Debating with Quanteda

In the second problem set, you will study the US presidential debates using the quanteda package. Since R is a collaborative software, you will start with a few snippets of code to clean the debate fragments, done by last quarter QTA students.

1) Install the quanteda package if you have not done so.
   [A quick introduction into quanteda is provided here: https://cran.r-project.org/web/packages/quanteda/vignettes/quickstart.html]

2) Use the starting code to read the presidential debate object.

3) A text fragment in the DEBATES corpus object is a response to a question or a rebuttal and can be thought of as a self-contained document. Explore to what extent the Heap's law applies for Trump vs Clinton. Is it stronger or weaker for either speaker? (tip: see the code used in class).

4) Analyze the evolution of lexical diversity across the candidates before and after they became their parties' respective candidates? [tip: the R script provided has a primary indicator variable for identifying the fragments that came from non-primary debates]

5) After exploring 4), do you have a hypothesis why patterns may be more or less pronounced between Trump vs Clinton? How could you test this?

6) Remove stopwords from the corpus.

7) Using the tokenize function, construct separate bi-grams for the Hilary Clinton / Donald Trump parts of the corpus. Tabulate the ten most frequent bigrams by speaker. Are these informative? Why or why not?

8) Using the collocation function in quanteda (which takes a tokenize object as argument), construct collocations based on Chi2 test for each speaker. Order by Chi2 test statistic. What do you notice or what is strange? Can you provide a formal reasoning relating to the Chi2 test statistic formula?

9) Using the code provided in the lectures, identify collocations that are distinct to Trump vs Clinton. Based on your perception of each candidate, do the results of this analysis make sense? Provide a brief answer where you summarize the most important results and your explanations.

Send your PDF version of the Markdown document to

thiemo_837e@sendtodropbox.com

Please use the following naming convention: HW1_Surname_FirstName.pdf
Deadline: by Friday 28th April, 2017.