Homework Assignment 1

Assignment Evaluate the dimensionality reduction techniques LSA and LDA. Use 2 data sets. The first data set is one of the standard research data sets, 20 NewsGroups. The second is the Yelp review data set from the Yelp data Challenge. Compute the LSA and LDA representation for documents from both data sets and evaluate if it help to achieve better performance for clustering. Total 200 points.

Page limit 10 pages. You will do quite a bit of work for this assignment but for your assignment report summarize everything and write only the final results and observations, not all of your steps. **Do not cut-and-paste screen shots** of all your work. Screen shots won't count as answers. Write brief explanations for each question below instead. Only use screen shots for 2-3 command lines to show what you used to compute LSA/LDA/kMeans and how you constructed the LSA and LDA vectors for clustering. You can also use screen shots to show the most frequent words. But make sure those screen shots are explained in your own words. Do not copy-paste from solutions by other students. It's ok to work in teams but write the solutions in your own words. If we see solutions that use the same words, all of them will get 0 points.

Submit your solutions as a PDF file on the BB. You may write the solutions to the questions in IV by hand, take a screen shot and insert into your solution. Do not submit it separately. Make sure that the solution is clearly written, the screen shot quality is good. Otherwise it won't be graded!

Details

- Use 20 NewsGroups data set http://qwone.com/~jason/20Newsgroups/
- Use Yelp Challenge data https://www.yelp.com/dataset_challenge
- Use R

Some helpful links about the processing provided at the end of the assignment.

Use the following outline for your assignment report. You should use this outline for your final project report as well.

I Data (15 points)

• Read about each data set, describe briefly in your own words what they are, write about how many documents they contains, what are the topics in that data (if it is known), how different are documents from one another, are there any semantic groups that you would expect just based on your world knowledge.

• For 20 News groups provide an overview of the number of news topics, number of documents per topic, and number of unique words for each, explain how it matters for your experiments. For the Yelp data, research if the reviews can be grouped into different thematic groups.

- If you understand these datasets well, it will be easier to give good answers to the following questions.

II Experiments

II.A Data Preprocessing (25 points)

• Determine the size of the data that you can process (based on the memory etc available on your laptop or lab computers. If you can use AWS or similar resources, that's great).

- You should be able to process the full 20 NewsGroups data. If not, select some groups out of 20, and use only documents from those groups. You should select at least 5 groups, but more would be better. Explain, what groups you selected and why. Analyze if they are similar or not and what results you would expect based on that.
- Use a sample from the Yelp data based on how much data you can process. You can select reviews at random or based on a particular theme, for example, cuisine, location etc. Explain, what reviews you selected and why. Analyze if they have common topics or not and what results you would expect based on that.

• Use your analysis from part I about the #docs, #words etc for each topics/group you pick. Discuss if the topics/groups you picked are similar to each other or not and how this will matter for your experiments. Based on what you know about clustering and LSA, LDA do you think they will perform well on your data just by analyzing the documents that you picked?

• Create document-term matrices
• Remove stop words (using the stop words list)
• Use stemming (you can use the document-term matrix without stemming first and see if stemming helps)

- Prune words by frequency – remove words that occur in very few documents (e.g. <4) or that occurred in too many documents (depends on the data, e.g. 300-400 for 20 NG). Discuss briefly the vocabulary size before your pruning and after. How does this affect LSA/LDA/clustering?
- Use tf-idf weights for your document-term matrices

II.B Clustering Experiments. Do the following steps for EACH of the two data sets. The points break down below is show for one data set. (45 points for each data set)

1. Cluster the document vectors from the tf-idf document-term matrix with kMeans to have a baseline for your comparisons. (15 points)
    - For each data set discuss how you will set the parameter **K**.
    - Use NbClust to determine the best number of clusters. (For the students who don't know about this technique I will upload help material).
2. Compute the LSA representation, cluster LSA document vectors (15 points)

    • Compute the SVD of the document-term matrix

    • Discuss briefly how you obtain the k-dimensional LSA document vectors and LSA word vectors from the SVD. Make sure you provide enough details to show that you select the correct matrices from the SVD decomposition to represent the documents. Here you can show 1-2 lines of the R commands you use to create LSA document vectors that you use as input to clustering.

    • Compute the d=50, 100, 200 dimensional representation for the term-document matrix

    • Cluster the d-dimensional documents and the words with kMeans, using the same **K** as for the tf-idf vectors to be able to compare the clustering results.

• For each of the top 5 concepts report the most representative words. We discussed in Lecture 2 how to do it.

3. Compute the LDA representation for the documents (15 points) See a note below how to compute LDA documents
   - Discuss briefly, what is the document representation that you will get after processing the data with LDA in R. Discuss briefly, what are the main parameters for LDA and how did you set their values.
   - Compute LDA representations for d topics, using different values for d. You can use the same values for d as you did for LSA.
   - Discuss and write equations of how you represent documents as d-dimensional vectors using the output of LDA.
   - Cluster LDA vectors with kMeans using the same **K** as for tf-idf
   - For each of the top 5 concepts report the most representative words

II.C Evaluation (30 points)

4. Use the SSE measure to evaluate and compare your clustering results. Do this for both data sets. Compare clustering results for documents represented with tf-idf and for the LSA, LDA document vectors. Use the SSE measure for clusters evaluation and comparison. (15 points)
   a. Remember that one should usually compare clustering results with the same or comparable number of clusters. Keep it in mind also when comparing clustering result for different setting and different values of d in d-dimensional vectors. Discuss how it matters. For this question, recall that SSE usually decreases with more clusters SSE during the clustering process.
   - For the 20NG you have the actual class labels for each document which is its news group assignment. Use the labels to compute the accuracy for your clusters. For each cluster pick the majority news group as your cluster class label and assign it to all documents in that cluster. Then compute Precision, Recall, F1 for each news group that you use. Also compute the confusion matrix. (15 points)

II.D Results Summary (10 points)

Present the results of your experiments. A summary table is usually a good way to summarize and compare results for your experiments on different data and for different numbers of clusters/ different LSA dimensions and different LDA parameters.

III Analysis (15 points)

This is the most important part of the assignment! Discuss what results you got, how do you evaluate the usefulness of LSA/LDA for this data and for your clustering problem for each of the data sets. Use your analysis of the data you did in Section 1 to explain the results and do the error analysis. Discuss the semantic spaces computed by LSA/LDA using the most representative words. Discuss what you learned.

IV LSA Derivation (15 points)

- Consider the n x m term-document matrix A. Compute the matrix products AA^T and A^TA in terms of the matrices of A's SVD decomposition. You won't have a numeric answer here, just replace A

in AA^T and A^TA by its SVD matrices and show how to simplify the product to get to a representation of AA^T and A^TA with 3 matrices. Explain what relationship you see between the SVD of A and the eigenvalue decomposition of AA^T and A^TA. (5 points)

- Using the SVD matrices, show what is the LSA document representation that you used in for your clustering experiments. Show 1 or 2 R commands that confirm that you did that. (10 points)

Useful links

- Tm for R: Text Mining Package

  http://cran.r-project.org/web/packages/tm/index.html

- One example of how to read the Yelp data with R

http://rstudio-pubs-static.s3.amazonaws.com/121639_3364a2eb69b54ed9b85faf1ecf21cd7f.html

- Some additional text mining libraries for LDA and visualization

http://tidytextmining.com/topicmodeling.html

http://www.imsbio.co.jp/RGM/R_rdfile?f=lda/man/newsgroups.Rd&d=R_CC

- How to compute LDA document representation

Assuming that LDA produced a list of topics and put a score against each topic for each document, you could represent the document and it's scores as a vector:

```
Document | Topic1 | Topic1 | Topic1 | TopicN | ...
   1       0.041    0.042      0.041          ...
   2       0.052    0.011      0.042          ...
```

Use the top N topics probabilities as new entries in your document vectors and cluster those vectors.