Content analysis: Worksheet 1

Trinity Term 2016

Regular expressions

Regular expressions allow you to match text in flexible ways; it can be useful for counting things in text, checking for the presence of things in text, dividing text into chunks, etc. Regular expressions are implemented in almost every programming language. There are many resources online for learning regular expressions in Stata or R (even more in Python, Ruby, Perl, etc).

1. 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.
    - Create a variable called `female` that records whether the candidate is male or female. (Hint: use a regular expression that focuses on the first word of the biography.)

2. Download the zipped CSV located at `http://andy.egge.rs/data/THC_candidates_1950.csv.zip`, unzip it, and load it into `R` or Stata. The file contains information about each of the major-party candidates standing in the 1950 election.

    - What proportion of candidates are female? (Hint: use a regular expression that focuses on the first word of the biography.)
    - Does the proportion of females differ across parties?

Web scraping

Web scraping is useful for gathering information from the web when gathering it manually would be possible but tedious.

1. 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".) See the lecture slides or Google for suggestions of packages for web scraping.

2. 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.

3. 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