PPOL 628: Text as Data – Computational Linguistics for Social Scientists

Class 5: Dictionaries and Sentiment Analysis

Today

- Follow-up questions about phrases, rJava, etc.
- Lecture: key points from readings
- Reading discussion
- Lab: dictionaries.R
- Website: github.com/matthewjdenny/PPOL_628_Text_As_Data

Dictionary Based Methods

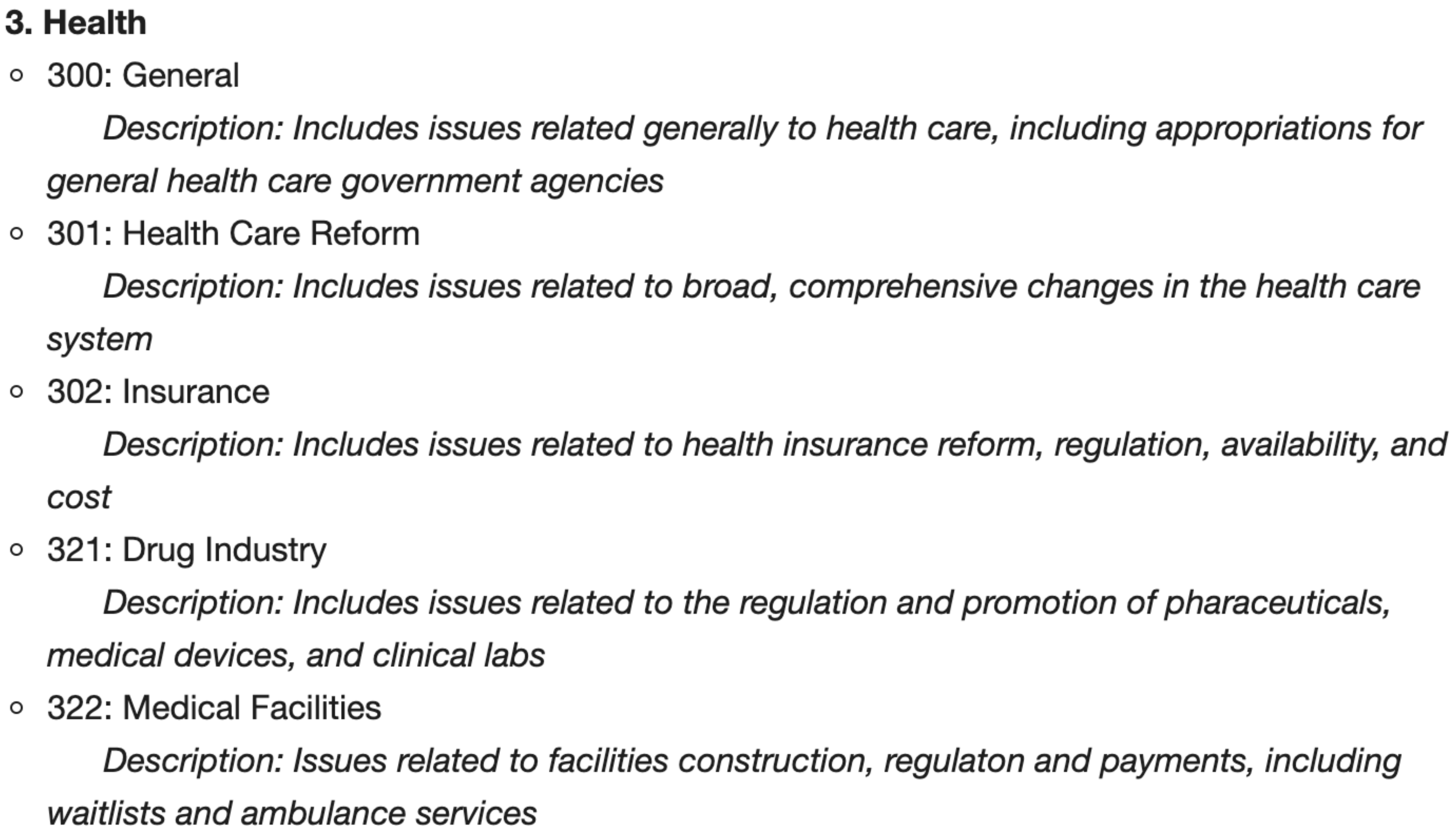
- How did we get here? Qualitative coding and analysis.
- The simplest dictionary method – grep.
- Constructing dictionaries.
- Supervised vs. dictionary vs. unsupervised methods.
- Sentiment analysis.
- Some takeaways.

Hand Coding →

- Research question →
  - Relevant information in text →
  - Coding guidelines →
  - Data collection and validation →
  - Analysis

- One natural extension is to use hand coded documents as a training set for supervised methods.

- Another approach is to manually formulate dictionaries, then use counts of those words in documents as scalable approach.



3. Health

- 300: General

  Description: Includes issues related generally to health care, including appropriations for general health care government agencies

- 301: Health Care Reform

  Description: Includes issues related to broad, comprehensive changes in the health care system

- 302: Insurance

  Description: Includes issues related to health insurance reform, regulation, availability, and cost

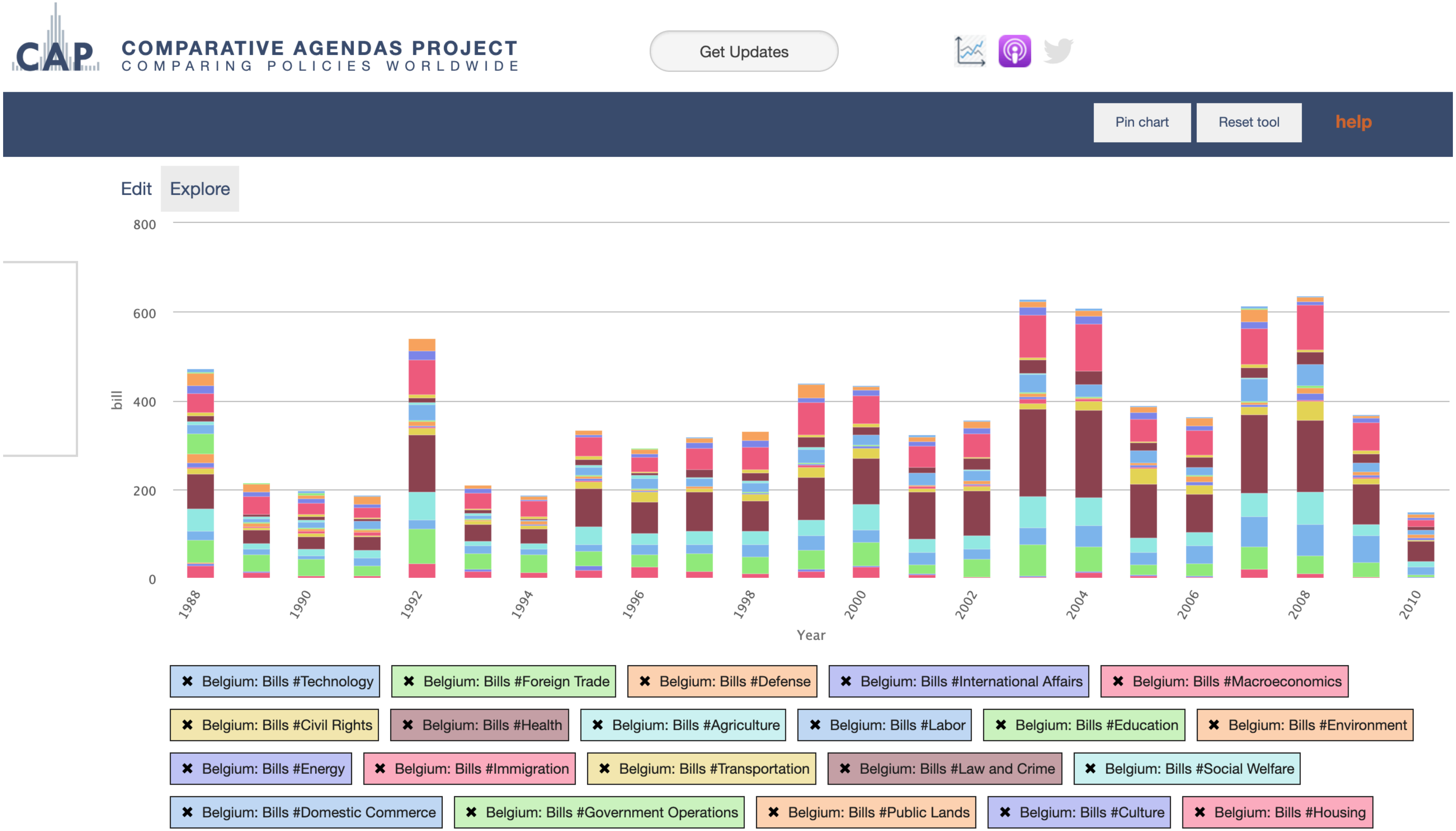
- 321: Drug Industry

  Description: Includes issues related to the regulation and promotion of pharaceuticals, medical devices, and clinical labs

- 322: Medical Facilities

  Description: Issues related to facilities construction, regulaton and payments, including waitlists and ambulance services

https://www.comparativeagendas.net/pages/master-codebook


CAP
COMPARATIVE AGENDAS PROJECT
COMPARING POLICIES WORLDWIDE
Get Updates
Pin chart
Reset tool
help
Edit
Explore
800
600
400
200
0
bill
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
Year
✖ Belgium: Bills #Technology
✖ Belgium: Bills #Foreign Trade
✖ Belgium: Bills #Defense
✖ Belgium: Bills #International Affairs
✖ Belgium: Bills #Macroeconomics
✖ Belgium: Bills #Civil Rights
✖ Belgium: Bills #Health
✖ Belgium: Bills #Agriculture
✖ Belgium: Bills #Labor
✖ Belgium: Bills #Education
✖ Belgium: Bills #Environment
✖ Belgium: Bills #Energy
✖ Belgium: Bills #Immigration
✖ Belgium: Bills #Transportation
✖ Belgium: Bills #Law and Crime
✖ Belgium: Bills #Social Welfare
✖ Belgium: Bills #Domestic Commerce
✖ Belgium: Bills #Government Operations
✖ Belgium: Bills #Public Lands
✖ Belgium: Bills #Culture
✖ Belgium: Bills #Housing

https://www.comparativeagendas.net/tool/fRWkfk8T

Dictionary Methods as Preprocessing and Analysis – grep

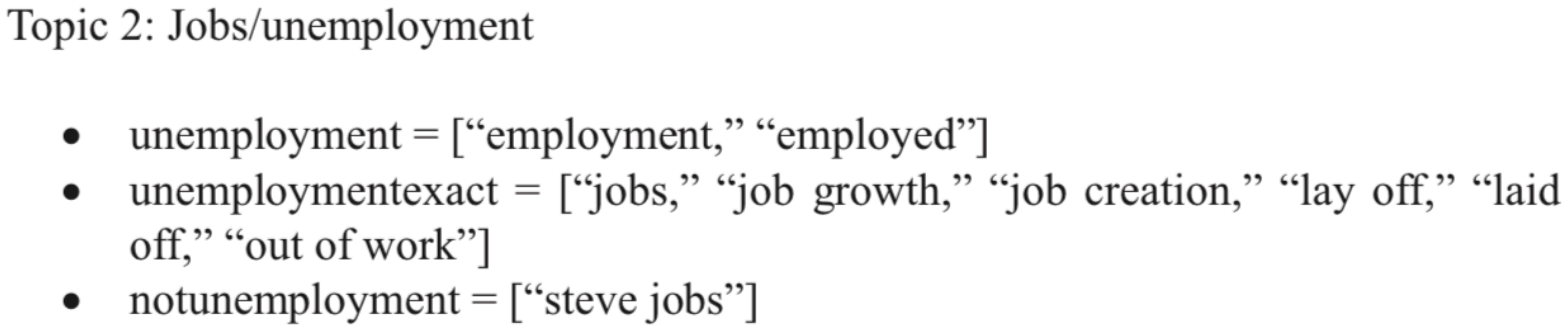
- grep has been around since 1974 – simplest form of dictionary method is to check for whether/number of times a single word appears in each document.

- Using grep()/stringr::str_extract_all() allows you to match arbitrary character sequences/regular expressions.
  - If all you care about is the use of a well defined set of terms, you just use this approach and have total control/flexibility over how terms are matched.

- Key challenge is in construction of dictionary.

Gou et al. Dictionary Creation

- A-priori – 16 major issue areas from literature review.
- Stem, remove punctuation, numbers, special characters, “stopwords”.
- Only look at words that appeared more than 1,000 times.
- Look at top terms and determine which ones should go with which issue areas:



Topic 2: Jobs/unemployment

- unemployment = [“employment,” “employed”]
- unemploymentexact = [“jobs,” “job growth,” “job creation,” “lay off,” “laid off,” “out of work”]
- notunemployment = [“steve jobs”]

Supervised vs. Dictionaries vs. Unsupervised

- **Supervised**: Hand coding documents and then training a model on features of those document to predict class of unseen documents.
  - If you have a good coding scheme and enough coders, works really well.
- **Dictionaries**: Come up with a list of terms and look for them.
  - Strongly depends on ability to generate good dictionary.
  - Better performance when looking for specific content over general classification.
- **Unsupervised**: Computer determines clusters of words that co-occur, user interprets.
  - Good for discovery, can feed into dictionaries, has its own problems we will learn about!

Sentiment Analysis

- **Basic Idea:** some words have positive/negative valence.
  - We can create a dictionary of words of each type.
  - Then count how many words of each type appear in each document.
  - Take difference in proportions.

- From Young and Soroka:
  - **"Net tone," our core measure of automated tone, is the proportion of positive words minus the proportion of negative words in an article**, that is: (# positive words/all words) − (# negative words/all words).24 So a score of −2.4 for crime means that, on average, in crime stories there is a 2.4-percentage-point gap between the number of negative words and the number of positive words.

Sentiment Analysis (continued)

- **Challenges and Limitations:**
  - Dictionaries mostly not portable to other languages.
  - Some words have different valence in different contexts (e.g. Twitter, News, Sports Television, young vs. old people, etc.)
  - Tricky to handle negation, sarcasm:
    - That was so **not cool**!
    - Yea that hat totally looks **“great”** on you.
- Emotional words may have little correlation to emotions in some domains (e.g. social media).
  - Beasley et al (2016) “Inferring Emotions and Self-Relevant Domains in Social Media: Challenges and Future Directions”

My Take

- With advent of topic models and other unsupervised methods, people have moved away from dictionary methods.
    - We should use dictionaries more often.
- Creating a dictionary can be an iterative process.
    - New field of CS research called “query expansion”.
    - Start with seed terms, use topic models for expansion, manual/KWIC checks.
- Dictionaries are especially useful when you have a limited number of categories you are interested in.
- Sentiment analysis via dictionaries: be careful with interpretation.
    - Still lots of work to do in this field, still very important.

The Readings This Week

- Young, L., & Soroka, S. (2012). Affective News: The Automated Coding of Sentiment in Political Texts. Political Communication

- Guo et al. (2016) Big Social Data Analytics in Journalism and Mass Communication: Comparing Dictionary-Based Text Analysis and Unsupervised Topic Modeling.