Text-as-Data Exercise

In this exercise, you will work with Twitter data attributed to Russia's Internet Research Agency (IRA). You will practice cleaning and processing text data, creating document term matrices, and learn dictionary and topic modeling methods discussed in your assigned readings. Please be sure to set your working directory to your folder on Dropbox.

Data

The following dataset is provided:

- `ira_tweets_csv_hashed.csv`: tweets identified by Twitter as belonging to a Russian/IRA influence operation, released publicly in October 2018. Twitter's Readme file is also included for your reference.

Guidelines

1 Pre-processing and Visualizing Twitter Takedown Data

(a) Import the `ira_tweets_csv_hashed.csv` dataset from Dropbox.

- How many tweets are there in total?
- How many English and non-English language tweets are there?
- How many users have self-reported locations? What are the top 5 self-reported locations among these accounts?
- How many times do the words "Black Lives Matter" or "BLM" appear in the tweets?
- How many times do the tweets mention or reply to the official Twitter accounts of either Sputnik News or Russia Today (RT)?

2 Creating a Document Term Matrix

(a) Using only English-language tweets, create a document term matrix with unigrams (single words). Be sure to:

- Discard punctuation, capitalization, and word order
- Apply the Porter Stemmer
- Remove stop words from this list: http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/a11-smart-stop-list/english.stop
- Can you think of any other types of characters you should remove when working with social media data like tweets? If so, remove those.

(b) How many total words appear across all English-language tweets?

(c) Create a figure of the top 20 words and the number of times they appear in the data. Also create a word cloud of the top terms. Do the same for the top hashtags used.

3 Dictionary Methods

(a) Load the following lists of positive and negative words from Neal Caren:

- https://raw.githubusercontent.com/nealcaren/quant-text-fall-2014/master/positive.csv
- https://raw.githubusercontent.com/nealcaren/quant-text-fall-2014/master/negative.csv

(b) Count the number of positive and negative words for each tweet using these dictionaries, as well as the difference between the two scores. Report the summary statistic of the sentiment of IRA tweets.

(c) Plot a time-series of the average tweet sentiment per month.

4 Topic Models

(a) What is Latent Dirichlet Allocation (LDA) and how does it work?

(b) Identify the 300 most common unigrams and create a N x 300 document term matrix where the columns count the unigrams and the rows are the tweets. Run a LDA on this DTM and print the top 10 words for each topic.

(c) What is your best guess of the different topics that appear in the tweets pushed out by IRA trolls based on the frequent words?