PPOL 628: Computational Linguistics for Social Scientists Final Exam

Due Date: 5/2/20 by Midnight EST. Please email to: md1723@georgetown.edu. Please include your name at the top of the document!
Page limit: 3 pages single spaced, with 1 inch margins, size 12 font. You do not need to hit this page limit. Acceptable answers can be shorter. That said if you find yourself only going half way down the second page, you may want to expand your answers a bit. Please don't spend more than a few hours on this. I am not out to get you with grading and don't want this to be a huge drag!
General Instructions: This is an open note exam. It is fine to look things up on the internet. However, this is not a collaborative assignment. Please do not discuss your answers with other students, or get help from friends or mentors. In general, please do not copy any material from websites or directly out of documents written by anyone else without an explicit citation.
Please answer all three questions.

For each question please keep in mind the following:

1. If a reading or lecture discussed something relevant to the question it is a great idea to cite it, but don't go overboard with citations.
2. **I really want you to provide your own intuitions and experience**: what would you do? Why would you do it? What have you learned in labs/your project that influenced your thinking? Providing a very brief example from your own work is also a good idea if it supports a point!
3. At the end of your answer it is ok to include a question or two you might want the answer to if you had to answer this question as part of your real life work.

Exam Questions

Question 1: Suppose a colleague comes to you with a corpus of 10,000 newspaper articles about politics from the New York Times. They would like you (the resident expert in computational linguistics) to preprocess the corpus for them into a document term matrix so they can statistically analyze it. **Discuss two preprocessing choices** (e.g. whether or not to remove punctuation, whether to keep numbers, etc.) that might have important consequences for their analysis, what your first intuition would be for these choices (e.g. "I would remove numbers because YYY"), and what you might want to know about the data to make these choices.

Question 2: This same colleague comes back to you after you have preprocessed their corpus of 10,000 newspaper articles about politics from the New York Times into a document term matrix. They have heard about this thing called "sentiment analysis" and think it sounds cool. They would like you to provide a **brief overview of dictionary-based sentiment analysis** and give them some pointers from your initial experience trying it out (even if it was just on the lab corpus). **What are some potential issues they might run into** (one or two is fine!), **and can you think of any potential solutions to these issues** (even if they would be experimental)?

Question 3: It's topic modelling time! Your colleague is back again with their corpus of New York Times articles and now they would like to understand the topical content of these documents. Please start by providing a very **brief intuition (in your own words) for how topic models work**. Next, they would like you to **suggest 2 hypotheses you could potentially explore with a (structural) topic model** (you can assume whatever you want about the authors, content, available metadata, and time frame of the newspaper articles). Finally, give them a tip or two for **validating their results, and/or something look out for that has come up in your own experience applying topic models**.