Data Analytics for Social Science
Project 3: Text data

Johan A. Elkink
jos.elkink@ucd.ie

due 9 May 2017

This assignment is about text data, where we perform statistical analysis on text, or more precisely, on document-term matrices that capture the frequency of occurrence of each term in each document. In this case we have State of the Union addresses by US Presidents in 1945. Note, however, that some of these are not State of the Union speeches proper - for example, while Trump has given a speech to the joint chambers of Congress, he has not yet given a real State of the Union speech, but I have included his speech. Some speeches were written only and sometimes there was a second speech in the same year. More details can be found at
`https://archive.org/details/State-of-the-Union-Addresses-1945-2006` and
`http://www.presidency.ucsb.edu/sou.php`.

In addition to the speech data, I have included a small data set with a few additional variables, namely the year, the president's name, the party of the president, and a label that is somewhat easier to read in plots. Both the speech data and the additional data are opened at each lab session. Be careful, though, that the speech data can include in some instances the Convention speeches by Cruz and Sanders, which in many cases should be excluded from the analysis.

The task for the project is roughly to look into two questions:

1. What do presidents talk about in speeches to joint sessions of Congress since 1945? Are there any noticeable patterns?

2. Can we use the statistical text data to detect the ideological positioning of presidents, which can include left-right, liberal-conservative, nativist-globalist, etc.?

You can choose from the various topics each week, cluster analysis, principal components analysis, factor analysis, wordscores, wordfish, or topic modelling, which are most relevant or interesting to you—there is no need at all to cover all of them. One or two will do. You can get away with only answering one of the two questions, if you can write an interesting essay around it.

In principle, there should be no need for analysis beyond what is done in the labs. The focus is on the interpretation and putting it in context, ideally with relevant references on, for example, the literature on ideological mapping.[1] Follow the outline in the

[1] Note that ideological space is also called "ideal point" and "spatial model of politics" in the literature,

syllabus: “Each essay should consist of a short introduction, a description and motivation of the data and methods used (approximately 25% of the essay), the analysis including necessary graphs and tables (approximately 35%), and an interpretation and conclusion (approximately 40%). Everything needs to be properly referenced.” The total length, excluding tables and bibliography, should be between 2,000 and 2,500 words.

since we are thinking of an ideological space, in which individuals and parties have a particular ideal point that reflects their ideological perspective on politics.