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

Class 7: Term-Category Associations

Today

- Class format for rest of semester.
- Lecture: key points from readings.
- Lab: term_category_associations.R
- Website: github.com/matthewjdenny/PPOL_628_Text_As_Data

Class Format Going Forward

- Plan is to record three videos per week and post to YouTube (unlisted) by Monday afternoon, so you have a chance to watch before Tuesday:
  - Lecture
  - Lab
  - Responses to Reading + General Questions

- Homework + final projects will continue as in the syllabus.

- In-class final will now be a 24 hour take-home with a strict (short) page limit to make sure you do not spend too much time on it.

Class Format Going Forward

- Live office hours on Tuesdays 7:00-8:30 EST.
  - You can ask questions on slack, via email, or over Bluejeans/hangout (by appointment).
  - Where necessary, I will try to record example videos if you have questions that require coding help in R.

- Reading Questions
  - Will now be due by Sunday morning so I have enough time to record a response video.
  - Can include coding/data/general questions as well.

Term-Category Associations

- Conditional on knowing categories for documents, can tell us what they are about, what makes categories different.
- PMI, TF-IDF, etc. all work just as well if not better with document categories vs. individual documents.
- Fightin' Words – model based inferences about which terms most strongly differentiate categories.
- ACMI – (possibly) principled "stop term" discovery?

Fightin’ Words

- We are given document categories, want to know the words that are most strongly associated with each category.
- Designed to work for comparing two categories – theoretically can generalize to multi-category case.
- Posits generative model for the text in documents, then ”reverses” that process during inference to back out estimates of generative process parameters.
- Dirichlet Prior is the most commonly used.

General Points – Dirichlet Model

- Useful for exploration and inference:
  - “what terms are the most Republican?”
  - Count terms that are statistically associated with a category across documents as a covariate of interest.

- As with TF-IDF (and as we will see, with topic models), the choice of prior can have a significant impact on results.
  - Means you need to come up with a prior before you conduct analysis so you are not p-hacking.

- The larger the prior (alpha), the more smoothing → model will select for words that are more exclusive to that category.

- Z-scores select for high frequency + exclusivity, odds ratio selects for exclusivity.

Term Contributions to Mutual Information

- Method for identifying domain stopwords.
- Builds on intuition that some terms will actually reduce Mutual Information of DTM/Contingency Table.
- Form a reasonable contingency table – want metadata defined categories that are strongly differentiated.
- Iteratively remove each term one by one and calculate change in MI.

DTM → Joint Distribution

- If we divide each (*i,j*) entry in a document-term matrix by the sum of counts of all terms in the dtm, we have an empirical joint distribution over documents and terms.
- We can think about an (*i,j*) entry in this join distribution as telling us the probability of picking word *j* in document *i* if we were to pick a random word from all words in all documents.

Category	"striking paragraph"	"opioid addiction"	"nuclear power"	"Affordable Care Act"	...
Democrat, Health Insurance	0.10	0.01	0.00	0.01	...
Republican, Health Insurance	0.15	0.01	0.00	0.03	...

Mutual Information

- Tells us for a given join distribution, how strongly the columns and rows are related.
  - The expected value of PMI over the entire joint distribution.
  - High MI means terms tend to give lots of information about documents/categories.

- In DTM context, can tell us about how “well” terms relate to documents.

$$I(C;V) = \sum_{c \in C} \sum_{v \in V} p(c,v) \log \left(\frac{p(c,v)}{p(c)\,p(v)} \right)$$

Mutual Information In Practice

- If terms generally have a stronger association with documents/categories, mutual information increases.
- If all terms used the same in all documents, then mutual information of zero. Bounded above by $I(C,V) \leq \min[H(C),H(V)]$.

Distribution 1

	"pursuant to section"	"fiscal year"
Democrat	0.25	0.25
Republican	0.25	0.25

$$I(C;V) = 4\ (.25 \times \log(1)) = 0$$

Distribution 2

	"repeal Obamacare"	"carbon tax"
Democrat	0.00	0.50
Republican	0.50	0.00

$$I(C;V) = 2\ (.5 \times \log(2)) = 0.693$$

Some Terms Reduce Information

- The inclusion of some terms in a document term matrix can actually decrease the mutual information of the joint distribution it implies.
- Possible method for identifying stop terms.

Distribution 1

	"section"	"birth control"	"insurance"
Democrat	0.36	0.08	0.00
Republican	0.20	0.02	0.14

$I(C;V) = 0.11$

Distribution 2

	"birth control"	"insurance"
Democrat	0.33	0.00
Republican	0.08	0.58

$I(C;V) = 0.428$

General Points

- Requires a specific data structure with some number of predefined topics, and then two different classes (like political parties) that we **expect to use words differently** conditional on topic.
  - Differential word use is the critical part, as is the two different categories.

- Only as good as the categories/assumptions holding.

- Higher recall, lower precision → possibly interesting preprocessing step for supervised learning.

- Simple to implement, complex to make feasibly fast enough.