Document Distance and Topic Models

Elliott Ash

Bocconi 2018

Big Data, Big Analytics

- Massive increase in unstructured text due to:
  - new social structures (the internet, email)
  - new/improved data collection
  - digitization efforts (govt documents, Google)
- Tools to analyze text advancing in parallel
  - text by itself is useless
  - importing methods from many different fields
  - new analysis techniques can drive new data availability

Different Methods for Different Goals

- Supervised: Pursuing a known goal.
  - human annotates a subset of documents
  - algorithm annotates the rest
  - usually associated with quantitative research
- Unsupervised:
  - algorithm discovers themes/patterns in the texts
  - human interprets the results
  - usually associated with qualitative research
- Both strategies amplify human effort, each in different ways

How to measure cosine similarity

$$\text{cos\_sim}(v_1, v_2) = \frac{v_1 \cdot v_2}{||v_1||||v_2||}$$

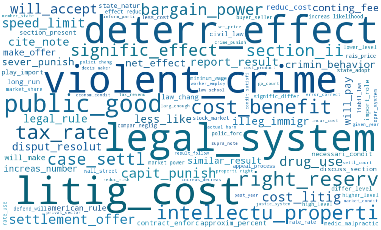
where v_1 and v_2 are vectors, representing documents.

```python
from sklearn.metrics.pairwise import cosine_similarity

sim = cosine_similarity(X[:100])
sim.shape
sim[:3,:3]

tsim = cosine_similarity(X_tfidf[:100])
tsim[:3,:3]
```

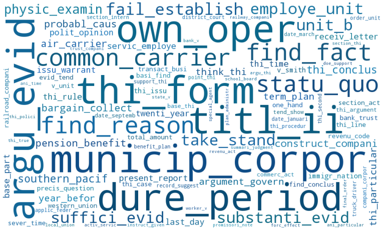
- Note that for n rows, this gives you $n \times (n-1)$ similarity scores.

Law-and-Economics Language (Ash-Chen-Naidu 2018)

- All available JSTOR articles with JEL K (Law and Economics) (1991-2008)
  - Highest and lowest frequencies for two-grams in $\geq$ 1000 cases:



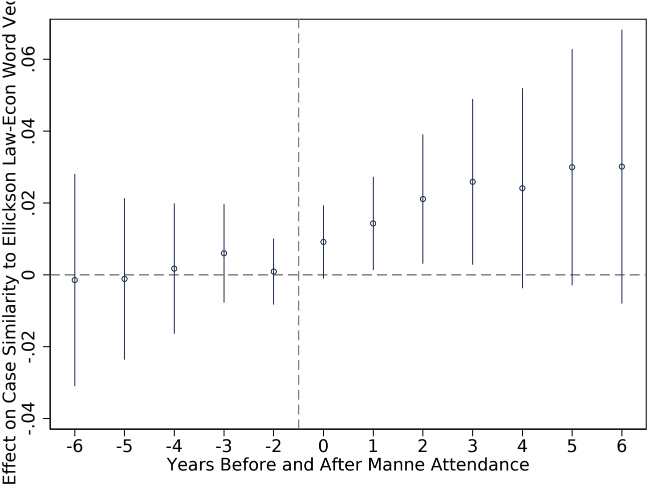
Most similar to Law-Econ Corpus



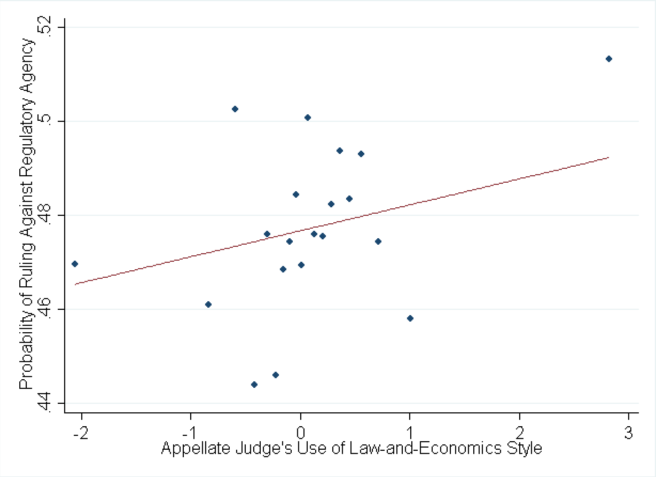
Least similar to Law-Econ Corpus

- Note: deterrent effect, cost-benefit, public goods, bargaining power, litigation costs

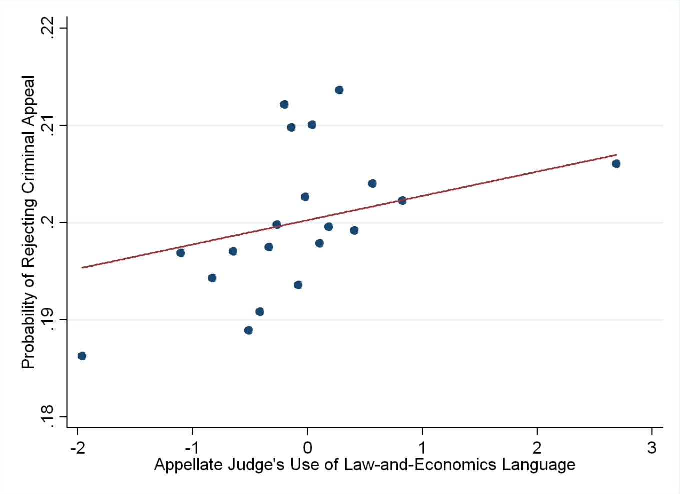
Economics Training Increases Economics Language Use



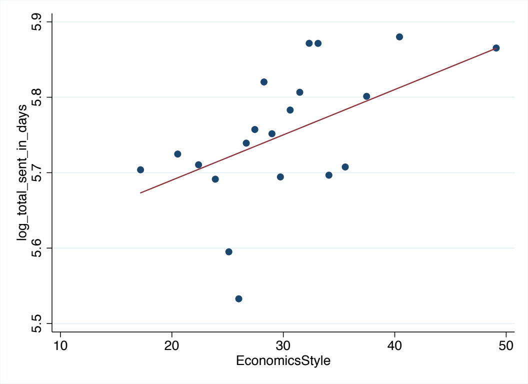
Judge Econ Language and Conservative Regulatory Decisions



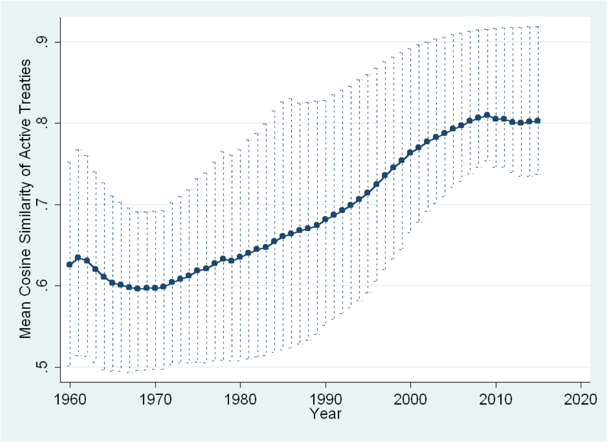
Judge Economics Language and Conservative Appellate Decisions in Crime Cases



Judge Economics Language and Sentence Length at Trial

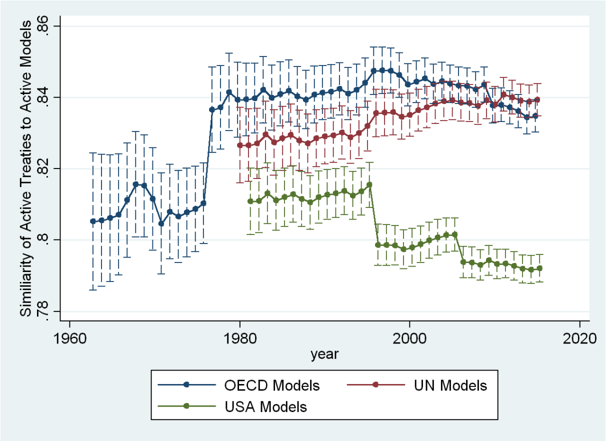


Tax Treaties have converged in their language over the last 100 years



Average cosine similarity between active treaties by year. Error spikes give 25th and 75th percentiles.

Influence of Model Treaties over Time



K-means Clustering

- With these notions of distance, we can use the K-means clustering algorithm to group related documents

- Given document vectors $\{\vec{q}_1, \vec{q}_2, ..., \vec{q}_P\}$, the algorithm chooses clusters $Q = \{Q_{1,} Q_2, ... Q_k\}$, $k > 1$, to minimize the within-cluster sum of squares:

$$\arg\min_{Q} \sum_{i=1}^{k} \sum_{\vec{q} \in Q_i} ||\vec{q} - \mu_i||^2$$

  where μ_i is centroid (mean vector) for cluster Q_i.

- Each cluster is a set of documents that are close to each other in the vector space (normally, they will be topically related)

K-Means Clusters

```python
from sklearn.cluster import KMeans
num_clusters = 100
km = KMeans(n_clusters=num_clusters, n_jobs=-1)
km.fit(X_tfidf[:1000])
doc_clusters = km.labels_.tolist()
dfs = df1[:1000]
dfs['cluster'] = doc_clusters
dfs[dfs['cluster']==1]['snippet']
```

- The advantage of clusters, rather than topics or embeddings, is that they provide discrete groups.
    - This might be useful depending on your research task.

Topic Models in Social Science

- Core methods for topic models were developed in computer science and statistics
  - used as a way to summarize unstructured text
  - use words within document to infer its subject
  - introduced as a form of dimension reduction
- A theory of use in social sciences
  - social scientists wanted to use topics as a form of measurement
  - we are often interested in how observed covariates drive trends in language
  - we want to tell a story not just about what, but how and why
  - topic models are more interpretable than other methods, e.g. principal components analysis.

Some example questions

- How do senators present their work to the public? What explains variation in representational style? (Grimmer 2013)
- Did electoral reform change the portfolio of issues addressed by politicians in Japan? (Catalinac 2016)
- What are the propaganda strategies of the Chinese government? (Roberts and Stewart)
- Do presidential candidates move to the center after the convention? (Gross et al 2013)
- How do central bankers respond to an increase in transparency over their discussions? (Hansen, McMahon, and Pray 2015)

Four Principles of Text Analysis (Grimmer/Stewart)

- Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful
  - Data generation process for text is unknown
  - Complexity of language:
    - Time flies like an arrow, fruit flies like a banana
  - Models necessarily fail to capture language, but may be useful for specific tasks
  - This brings emphasis on **validation**, to demonstrate that methods perform the desired task.

Four Principles of Text Analysis (Grimmer/Stewart)

- Principle 2: Quantitative Methods Augment Humans, Not Replace Them
  - Computer-Assisted Reading
  - Quantitative methods organize, direct, and suggest
  - Humans: read and interpret

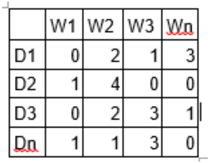
Four Principles of Text Analysis (Grimmer/Stewart)

- Principle 3:
  - There is no Globally Best Method for Automated Text Analysis
    - Supervised methods – known categories
    - Unsupervised methods – discover categories

Four Principles of Text Analysis (Grimmer/Stewart)

- Principle 4: Validate, Validate, Validate
  - Quantitative methods have variable performance across tasks
  - Few theorems to guarantee performance
  - So: Apply methods and validate
  - Avoid: blind application of methods

Document-term Matrix X

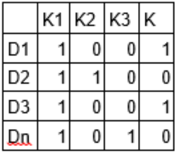


	W1	W2	W3	Wn
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
Dn	1	1	3	0

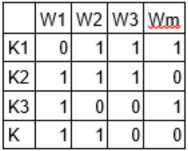
- A corpus of N documents $D1, D2, D3 \ldots Dn$
- Vocabulary of M words $W1, W2..Wm$.
- The value of i, j cell gives the frequency count of word Wj in Document Di.

Matrix factoring

- LDA converts the document-term matrix into two lower-dimensional matrices, $M1$ and $M2$:



	K1	K2	K3	K
D1	1	0	0	1
D2	1	1	0	0
D3	1	0	0	1
Dn	1	0	1	0



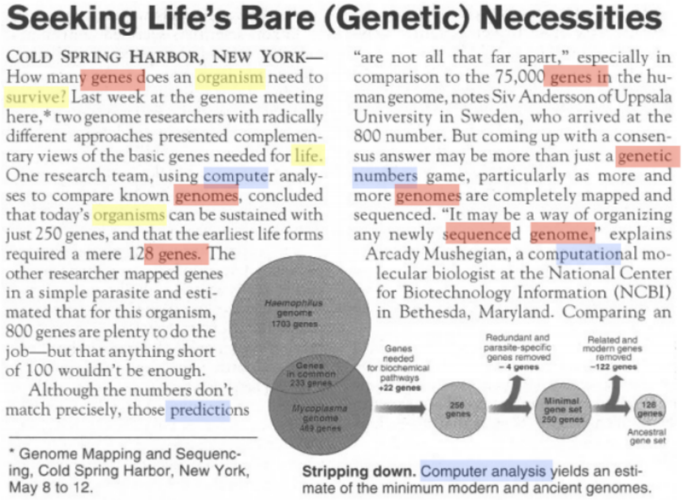
	W1	W2	W3	Wm
K1	0	1	1	1
K2	1	1	1	0
K3	1	0	0	1
K	1	1	0	0

- $M1$ is a $N \times K$document-topic matrix
- $M2$ is a $K \times M$ topic-term matrix.

Latent Dirichlet Allocation (LDA)

- Idea: documents exhibit each topic in some proportion.
  - Each document is a distribution over topics.
  - Each topic is a distribution over words.
- Latent Dirichlet Allocation estimates:
  - The distribution over words for each topic.
  - The proportion of a document in each topic, for each document.
- Maintained assumptions: Bag of words/phrases, and fix number of topics ex ante.

A statistical highlighter



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set

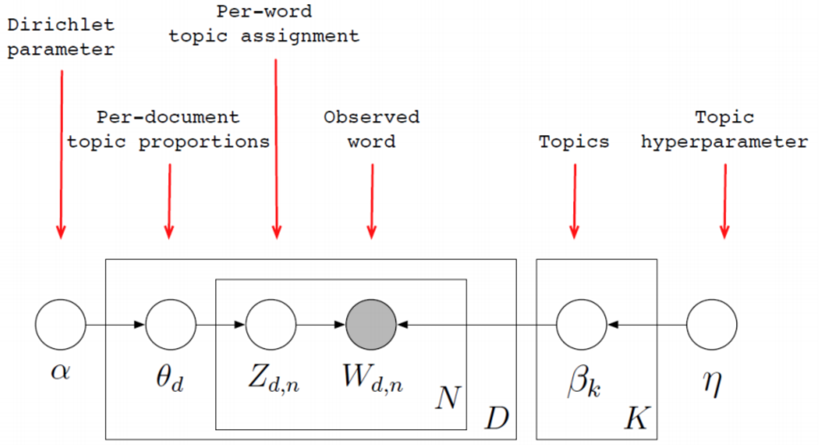
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Image from Hanna Wallach

A Bayesian Model

Figure: Plate Notation of Latent Dirichlet Allocation



LDA Parameters

- α: document-topic density
  - higher α means documents contain more topics, lower α means documents contain fewer topics
- β: topic-word density
  - higher β means topics have more words, while lower β means topics have fewer words
- Number of topics:
  - this is specified in advance, or can be chosen to optimize model fit.
  - the “statistically optimal” topic count is usually too high for the topics to be interpretable/useful.

Why does this work? Co-occurrence

- Where is the information for each word's topic?
  - We are learning the pattern of what words occur together.
- The model wants a topic to contain as few words as possible, but a document to contain as few topics as possible.
  - This tension is what makes the model work

LDA in Python

- gensim provides the best impementation of LDA in Python (gensim.models.LdaModel)
  - streaming (so works on arbitrarily large corpora), intuitive, fast (cython, parallelized)

Prepping data for gensim LDA

```
# randomize document order
from random import shuffle
shuffle(doc_clean)

# creating the term dictionary
from gensim import corpora
dictionary = corpora.Dictionary(doc_clean)

# creating the document-term matrix
doc_term_matrix = [dictionary.doc2bow(doc) for doc in doc_clean]
```

Running LDA

```
# train LDA with 10 topics and print
from gensim.models.ldamodel import LdaModel
lda = LdaModel(doc_term_matrix, num_topics=10,
               id2word = dictionary, passes=3)
lda.show_topics(formatted=False)
```

- "passes" is number of times to go through the corpus
  - probably doesn't matter for large corpora
  - if your topics differ significantly across runs, you need more passes.
- Once trained, can easily get topic proportions for a document:

```
lda[doc_term_matrix[0]]
```

Visualizing Topics

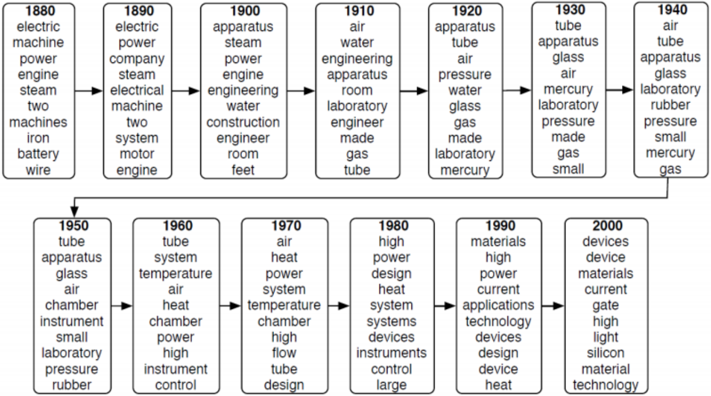
```
from wordcloud import WordCloud
for i,weights in lda.show_topics(num_topics=-1,
                                 num_words=100,
                                 formatted=False):
    wc = WordCloud().generate_from_frequencies(dict(weights))
```

Extensions

- There are a ton of extensions/variants LDA.
  - But almost all of them are very context-specific.
  - LDA is great because it works so well across different domains.

Dynamic Topic Model

Figure: Topic Evolution over Time

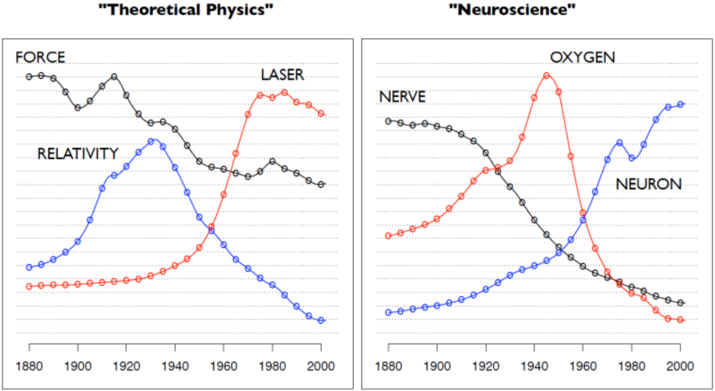


1880	1890	1900	1910	1920	1930	1940
electric	electric	apparatus	air	apparatus	tube	air
machine	power	steam	water	tube	apparatus	tube
power	company	power	engineering	air	glass	apparatus
engine	steam	engine	apparatus	pressure	air	glass
steam	electrical	engineering	room	water	mercury	laboratory
two	machine	water	laboratory	glass	laboratory	rubber
machines	two	construction	engineer	gas	pressure	pressure
iron	system	engineer	made	made	made	small
battery	motor	room	gas	laboratory	gas	mercury
wire	engine	feet	tube	mercury	small	gas

1950	1960	1970	1980	1990	2000
tube	tube	air	high	materials	devices
apparatus	system	heat	power	high	device
glass	temperature	power	design	power	materials
air	air	system	heat	current	current
chamber	heat	temperature	system	applications	gate
instrument	chamber	chamber	systems	technology	high
small	power	high	devices	devices	light
laboratory	high	flow	instruments	design	silicon
pressure	instrument	tube	control	device	material
rubber	control	design	large	heat	technology

Dynamic Topic Model

Figure: Word Use in Topics Over Time



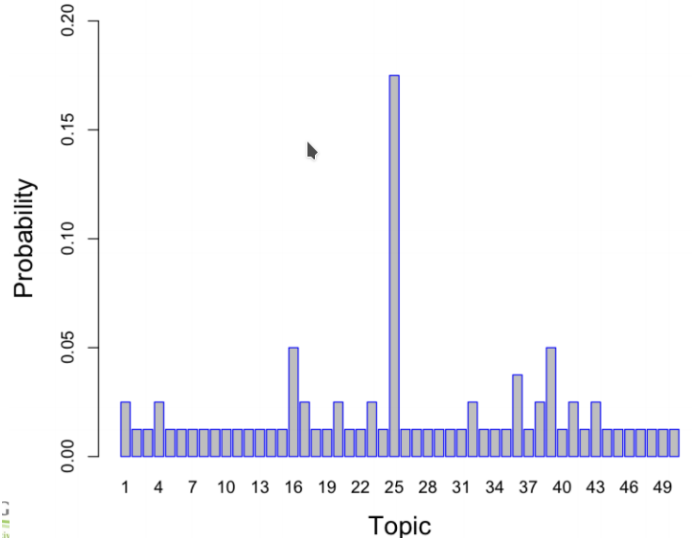
Non-text applications

- LDA can be used on any type of co-occurence data.
    - Bandiera, Hansen, Prat, and Sadun (2017): use LDA to do dimension reduction on CEO tasks using time allocation data.
        - model endogenously identifies a "leader" CEO topic, and a "micro-manager" CEO topic
    - Draca and Schwarz (2018): use LDA to dimension-reduce the features of political attitudes
        - model endogenously identifies "conservative" and "liberal" attitudes clusters.

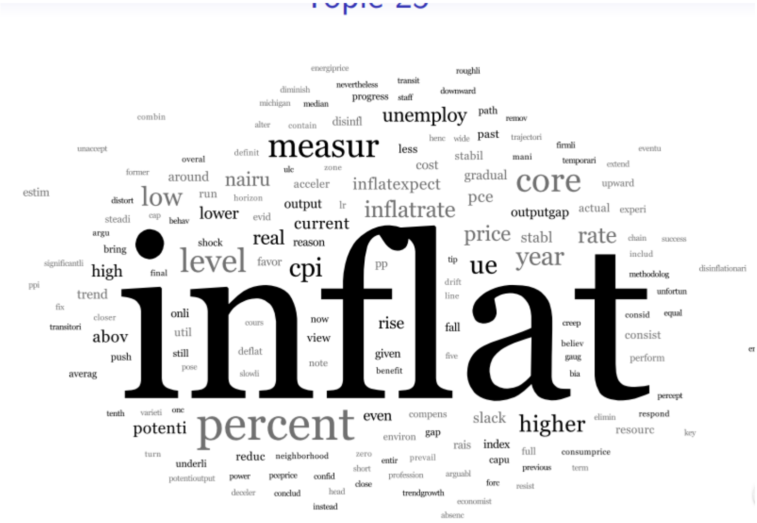
LDA on FOMC

- Hansen, McMahon, and Prat use LDA to analyze speech at the FOMC (Federal Open Market Committee).
  - 150 meetings, 20 years, 46000 speeches
- They take the standard pre-processing steps and train LDA.

Distribution of Attention



What is topic 25?



Overview of results

- They show that increasing transparency results in:
  - higher discipline / technocratic language (which is beneficial)
  - higher conformity (which is costly)
- Highlights tradeoffs from transparency in bureaucratic organizations.

Categorization of Union Contract Clauses

- Ash, MacLeod, and Naidu (2017): Recall the setup from earlier, where we represented contracts as a list of clauses:
  - $< Agent >< Obligation/Entitlement/Other >< Action >$.
- How to encode actions as data?
- LDA Approach:
  - We used the "action" segment of a clause (other connected pieces of the parse tree besides the subject, modal, special verbs, and stopwords) to classify each statement by topic using Latent Dirichlet Allocation (LDA, see e.g. Blei 2012).
  - We got good results with 20 topics.

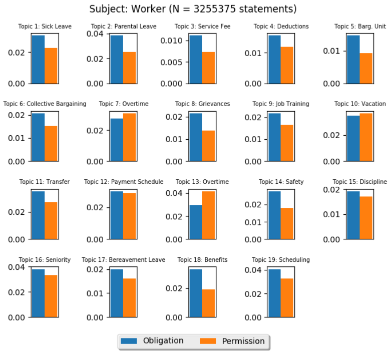
LDA Topics (1 of 2)

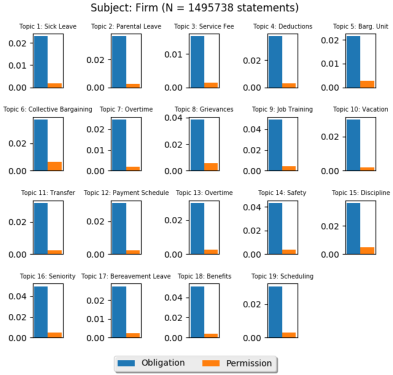
- 1 -- **"Sick Leave"** -- period month sick leave six probationary credit three complete employment twelve absent completion accumulate date exceed consecutive professional
- 2 -- **"Parental Leave"** -- leave absence pay request date grant prior week parental commencement pregnancy write maternity duty witness advance approve notice
- 4 -- **"Payroll"** -- change due result deduction amount status deduct monthly payroll reduction affect cheque technological fee employment orientation statement
- 5 -- **"Bargaining Unit"** -- unit bargaining person appointment appoint employ outside activity membership represent agent terminal sole select exercise ontario bargain behalf
- 7 -- **"Overtime"** -- hour shift work schedule overtime period call rest meal half minute start end break duty sunday weekend saturday two friday
- 8 -- **"Grievances"** -- grievance party procedure arbitration writing decision write step matter arbitrator committee complaint submit final dispute request name process
- 9 -- **"Job Training"** -- requirement operation training require equipment individual meet service responsibility provide program area manner performance" business duty operational
- 10 -- **"Vacation Leave"** -- year vacation service pay date employment week continuous effective two annual entitlement percent january salary earn termination period follow

LDA Topics (2 of 2)

- 14 – "**Medical Leave/Injuries**" medical reasonable illness reason certificate unable duty injury course require due provide information circumstance accident personal condition examination reasonably
- 15 -- "**Discipline/Firing**" -- school act safety committee health action discharge labour cause discipline disciplinary file application canada public relations suspension regulation authority accordance
- 16 -- "**Seniority**" -- seniority lay position list layoff vacancy recall transfer post temporary qualification permanent job hire fill date provide ability copy basis
- 17 -- "**Work-Related Deaths**" – article accordance law child spouse pursuant family death include immediate parent purpose require city office paragraph funeral
- 18 -- "**Insurance/Benefits**" -- benefit plan insurance payment cost premium eligible provide receive compensation disability pay coverage pension receipt term amount
- 19 -- "**Scheduling**" -- work hour day week schedule two return perform normal regular report normally excess regularly require notice eight teaching available emergency

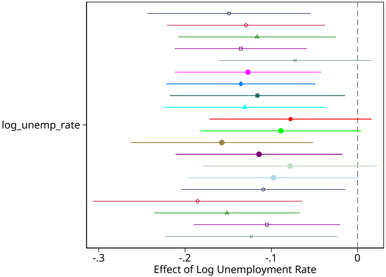
Workers Have More Entitlements Relative To Obligations





- Workers get more authority at work than employers, consistently across work areas.

Which contracts topics change in response to change in local unemployment?

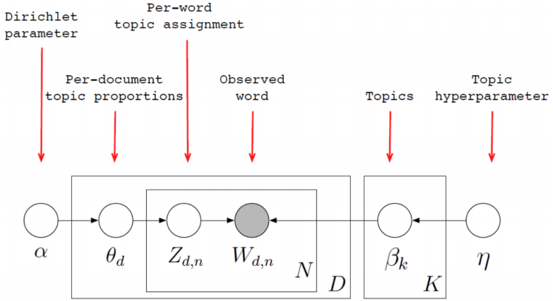


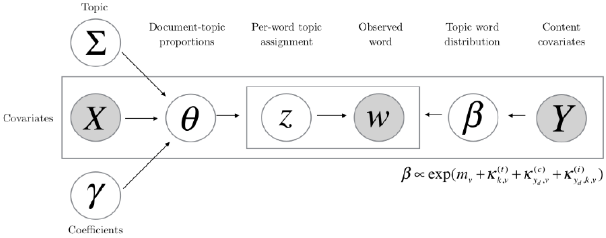
- Topic 7 (turquoise, square) is “Payment Schedule”; Topic 9 (bright red) is “Job Training”; Topic 10 (bright green) is “Vacations”; Topic 16 (dark red, diamond) is “Seniority”.

Structural Topic Model = LDA + Metadata

- STM provides two ways to include contextual information:
  - Topic prevalence can vary by metadata
    - e.g. Republicans talk about military issues more then Democrats
  - Topic content can vary by metadata
    - e.g. Republicans talk about military issues differently from Democrats.

LDA vs. STM – Illustration





$$\beta \propto \exp(m_v + \kappa^{(t)}_{k,v} + \kappa^{(c)}_{y_d,v} + \kappa^{(i)}_{y_d,k,v})$$

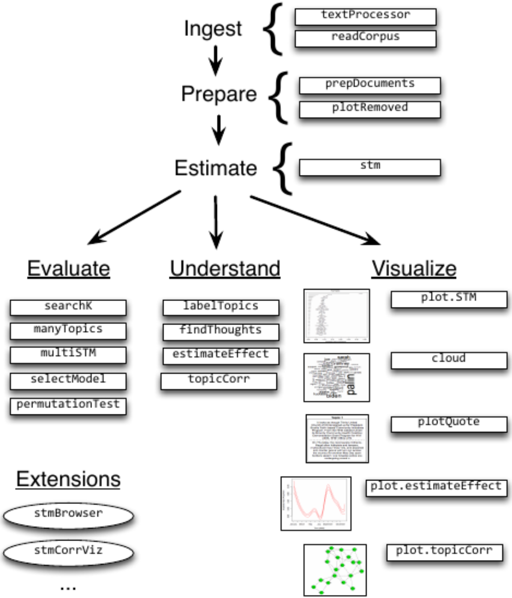
stm Package in R

- Complete workflow: raw texts → figures
- Simple regression style syntax using formulas

```
mod.out <- stm(documents,vocab, K=10, prevalence= ~
paper + s(time), data=metadata, init.type="Spectral")
```

- many functions for summarization, visualization and checking
- Complete vignette online with examples

stm has great functions/features



Ingest
- textProcessor
- readCorpus

Prepare
- prepDocuments
- plotRemoved

Estimate
- stm

Evaluate
- searchK
- manyTopics
- multiSTM
- selectModel
- permutationTest

Understand
- labelTopics
- findThoughts
- estimateEffect
- topicCorr

Visualize
- plot.STM
- cloud
- plotQuote
- plot.estimateEffect
- plot.topicCorr

Extensions
- stmBrowser
- stmCorrViz
- ...

Caveats

- Structural topic model is not a prediction model:
  - it will tell you which topics or features correlate with an outcome, but it will not provide an in-sample or out-of-sample prediction for an outcome
- STM does not work with streaming data (yet)
  - have to load the whole corpus into memory