DATA.STAT.840 Statistical Methods for Text Data Analysis

Exercises for Lecture 6: Topic models

Exercise 6.1: Latent semantic analysis

(a) Download the 20 Newsgroups data set from http://qwone.com/~jason/20Newsgroups/.

(b) In this exercise we consider only four of the newsgroups: rec.autos, rec.motorcycles, rec.sport.baseball, and rec.sport.hockey. Process the documents of the four newsgroups using the pipeline described on the lectures, including vocabulary pruning.

(c) Create a TF-IDF representation for the documents, using Length-normalized frequency (TF) and Smoothed logarithmic inverse document frequency (IDF).

(d) Apply latent semantic analysis to the TF-IDF matrix, to find 10 underlying factors.

(e) Describe the resulting factors: list the 10 words with highest (absolute) weight in each factor. Do the factors seem related to individual newsgroups? Does their content seem meaningful?

(f) Do the same with 15 factors (the first 10 factors will be the same). Do the new 5 factors seem more or less meaningful?

Report your analysis results and your code.

Exercise 6.2: Probabilistic latent semantic analysis

(a) Using the same data as in Exercise 6.1 (four newsgroups), create a term frequency matrix of raw term counts for the documents.

(b) Apply PLSA to the term frequency matrix to find 10 underlying factors.

(c) Describe the resulting factors: list the 10 words with highest probability in each factor.

(d) Find, for each factor, the document (message) with highest probability of that factor, and print its 100 first words.,

(e) Do the factors seem related to individual newsgroups? Does their content seem meaningful?

(f) Optional: let's see if an AI chatbot can analyze these factors. See exercise 2.4 regarding possible chatbots; report what chatbot you used.
Take one of the factors, e.g., the highest probability one, and give its list of top words to an AI chatbot; you can tell it that the factor has been learned from the four newsgroups. Ask it to interpret the factor based on the word list. Ask it also to generate a text that could be drawn from the factor. Does its interpretation seem meaningful and likely correct?

Report your analysis results and your code.

Exercise 6.3: Latent Dirichlet allocation

(a) Using the same data as in Exercise 6.1 (four newsgroups), create a term frequency matrix of raw term counts for the documents.

(b) Apply Latent Dirichlet Allocation to the term frequency matrix to find 10 underlying topic.

(c) Describe the resulting factors: list the 10 words with highest probability in each topic.

(d) Find, for each topic, the document (message) with highest probability of that topic, and print its 100 first words.

(e) Do the topics seem related to individual newsgroups? Does their content seem meaningful?

(f) Carry out (b)-(d) with 15 topics instead. Are some of the topics the same as before? What are the differences?

(g) Optional: let's see if an AI chatbot can analyze these topics. See exercise 2.4 regarding possible chatbots; report what chatbot you used.
Take one of the topics, e.g., the highest probability one, and provide its list of top words to an AI chatbot; you can tell it that the topic has been learned from the four newsgroups. Ask it to interpret the topic based on the list of words. Ask it also to generate a text that could be drawn from the topic. Does its interpretation of the topic seem meaningful?

Report your analysis results and your code.