Assignment 3 Writeup and Method Comparison
Xinyu Zheng

1. Run LDA on a sample set

I run LDA on the 10 original txt documents (6334220.txt to 6334229.txt) from assignment 2. I cleaned each document (removed stop words, stemming, digits, single letters, and special characters), generated a term document matrix, and run LDA model on the matrix with 1000 iterations to get the 10 topics. I allowed 9 words for each topic and the generated topics are as follows:

Topic 0: meet main defin fulli loosen end conduct adjust thereof
Topic 1: screw sock height thereof foot fulli opposit seat-support person
Topic 2: foot sidewal thereof semirigid displac either peripheri rim 40:1
Topic 3: either brace cloth method move strip loop nut inclin
Topic 4: close rest wall amount main remov rim adhes displac
Topic 5: screw remov predetermin torqu within fan effect plate strap
Topic 6: part posit main edg upon sheet guid strap fix
Topic 7: main brace move top guid screw recit convers ledg
Topic 8: 50.80 inhibit maintain main gener adjust fulli rang collect
Topic 9: opposit screw rim ride an rubber main meet inhibit

2. Build a toy data set of text documents and relationships between the text documents

I built a toy data set of 5 documents. toy_doc_1.txt and toy_doc_2.txt fall into group one, which is related to the topic 'apple'. toy_doc_3.txt, toy_doc_4.txt, and toy_doc_5.txt fall into group two, which is related to the topic 'book' and 'reading'. The five documents are attached.

3. Calculate the jargon distance between documents using the method in the Jargon Paper

According to my results, H value for group one is 4.306138533 and H value for group two is 4.73389465394. This indicates that the expected message length per phrase in group one is about 4.3 bytes while that in group two is about 4.7 bytes.

In addition, Q value of group one to group two is 12.4348132056 while Q value of group two to group one is 11.9301882073. This result indicates that when the writer is from group one and the reader is from group two, the expected length of the writer's message per phrase is about 12.43 bytes. However, when the writer is from group two and the reader is from group one, the expected length of the writer's message per phrase is about 11.93 bytes.

Furthermore, E value from group one to group two is 0.346297001957 while that from group two to group one is 0.396799662477. This shows that the quantified efficiency of communication from group one to group two is about 34.63% while that from group two to group one is about 39.68%.

At last, the estimated average cultural hole around group one is 0.653702998043 while that around group two is 0.603200337523.

4. Run LDA on toy data sets

I run LDA model on my toy data sets and generated two topics shown as follows. There are 14 words in each topic and the iteration is 1000.

Topic 0: reader it make respons almost public borrow fruit week energet profit favorit love new
Topic 1: appl morn peter note everi eat keep fresh farm grow lot work harvest especi

From the above result I can tell that the two topics generated by LDA model are basically about 'apple' and 'reader' – very close to my idea when I created the five documents that fall into two groups.

5. Compare methods (1-2 page paper) 6. Turn in paper, toy data set, code and LDA results

I run both Jargon Distance and LDA on my toy data set and here are my thoughts on the comparison of them.

Jargon Distance is applied when the document groups are known. It runs on different groups and calculates the efficiency of communication from one group to another group. When it calculates the efficiency of communication, Jargon Distance takes the probability of each word in each group and the whole corpus. The results of Jargon Distance are average culture hole around each group, indicating the degree of convenience of communication between a group and others around it.

LDA runs on a corpus that incudes a bunch of documents but does not need to know which document falls into which group. In order to run LDA, we should first generate a term-document matrix that shows the frequency of each term in every document. With the term document matrix we generate certain number of topics that describe the corpus. LDA is used to help us understand the main information conveyed by a corpus.

To summarize, Jargon Distance and LDA are different methods applied in text analysis. Jargon Distance helps us understand the efficiency of communication between groups of documents while LDA helps us understand the main information conveyed by a corpus. Therefore, we should combine these two methods when we do text analysis.