POLI SCI 490: Machine Learning & Text-as-Data

HW2

Sarah B. Bouchat

DUE 7 February 2018

1 Data collection

1.1 Scraping

a. Install the `xml2` and `rvest` packages. Use the `read_html` function to get the code of the ASEAN membership categories page here.
b. Use the SelectorGadget plugin and `html_nodes` to get the "nodes" for the links to the country pages.
c. Make a dataframe with the country names and the links to the pages (using `html_text` and `html_attr`)
d. Using those links, write a loop that 1) gets the code for the page (using `read_html`), 2) extracts the text from the paragraphs on the page, 3) collapses it into a single string, and 4) saves it to the dataframe.

2 Pre-Processing & Word Frequency Analysis

2.1 Pre-Process

Use Trump's tweet data (on Canvas) for this section.[1]

a. Load the Trump tweets into R.
b. Pre-process these data using either `tm` or `tidytext`. (Discard punctuation, remove capitalization, remove stopwords, remove sparse terms to .01, tokenize, stem)
c. Contruct a document-term matrix.
d. `Tidy` the term matrix or otherwise standardize it for analysis.
e. Create a tf-idf matrix.

2.2 Word Frequency/Dictionary Methods

a. Plot the 20 most commonly occurring terms across the tweets.
b. Split the data into pre/post-election sets. Now re-analyze and plot the 20 most common terms for each set. How do they differ?
c. Suppose now that you'd like to assess the frequency with which Trump uses specific hashtags. Notice that the `#` that signals a hashtag was removed in your preprocessing step that eliminated punctuation. Regret this immensely. Pre-process the data again to preserve only `#` and eliminate other punctuation (.,; etc.).
d. With your differently pre-processed DTM, evaluate the frequency *only* of hashtags Trump has used: what are the top 5 most-used over the entire time period?
e. Plot the frequency of these top 5 hashtags over time using `ggplot2`.
f. Using bigrams rather than unigrams, report the frequency with which Trump used the phrase "Crooked Hillary" over time (by month).
g. Suppose I want to know if the words associated with the greatest number of "likes" of a tweet are different from the words associated with the greatest number of retweets.

[1] You can view a great tutorial here about how these data were collected from a New York Times article. Links to the analogous tutorial for python are also instructive if you'd like to try your hand at that.

- Generate the number of tweets in which a given word appears.
- Generate variables that store the number of RTs and number of “likes” for the tweets each word appears in. (*Hint*: Be careful about words that occur > 1 time per tweet.)
- Generate variables that store the average RT and “like” rate for a word.
- Report the top 10 words associated with the greatest average retweet and like rate respectively. How do they differ?