Quantitative Text Analysis

Bruno Castanho Silva

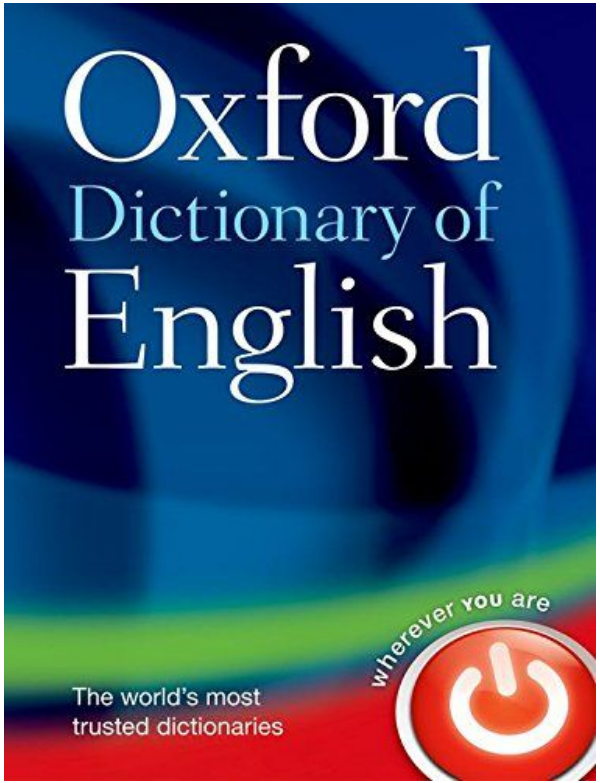
Day 2 - Dictionaries and sentiment analysis

Cologne Center for
Comparative Politics

Any questions from yesterday?

Dictionaries and sentiment analysis

What are they good for?



What are they good for?

- Classify texts into
  - Topics
  - Categories based on ideology/rhetoric
  - Style
  - Emotions
  - …
- Crafting and validating a dictionary takes time
- Implementation to corpus is quick and easy to scale

Dictionary-based approaches

- Counting pre-defined words associated with specific meanings
- Two components:
  - key: the label for the equivalence class for the concept or canonical term (e.g. “Economy”; “Anger”; or “Populism”)
  - values: (multiple) terms or patterns that are declared equivalent occurrences of the key class
- General idea: list of words that signal whether a text/sentence/paragraph deals with a certain topic

Constructing dictionaries

- A dictionary assigns a set of relevant values to each key (exclusivity)
- Concern: dictionary should have words that are indicative of the relevant category (precision), and have enough words so that the documents to be analyzed contain the dictionary words (recall)
- There is a trade-off between precision and recall.

Constructing Dictionaries - Strategies

- Hierarchy (e.g. domain, subdomain, etc.)
- Brainstorm words you think apply to a given key
- Identify extreme texts (e.g. policy speeches, positive/negative text, etc) - check if you missed something
- Identify discriminating words, look up texts to check sensitivity (do I miss relevant language?) and specificity (do I include irrelevant language?)
- Check if stemming/wildcarding is required (e.g. ”terror*”), use these words for dictionary categories

Constructing Dictionaries - Using them

Measures

- Proportion of words: How much of the text is about policy?
- Proportion of category matches: How much does the speaker speak about the environment vs the economy?

Results

- Classification: Is a text about the environment?
- Sentiment: How positive is a text?
- Personalisation: How personal (vs. professional) is a text?

Issues with dictionaries

- Humans are very good at recognizing correct answers but not great at listing them (this is where bias comes in)
- Dictionaries are often time and context dependent: words that signal policy areas might change over time and might differ across languages/countries
- It is easy to introduce bias by the researcher in constructing a dictionary. Example: Constructing an immigration policy dictionary and considering terms such as ”threat” and ”security”, but not ”help” and ”asylum” (or the other way round).

To R

Sentiment

Sentiment analysis

- Generally two or three categories: positive, (neutral), negative
- Widely applied in marketing: what are people saying of my brand on social media?
- Also useful for:
  - Negativity in media;
  - Measuring positions on certain issues;
  - Parliamentary debates and studies of coalitions;
  - Diplomatic texts
- Several different approaches are available

One (easy) option: Lexicoder dictionary

- Two categories: positive and negative
- Dictionaries contain words indicative of each category (e.g. *abhorrent*, *deceptive* for negative, *trustworthy*, *improving* for positive)
- Counts of words in each category.
- Aggregate: best to use (logged) ratio of positive to negative terms (confrontational categories, independent of document length)

Lexicoder

$$\text{Sentiment} = log\ \frac{0.5 + positive}{0.5 + negative}$$

- Only relative positivity and negativity matters
- Zero is either unemotional or equally positive and negative language
- Logged ratio leads to additional positive or negative words having decreasing marginal impact (Going from one to two negative words is more important than from 101 to 102).

Advantages

- Easily scalable for large corpora
- Limited set of clearly defined terms/values
- Works across languages (as opposed to left-right ideology or scaling generally)

Disadvantages

- Insensitive to short, very negative statements in the middle of a positive speech
- Relying on a clearly defined context: What are you supporting/opposing?
- Resulting scale not clearly interpretable for many political science questions

Combining with other dictionaries

- E.g. Heidenreich et al., 2020 - use a dictionary to filter sentences related to **immigration**, then sentiment to identify if speaker is in favor or against (sentiment as position)
- Same can be done with **EU** keywords, or **Covid restrictions**

Back to R