Project Outline

60% of your grade for this course will come from a final project on which you should work individually (subject to clarifications below). There are two required outputs: a paper and source code.

1 Paper

The first output should be a 10-20 page (shorter is better all else equal) paper with the following sections:

Data
Describe the text data you have collected. For example, how many documents, what is their associated metadata, the source of the data, the time period, etc. You may either use the data you collected for the homework assignments, or else download new data should you wish. Also, this section can be common among members of the groups you formed for your homework assignments.

Question
Ultimately, a theme of the course was to get you to think about not just how to describe text data, but also how such descriptions may be useful in the social sciences. Think about what interesting question your data might allow you to explore, and explicitly state what idea or behavioral model you are attempting to test. Also, argue why someone should care about your question: what implications does it have for how we think about the world?

Extracting Content
The text mining tools in this course all relied on a bag-of-words representation of text, and focused on content analysis. With this in mind, and staying focused on the question you are trying to answer, apply what we have learned to your text data.

First, pre-process the raw text data. Describe which stop words you use, whether you stem (or lemmatize), whether and how you choose to drop rare and common words, etc.

Second, choose one (or more) or the following content analysis approaches we discussed in class to apply to your data:

1. Boolean methods
2. Dictionary methods with and without tf-idf weighting
3. Vector space model / cosine similarity
4. Latent Semantic Analysis
5. Multinomial mixture model
6. Latent Dirichlet Allocation (posterior inference with Gibbs sampling and VBEM)
7. Supervised learning: Multinomial Inverse Regression; Naive Bayes; Supervised LDA.

Why do you choose the particular method(s) you do? Here I don't care so much about their computational properties, as I do about why they are, or aren't, appropriate for the issue you are exploring. Apply the method(s) you choose to generate a quantitative representation of each document.

NB: If you feel that an algorithm we haven't discussed is better than the ones we have, you may use it. In this case, your paper will need to include an additional section describing this algorithm so I understand its properties. In this case, your reward will be intellectual, and I will apply the same criteria for grading your project as I do for those that use algorithms discussed in the course.

Addressing Your Question
Now use your output to address the question you posed in the above section. There are several possibilities here that I leave up to you, but a basic approach would be to regress your content measures on various metadata fields. If you think there is exogenous variation in your data, argue why this is the case, and attempt to exploit this in your analysis.

Conclusion
To what extent do you think your analysis has successfully answered your question? What more data would you like to have? What additional text mining algorithms might need to be developed to provide additional evidence? In light of your results, state one message that you could comfortably state to a non-academic that would convince them your project is interesting.

2 Code

The code you use for the "Extracting Content" section should be written in Python. It should take as an input raw text data and metadata in some organized file structure, pre-process it, and generate content. You may use whichever libraries you wish, either your own or those written by others.

The code you use for "Addressing Your Question" can be written in Python or R, and again rely on whichever libraries you wish. It should take as an input the output of the "Extracting Content" code, perform statistical analysis, and generate necessary tables and graphs. Tables and graphs in your paper that are not produced by this code will not be treated as legitimate.

Code for both sections should be properly documented so that someone else can understand what is happening in each section. Code that will not run on my machine is not acceptable, so any additional instructions should be included in a README.

3 Deadline

The deadline for turning in the project is 5:00pm on Monday 27 June. Each person should create a sub-folder in a Box folder (to be created) with the name "project_lastname_firstname." Within this file, place your paper and your source code.