operations will return all the files containing the words in the query, including the file names, weights and numbers. It also returns the best document which is the most relevant document to the query. The three best document is beng\TomlinsonM.txt, stop\TomlinsonM.stp and stem\YiuMLM.stm. The Sim are 0.152037, 0.152309 and 0.261187 respectively.

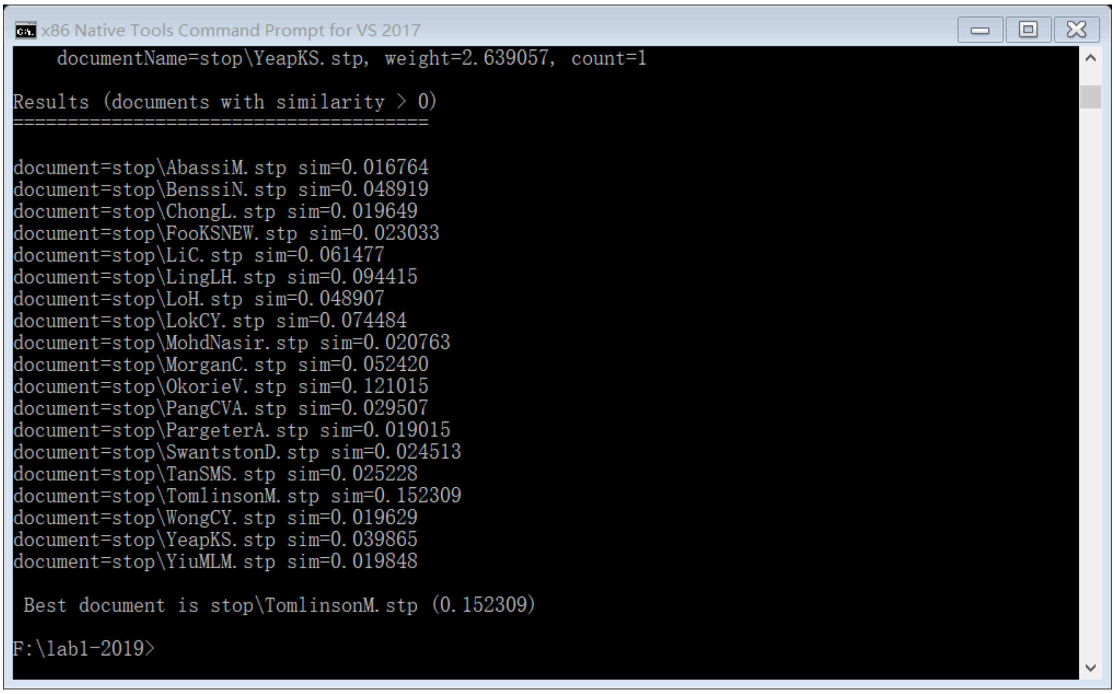


```
Results (documents with similarity > 0)
=====================================

document=beng\AbassiM.txt sim=0.016754
document=beng\BenssiN.txt sim=0.048885
document=beng\ChongL.txt sim=0.019179
document=beng\FooKSNEW.txt sim=0.022967
document=beng\LiC.txt sim=0.061421
document=beng\LingLH.txt sim=0.094402
document=beng\LoH.txt sim=0.048717
document=beng\LokCY.txt sim=0.074380
document=beng\MohdNasir.txt sim=0.020741
document=beng\MorganC.txt sim=0.052388
document=beng\OkorieV.txt sim=0.120675
document=beng\PangCVA.txt sim=0.029406
document=beng\PargeterA.txt sim=0.018981
document=beng\SwantstonD.txt sim=0.023945
document=beng\TanSMS.txt sim=0.025166
document=beng\TomlinsonM.txt sim=0.152037
document=beng\WongCY.txt sim=0.019549
document=beng\YeapKS.txt sim=0.039859
document=beng\YiuMLM.txt sim=0.019832

 Best document is beng\TomlinsonM.txt (0.152037)

F:\lab1-2019>
```



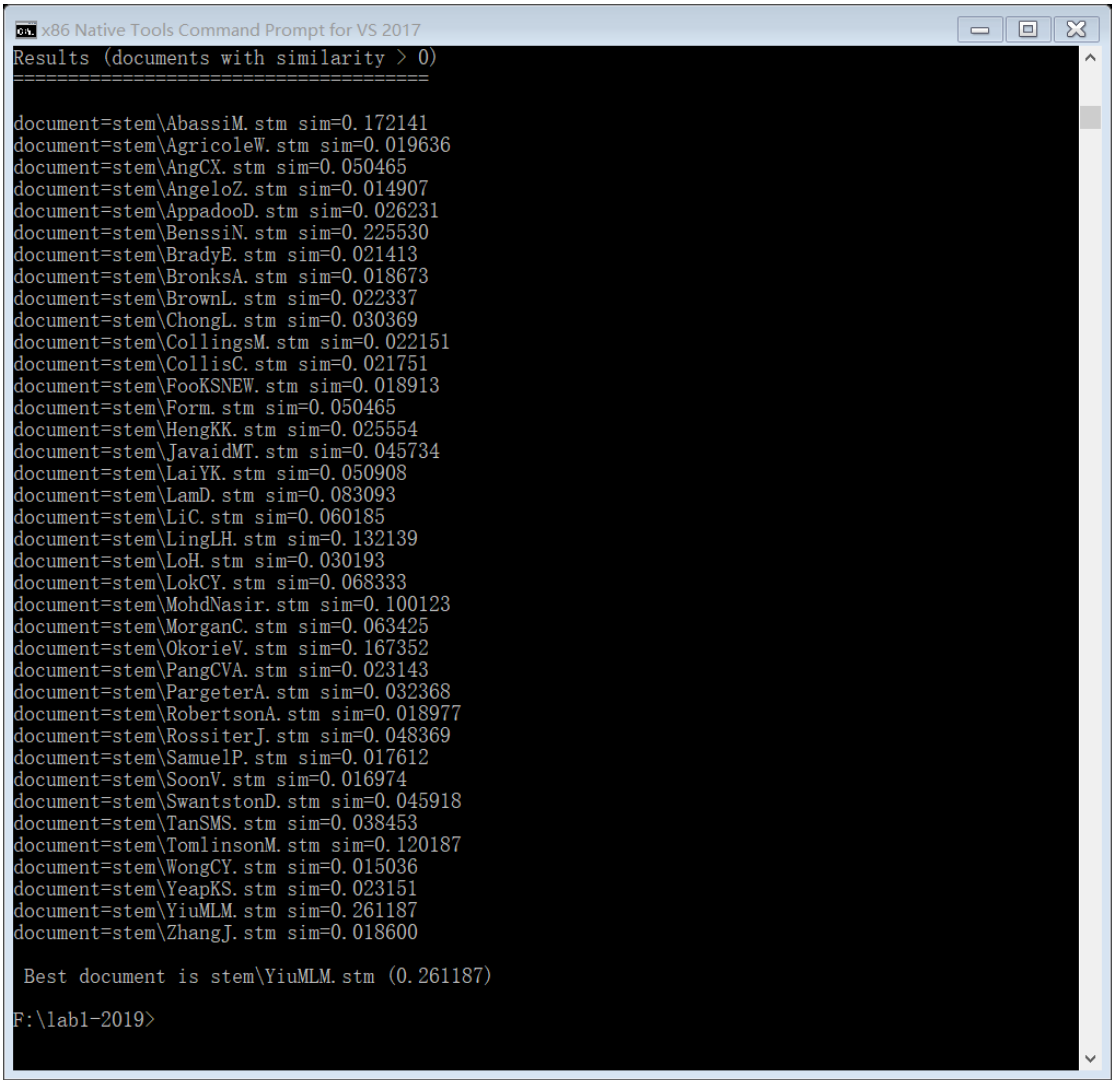
```
    documentName=stop\YeapKS.stp, weight=2.639057, count=1

Results (documents with similarity > 0)
=====================================

document=stop\AbassiM.stp sim=0.016764
document=stop\BenssiN.stp sim=0.048919
document=stop\ChongL.stp sim=0.019649
document=stop\FooKSNEW.stp sim=0.023033
document=stop\LiC.stp sim=0.061477
document=stop\LingLH.stp sim=0.094415
document=stop\LoH.stp sim=0.048907
document=stop\LokCY.stp sim=0.074484
document=stop\MohdNasir.stp sim=0.020763
document=stop\MorganC.stp sim=0.052420
document=stop\OkorieV.stp sim=0.121015
document=stop\PangCVA.stp sim=0.029507
document=stop\PargeterA.stp sim=0.019015
document=stop\SwantstonD.stp sim=0.024513
document=stop\TanSMS.stp sim=0.025228
document=stop\TomlinsonM.stp sim=0.152309
document=stop\WongCY.stp sim=0.019629
document=stop\YeapKS.stp sim=0.039865
document=stop\YiuMLM.stp sim=0.019848

 Best document is stop\TomlinsonM.stp (0.152309)

F:\lab1-2019>
```



```
Results (documents with similarity > 0)
=====================================

document=stem\AbassiM.stm sim=0.172141
document=stem\AgricoleW.stm sim=0.019636
document=stem\AngCX.stm sim=0.050465
document=stem\AngeloZ.stm sim=0.014907
document=stem\AppadooD.stm sim=0.026231
document=stem\BenssiN.stm sim=0.225530
document=stem\BradyE.stm sim=0.021413
document=stem\BronksA.stm sim=0.018673
document=stem\BrownL.stm sim=0.022337
document=stem\ChongL.stm sim=0.030369
document=stem\CollingsM.stm sim=0.022151
document=stem\CollisC.stm sim=0.021751
document=stem\FooKSNEW.stm sim=0.018913
document=stem\Form.stm sim=0.050465
document=stem\HengKK.stm sim=0.025554
document=stem\JavaidMT.stm sim=0.045734
document=stem\LaiYK.stm sim=0.050908
document=stem\LamD.stm sim=0.083093
document=stem\LiC.stm sim=0.060185
document=stem\LingLH.stm sim=0.132139
document=stem\LoH.stm sim=0.030193
document=stem\LokCY.stm sim=0.068333
document=stem\MohdNasir.stm sim=0.100123
document=stem\MorganC.stm sim=0.063425
document=stem\OkorieV.stm sim=0.167352
document=stem\PangCVA.stm sim=0.023143
document=stem\PargeterA.stm sim=0.032368
document=stem\RobertsonA.stm sim=0.018977
document=stem\RossiterJ.stm sim=0.048369
document=stem\SamuelP.stm sim=0.017612
document=stem\SoonV.stm sim=0.016974
document=stem\SwantstonD.stm sim=0.045918
document=stem\TanSMS.stm sim=0.038453
document=stem\TomlinsonM.stm sim=0.120187
document=stem\WongCY.stm sim=0.015036
document=stem\YeapKS.stm sim=0.023151
document=stem\YiuMLM.stm sim=0.261187
document=stem\ZhangJ.stm sim=0.018600

 Best document is stem\YiuMLM.stm (0.261187)

F:\lab1-2019>
```

Task 7: This task needs to create two query files, query1 and query 2, including the text : ‘requirements transducer’ and ‘theoretical and amplifier’. For query1, the three best document is beng\PargeterA.txt, stop\PargeterA.stp and stem\PargeterA.stm. The Sim are 0.182517, 0.182845 and 0.237534 respectively. For query2, the three best document is beng\TngL.txt, stop\TngL.stp and stem\TngL.stm. The Sim are 0.258413, 0.259259 and 0.239817 respectively.

PART 2: LATENT SEMANTIC ANALYSIS

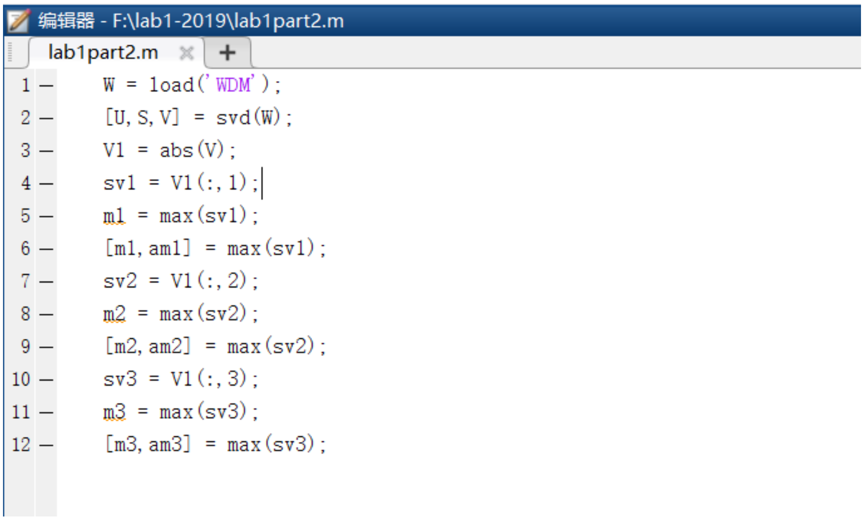
Task 1: This creates a document vector for each document in the stem folder and stacks them to create the matrix in the file WDM. It is necessary to compile doc2vec.c (cl doc2vec.c) firstly and change the type of stemFileList file, then run the following command:

doc2vec stemFileList.txt > WDM

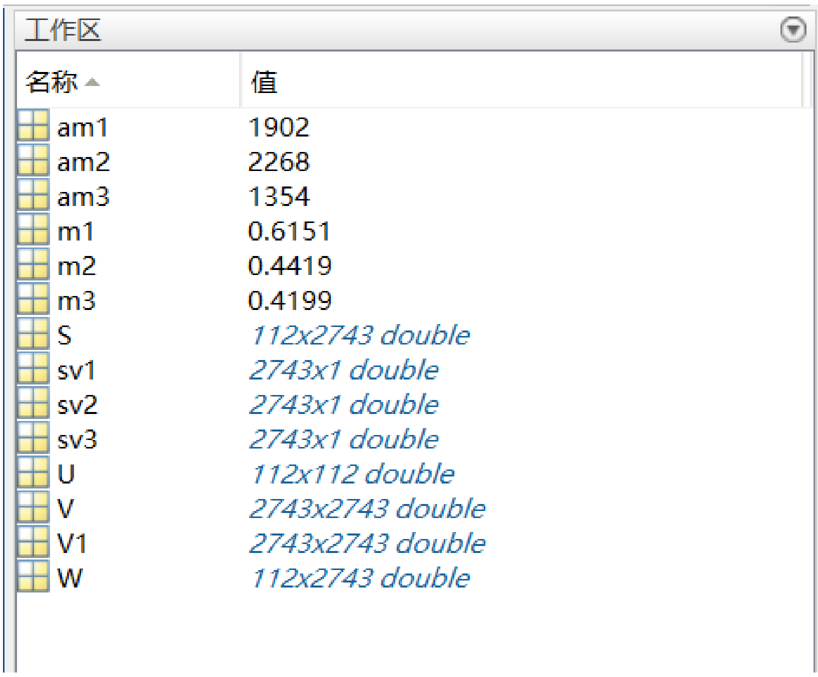
Task 2: Apply Singular Value Decomposition (SVD) to the word-document matrix. It could be done by using SVD functions in Matlab:

>>[U,S,V]=svd(A).

This runs SVD on W, decomposing it as W = USV^T.



```
W = load('WDM');
[U,S,V] = svd(W);
V1 = abs(V);
sv1 = V1(:,1);
m1 = max(sv1);
[m1,am1] = max(sv1);
sv2 = V1(:,2);
m2 = max(sv2);
[m2,am2] = max(sv2);
sv3 = V1(:,3);
m3 = max(sv3);
[m3,am3] = max(sv3);
```



名称	值
am1	1902
am2	2268
am3	1354
m1	0.6151
m2	0.4419
m3	0.4199
S	*112x2743 double*
sv1	*2743x1 double*
sv2	*2743x1 double*
sv3	*2743x1 double*
U	*112x112 double*
V	*2743x2743 double*
V1	*2743x2743 double*
W	*112x2743 double*

Question 1: U is a 112*112 matrix (112 is the number of documents) and V is a 2734*2734 matrix (2734 is the number of terms). The matrices U and V represent two mutually orthogonal matrices, and S represents a diagonal matrix. As it can be seen from the results of the operation, U and V are indeed orthogonal matrices. Therefore, the matrices U and V are as we would expect.

Question 2: The values of the first 3 diagonal entries in S is 274.596487768020, 53.7182065440781 and 47.1338164059136.

Then the first column of V should be written into the vector sv1.And do this for the first 3 columns of V, creating singular vectors sv1, sv2 and sv3. The most important word for the interpretation of SV1 is the word with the largest coordinate of sv1.

Question 3: The three most significant words for each of the singular vectors sv1, sv2 and sv3 are project (1902), data (608) and speech (2268) separately. We can find them in the stemindex file with the corresponding numbers. Each column of V is a document vector corresponding to a semantic class in the corpus. Each row in matrix U represents a type of word that is related, with each non-zero element representing the importance (or relevance) of each of these words. The larger the value, the more relevant. Each column in matrix V represents a type of article of the same topic, where each element represents the relevance of each article in such an article. The matrix S represents the correlation between the class and the article class.