Homework 5

STAT-GB.4310: Statistics for Social Data
Instructor: Patrick O. Perry

Due March 8, 2016

Application

We will fit a topic model to the `yelp-nyc-business.json` corpus. Or, optionally, apply the following steps to any other text corpus. If the corpus is too big, you may need to choose a random subset of the documents rather than using all documents in the corpus. You can refer to the lecture notes from March 1 when completing the assignment.

1. For each document in the corpus, remove all punctuation and numbers, and case-fold the text. You can do this by modifying the following commands:

   ```
   library("stringi")

   # convert to canonical case (lowercase for most languages);
   # normalize the unicode representation
   text <- stringi::stri_trans_nfkc_casefold(text)

   # remove punctuation and digits
   text <- gsub("[[:punct:][:digit:]]", "", text)
   ```

   If you would like, you can also choose to filter out certain words based on their frequencies or based on their POS tags.

2. Fit a topic model with 8 topics. Report the top 10 words in each topic. Based on these words, try to assign meaningful labels to the topics.

3. Use the document topic matrix to cluster the documents, using `kmeans`. (You will need to decide how to choose the number of clusters; there are many reasonable ways to do this.)

4. Pick a few of the clusters found by k-means, and look at some of the documents in each cluster. Do these groupings make sense to you? Why or why not?

5. Fit another topic model with 8 topics by running the command again. Did the topics change much? Try to come up with a rigorous measure for how much the topics changed.

6. Run `kmeans` again, this time using the second topic model. Using the `table` command, quantify the agreement between the clusterings from the two topic models.