Quantitative Content Analysis: Lecture 9

Matthias Haber

12 April 2017

Today's outline

- Regular Expressions
- Dictionary Approaches
  - Deriving a dictionary, "Wordscores version 0.1"
  - External dictionaries
  - Application
- Text analysis workflow
- Creating a text corpus
- Preprocessing
- Regular Expressions

Regular expressions

Regular Expressions (regex) are a language or syntax to search in texts. Regex are used by most search engines in one form or another and are part of almost any programming language.

You could use regex to e.g.:

- Count the occurence of certain persons/organization etc. in text
- Calculate the sums of fund discussed in legislation
- Chose your texts based on regexes

In textpreparation, regex are used to remove certain unwanted parts of text.

Rational for dictionaries

- Rather than count words that occur, pre-define words associated with specific meanings
- Two components:
  - **key** the label for the equivalence class for the concept or canonical term
  - **values** (multiple) terms or patterns that are declared equivalent occurences of the key class
- Frequently involves lemmatization: transformation of all in ected word forms to their "dictionary look-up form" – more powerful than stemming

Dictionary approaches

Dictionaries help classifying texts to categories or determine their content of a known concept.

- Which text pertain to which categories?
- Which texts contain how much of a concept?
- Compared to e.g. CMP
  - Dictionaries require knowing the semantic form of the concept
  - i.e. one would need a complete dictionary of left or right statements

Creating dictionaries

Creating Dictionaries

- Scheme of classification
- Documents with known properties or classification
  - Training Set: Used to construct a dictionary
  - Test Set: Used to test dictionary (properties/classification is known)
  - Classification Set: Text to be classified/scaled with the dictionary

Creating dictionaries (II)

Sequence of steps

- Collect the words that discriminate between categories/concepts, i.e. create a dictionary
  - Existing dictionaries
  - Creating a dictionary
- Quantify the occurence of these words in texts
- Validate

Creating dictionaries (III)

Methods (though not exhaustive)

- By hand
  - Based on a Training Set (Laver & Garry)
  - Based on a previously existing list or external Sources (Dodds & Danforth)
- Automatically (Wordscores)
  - Replaces the creation of a dictionary as in Laver and Garry 2000

Estimating Policy Positions from Political Texts

Laver & Garry

- Goal: Generating party positions for British and Irish manifestos
- Coding scheme similar to the CMP's
  - More hierachical, larger number of categories
  - Each category has a pro-, con- and neutral variant

Estimating Policy Positions from Political Texts (II)

Training Set

- Manifestos of Labour and Cons (UK) in 1992
  - Pool of 'keywords'
  - $N_L \geq 2N_R => $ Dictionary element left
  - $N_R \geq 2N_L => $ Dictionary element right
- Allocate selected words to the coding scheme's categories

Estimating Policy Positions from Political Texts (III)

- Count the occurence of the elements in the dictionary in manifestos
  - Britain (1992 & 1997)
  - Ireland (1992 & 1997)
- Left-right-scaling: $\frac{R-L}{R+L}$ (see Session 7 and assignment 2)
  - "Updating process"
  - $Econ_{LR}$
  - Soc_{LR}

Estimating Policy Positions from Political Texts (III)

- Test-Set: Crossvalidation
  - Expert Surveys
  - CMP Coding/Revised CMP Coding

TABLE 3 Pearson Correlations between Alternative Estimates of Economic Left-Right Scale Positions, Britain and Ireland 1992–97

	Computer Codings	Revised Expert Codings	Original MRG Codings	Expert Surveys
1992				
Computer codings	1.00			
Revised expert codings	0.85	1.00		
Original MRG codings	0.72	0.94	1.00	
Expert surveys	0.75	0.95	0.99	1.00
1997				
Computer codings	1.00			
Revised expert codings	0.94	1.00		
Expert surveys	0.91	0.95	n.a	1.00

Next Session

- Regular expressions
- Dictionary-based content analysis