Social Science Methods for Legal Scholars

A Survey of Quantitative Text Analysis Methods

Joshua C. Fjelstul, Ph.D.

Post-Doctoral Research Fellow, University of Geneva

Researcher, ARENA Centre for European Studies, University of Oslo

What are we going to cover in this workshop?

- What is text analysis? What methods does it include?
- How does text analysis relate to statistics, machine learning (ML), and natural language processing (NLP)?
- How can we use text analysis to study international legal texts?
- How can we implement some basic text analysis tools in R?

What is quantitative text analysis?

- Quantitative text analysis is the use of quantitative methods to analyze the content of documents

Levels of analysis

- Corpus
- Documents
  - Books, documents, judgments, speeches, tweets
- Paragraphs
- Sentences
- Words

Assumptions

- The content of text reveals something meaningful and interesting about the author that we can use to answer research questions
- Text can be represented by features/tokens (words, lemmas, *n*-grams)
- The relative distribution of words in documents captures meaningful variation in topics and substantively important latent dimensions
  - We make a matrix called a document feature matrix (DFM) with words in columns, documents in rows, and frequencies in cells that describe these distributions
  - This might seem unsatisfactory
  - Some kinds of neural networks can take into account word order
  - Taking into account word order adds a lot of complexity and the ”bag-of-words” approach usually works well

Types of quantitative text analysis

- Frequency analysis
  - What words/tokens are important?
  - Descriptive
- Similarity
  - How similar are documents/paragraphs/sentences?
  - Descriptive

Types of quantitative text analysis

- Scaling
  - How do documents/paragraphs/sentences compare on a continuous latent dimension?
  - Inferential (statistical models, machine learning models)
- Classification
  - How do documents/paragraphs/sentences cluster into discrete groups?
  - Inferential (statistical models, machine learning models)

Natural language processing (NLP)

- Text analysis overlaps with NLP — using computers to analyze language
  - Language processing
    - Optical character recognition (OCR)
    - Speech recognition
  - Morphology
    - Lemmatization
    - Stemming
    - Part-of-speech (POS) tagging
  - Semantics
    - Sentiment analysis
    - Named entity recognition
    - Word-sense disambiguation
  - High-level tasks
    - Machine translation
    - Natural language generation
    - Question answering

Machine learning (ML)

- Machine learning is the use of algorithms to predict outcomes based on data
- Algorithms learn the relationship between input data and labels based on training data and then make predictions for unseen data
- Quantitative text analysis is an application of ML to text data

Learning

- Supervised learning
  - Train a scaling/classification model on a pre-defined dimension using pre-coded training data
- Unsupervised learning
  - Train a scaling/classification model to endogenously learn an underlying dimension or categories using training data

Learning

- Supervised learning
  - Advantages
    - You already know the topics/dimensions so you know how to interpret your estimates
  - Disadvantages
    - It requires labeled training data (a lot of work)
    - You have to know ahead of time what the relevant topics/dimensions are

Learning

- Unsupervised learning
  - Advantages
    - You can explore naturally occurring topics and primary latent dimensions
    - You don't have to create labeled training data
  - Disadvantages
    - You have to validate that the topics/dimensions you uncover are meaningful and you have to figure out how to interpret them
    - You might not find meaningful topics/dimensions (or the ones you expect)
    - Unsupervised methods are harder to learn

Parametric vs non-parametric models

- Scaling and classification models can be parametric or non-parametric
- Parametric models overlap with statistics and involve estimating the values of parameters in statistical models that you can interpret to learn about the content of documents
- Non-parametric methods overlap with machine learning and depend on algorithms to learn the relationship between text data and labels

Classification

- Supervised methods for text
  - Naive Bayes classifier (non-parametric)
  - Regression classifiers (parametric)
  - Random forests (non-parametric)
  - Neural networks (non-parametric)
- Unsupervised methods for text
  - Latent Dirichlet allocation (LDA) (parametric)
  - Seeded LDA (parametric)
  - Structural topic models (STM) (parametric)

Scaling

- Supervised
  - Wordscores (parametric)
- Unsupervised
  - Correspondence analysis (non-parametric) (dimensionality reduction)
  - Latent semantic analysis (non-parametric) (dimensionality reduction)
  - Wordfish (parametric)

Topic models

- Topics models are unsupervised methods for identifying naturally occurring topics in documents
- The most common type of topic model is latent Dirichlet allocation (LDA)
- It assumes that documents are mixtures of topics and that topics are mixtures of words
- Documents are not sorted into discrete categories
- A topic is a distribution of words
- The goal is to figure out which words are associated with which topics and which topics make up each document

Topic models

- The model estimates the probability that each topic applies to each document
- You have to tell the model how many topics to find
- You can look at the words that are most strongly associated with each topic
- Based on that list of words, we can label each topic
- Some topics will be more distinct that others
- If topics overlap too much, we may need fewer topics
- It topics are not distinct, we may need more topics

Word scores

- Supervised scaling method
- You start with a set of reference texts
- These need to represent the two poles of your latent dimension
- The reference texts are like a training set
- You calculate word scores based on the reference text and then use them to score the rest of the texts
- Each document will have a single score that represents its position on the latent dimension

Word scores

- Advantages
  - After you choose the reference text, it's fully automated
  - It scales all documents between the references texts at each end of the dimension
- Disadvantages
  - It really matters which documents you use as the reference texts
  - The dimension you define by choosing the reference texts may not be the dimension that explains the most variation in the content of your documents
  - It's hard to choose the most extreme documents without a lot of knowledge about the content of documents (hard when there are a lot)

Wordfish (intuition)

- Unsupervised scaling method
- The input data is a DFM
- We don't have to know the underlying dimension ahead of time
- So we have to show that our estimates capture a meaningful latent dimension
- Based on a poison distribution
  - The poisson distribution models counts of discrete events — like the occurrence of words in a document
  - Wordfish is a type of poisson scaling model

Wordfish (equation)

- The model equation is:

$$log\lambda_{ij} = \alpha_i + \theta_i\beta_j + \phi_j$$

- Parameters:
  - λ_{ij} is the expected frequency of a word *i* in document *j*
  - *i* indexes documents
  - *j* indexes words
  - θ_i is the latent position of document *i* that we want to estimate
  - β_j is the latent position of word *j* (strength of the relationship between the latent dimension and the frequency of the word)
  - α_i is a document fixed effect (controls for how long each document is)
  - ϕ_j is a word fixed effect (controls for how frequent each word is)

Wordfish (estimation)

- On the right-hand side of a regression equation, we have data and parameters
- The data is constant and we estimate the parameters
- But here, there's no data on the right-hand side, so what do we do?
  - We start with random values for all of the parameters
  - We hold ϕ and β (the word parameters) constant and estimate α and θ (the document parameters)
  - Then we hold α and θ (the document parameters) constant and estimate ϕ and β (the word parameters)
  - And we iterate back and forth
  - This is called expectation-maximization
  - Eventually, we'll converge to good estimates of all parameters

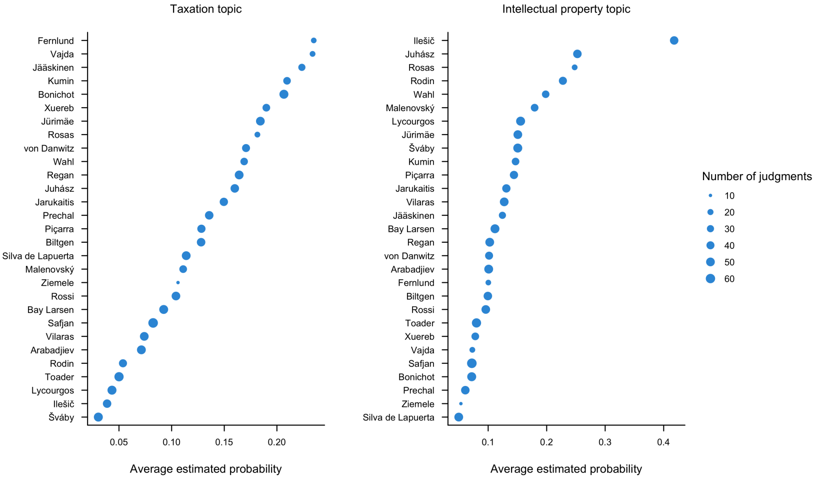
Wordfish (interpretation)

- We can interpret the θ estimates as the position of each document on our latent dimension. This is the main thing we're interested in
- But this is an unsupervised model, so how do we know what the latent dimension is?
  - We can plot the word fixed effects (y-axis) against the word positions (x-axis) to get an idea of what the dimension is
  - Rarer words (lower fixed effect) will be more discriminatory (will provide more information about how documents differ)
  - Words with more extreme positions (x-axis) will define the substantive content of each pole of the latent dimension
  - You have to convince your audience that your latent dimension is meaningful and interesting

Research questions

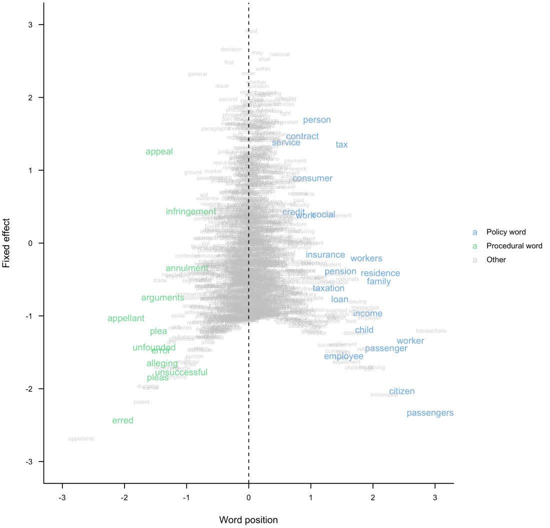
- We'll answer two research questions about the content of judgments delivered by the Court of Justice of the European Union (CJEU)
  - To what extent do judges specialize in certain areas of law?
  - What is the primary latent dimension in CJEU judgments? Is it a left/right dimension? Is it a pro-/anti-European integration dimension? Or is it something else, like a policy/procedure dimension?
- We'll use unsupervised and semi-supervised topic models to address the first question and an unsupervised scaling model to address the second question

Seeded LDA model



Wordfish (word positions)

Wordfish estimates for the positions of words in CJEU judgments (2019-2021)



Wordfish (document positions)