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

Class 5: Corpus Description, TF-IDF

Today

- Lecture: key points from readings
- Reading discussion
- Lab: Term_Weighting.R
- Website: github.com/matthewjdenny/PPOL_628_Text_As_Data

Corpus Description

- document-term matrix → terms as document covariates.

- What do terms tell us about documents?
  - What terms are “important”/“informative” in this corpus?

- Quantitative measures of term importance:
  - Information theory – how much information do terms give us about documents?
  - Term scoring/weighting for queries – which document is most associated with a given term? ←→ what terms are most associated with a given document?

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	...

A View from Information Theory

- Information theory: study of quantification, storage, communication of information.
- Deep ties to probability theory → way to understand joint distributions, marginal distributions, relationships between random variables.
- Using tools from information theory, we can ask how much information the terms in a dtm give us about what documents they belong to, and vice-versa.
- We can also ask if a term is unusually associated with a given document, or another term.

Pointwise Mutual Information (PMI)

- Statistical association between two discrete random variables.
  - Can think of this as documents/categories $c \in C$ and vocabulary terms $v \in V$

- The PMI between any two categories tells us about how strongly associated they are.
  - High PMI means strong association, zero PMI means independence, negative PMI means anti-associated.

$$PMI(c;v) = \log\left(\frac{p(c,v)}{p(c)\,p(v)}\right)$$

Entropy of a Marginal Distribution H(x)

- Tells us about how “spread out” a marginal distribution is.
  - Higher entropy means a more uniform distribution, lower entropy mean more concentrated.

- In the context of a document term matrix, low entropy words are highly associated with an individual document/category, while high entropy terms tend to be stopwords.

$$H(X) = -\sum_{i} P_X(x_i) \log_b P_X(x_i)$$

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

$I(C;V) = 0.428$

A View from Information Retrieval

- If I am designing a system to help a user get the best search results, which documents should I show them for a given search term?
- Which terms are the most distinctive to documents? Which terms are most “interesting” to look at?
- Naïve approach would be to use raw word counts:
  - The document where a word occurs the most is likely to be the one that is most associated with that term.
- More sophisticated approach is to weight terms (either globally or within a document) based on how uniquely associated they are with that document.

Term Frequency

- The i,j entries in the document term matrix.

$$tf_{t,d} = \text{count of term } t \text{ in document } d$$

- Alternate weightings:
  - Perhaps we think that term “importance” should increase with the log of its raw count -- similar to logging GDP in economics analyses.
  - May also only care that a term appears at least once in a document (boolean counting)
  - May want to normalize against the average number of times terms that appear in that document appear.

Inverse Document Frequency Weighting

- A common way to emphasize likely meaningful terms is to weight term counts by the inverse of the number of documents that term appears in.
  - Terms that appear in many documents will have counts/scores down-weighted more than unusual/infrequent terms.

- Standard formulations:

$$\text{idf}_t = \log\left(\frac{N}{\text{df}_t}\right)$$ where N is num docs and df_t is num docs containing term t.

TF-IDF

- The tf-idf weighting scheme assigns to term **t** a weight in document d given by:

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t$$

  In other words, $\text{tf-idf}_{t,d}$ assigns to term **t** a weight in document **d** that is:
  - highest when t occurs many times within a small number of documents (thus lending high discriminating power to those documents).
  - lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal).
  - lowest when the term occurs in virtually all documents.

Manning et al. (2009), p. 119

IDF Formulations

- TF-IDF (no log) -- $dtm[i,j] \times \left(\frac{\text{num docs in corpus}}{1+\text{doc counts}[j]}\right)$
- TF-IDF -- $dtm[i,j] \times \log\left(\frac{\text{num docs in corpus}}{1+\text{doc counts}[j]}\right)$
- TF-IDF (smooth) -- $dtm[i,j] \times \log\left(1 + \frac{\text{num docs in corpus}}{1+\text{doc counts}[j]}\right)$
- TF-IDF (max smoothing) -- $dtm[i,j] \times \log\left(1 + \frac{\max(\text{doc counts})}{1+\text{doc counts}[j]}\right)$
- TF-IDF (prob log) -- $dtm[i,j] \times \log\left(1 + \frac{\text{num docs} - \text{doc counts}[j]}{1+\text{doc counts}[j]}\right)$

TF-IDF variants

Term frequency		Document frequency	
n (natural)	$\mathrm{tf}_{t,d}$	n (no)	1
l (logarithm)	$1 + \log(\mathrm{tf}_{t,d})$	t (idf)	$\log \frac{N}{\mathrm{df}_t}$
a (augmented)	$0.5 + \frac{0.5 \times \mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N - \mathrm{df}_t}{\mathrm{df}_t}\}$
b (boolean)	$\begin{cases} 1 & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$		
L (log ave)	$\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t \in d}(\mathrm{tf}_{t,d}))}$		

Manning et al. (2009, p.128)

General Points

- Metrics/measures from information theory can give us information about the consequences of different preprocessing specifications.
  - Potentially rigorous way to think about stop terms.
  - PMI a useful exploratory tool to find top terms.

- TF-IDF is a widely used weighting for DTMs in production tasks.
  - Will often improve classifier performance.
  - Tends to down-weight terms we think of as stop terms.
  - Flexibility in formulation allow for different weighting of terms to suit different applications – but less “rigorous” statistical foundations.

- These methods can be applied to documents or categories of documents.

The Readings This Week

- OConnor, B. (2014). MITEXTEXPLORER: *Linked brushing and mutual information for exploratory text data analysis.*
- Lewis (1992). *Feature selection and feature extraction for text categorization.*
- White (2016). *Bag of Works Retrieval: TF-IDF Weighting of Co-cited Works.*
- Also highly recommend: Manning et al. (2009). *An Introduction to Information Retrieval*. http://www-nlp.stanford.edu/IR-book/

Some Questions + Answers

- How confident are we about TF-IDF that it's not dropping important keywords that appear frequently and commonly in all documents?
  - Depends on use case. Good to look at raw TF as well as robustness check.
- Unigrams vs. Phrases vs. combination for top terms analysis via PMI, TF-IDF?
  - Balance between TF and IDF
- MiTextExplorer interoperability/R package?
  - We can replicates some functionality in R (in our lab!)
- How well does TF-IDF work with a small corpus vs. large?
  - These methods always work better with large corpus.