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 100 points.

Details

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

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 (5 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.

II Experiments

II.A Data Preprocessing (10 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).

- If you can process the full 20 NewsGroups data, use 20 news groups. If not, select some groups out of 20, and use only documents from those groups. You should select at least 3 groups, but more would be better.
- 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.

• 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).
• Using counts of word occurrence

• Using tf-idf
• 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 clustering?

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.

1. Cluster the original documents vectors to have a baseline for comparison (5 points)

   • Cluster the document vectors from the input documents word matrix with kMeans. 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 (10 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.

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

   • Cluster the d-dimensional documents and the words with kMeans

   • For each of the d concepts report the most representative words

3. Compute the LDA representation for the documents (10 points)
   - Discuss briefly, what is the document representation that you will get after processing the data with LDA in R.
4. Compare clustering results for documents represented with tf-idf and for the LSA, LDA documents. Use the SSE measure for clusters evaluation and comparison. (10 points)
   a. Remember that one should usually compare clustering results with the same or comparable number of clusters. Keep it min when comparing clustering result for different setting.

II.C Results Summary (5 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 and LSA dimensions.

III Analysis (10 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.

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