Content analysis: Worksheet 1

Trinity Term 2015

1. Use the Google n-gram viewer to assess Bloom et al's measure of policy uncertainty.

   - Are the patterns in word usage they find reflected in fiction?
   - Would it invalidate their measure if so?

2. Load the CSV located at `http://andy.egge.rs/data/candidates.csv` into `R` or Stata. The file contains the surname, party, vote total, and brief biography of each candidate in the 1950 Bexley general election.

   - Create a variable called `banker` that takes the value 1 if the word "bank" appears in the candidate's bio and otherwise 0.
   - Create a variable called `age` that records the age in years of each candidate, if that is given in the bio.
   - Create a variable called `yob` that records the year of birth of each candidate, if that is given in the bio.

3. Write code (in `R`, `python`, or another language of your choice) to conduct a Google search for several search terms and record the number of hits in a delimited text file (e.g. CSV). (Hint: note that visiting the url `https://www.google.co.uk/?#q=search+term` will give the search results for "search term".)

4. Write code (in `R`, `python`, or another language of your choice) to download and store the contents of the Wikipedia page for each Oxford College. (You should be able to do this with a loop.) Store the output in a delimited text file (e.g. CSV), with one row per college.

5. Load the spreadsheet of college data into `R` or Stata and, using regular expressions,

   - count the number of times the word "undergraduates" (plural) appears in the document
   - count the number of times the word "undergraduate" (singular) appears in the document
   - count the number of times the word "graduate" appears in the document

- count the number of words in the document, apart from html tags like `<td>`
- extract the number of undergraduates