Topic Modeling

- Difference with what we've done so far: EDA: Exploratory Data Analysis Topic Modeling -- semantics Stylometry -- inconspicuous elements of writing
- Intro to modeling, McCarty
- The intuition behind topic modeling using the cook metaphor
- NMF explained
- Colloquium

https://www.scottbot.net/HIAL/index.html@p=19113.html https://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/ https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

Lego Train Metaphor

- Descriptive Model (Right): captures and represents the main features of a complex system. -- LEGO train: show how a real train's characteristics are captured in a model. Abstraction is made. -- We can use it to understand and explain the system, in this case a train.
- Prescriptive Model (Left): provides recommendations for action, optimizing a certain outcome. -- LEGO instructions are also a model: a step-by-step guide prescribes the process to achieve the final product. -- Here: there is role for the model in guiding actions to achieve desired results

! importance of recognizing which type of model to use depending on the objective.

Examples of models

models in humanities as tools for organizing information and understanding complex concepts

- **Grammars** (Descriptive Model):

  -- grammars are models of language that describe how words and sentences are formed. They help us understand the structure of language and are descriptive because they explain how language is used.

- **Periodization in Literature** (Descriptive Model):

  -- a model that organizes literary works into different time periods based on common characteristics. It is a descriptive model that helps categorize and understand literature in the context of historical and cultural periods.

- **Index** at the Back of Books (Prescriptive Model):

  -- The index is a prescriptive model because it guides the reader on where to find specific information within the text. It prescribes a way to navigate a book efficiently.

Minsky quote

Minsky, big name in AI, PhD from Princeton, founder of MIT's AI lab. According to Minsky, a model's usefulness is determined by its ability to provide answers to an observer's questions about the original object or concept.

- models **simplify complexity**, making it easier for researchers to study and understand phenomena
- models are **tools for prediction, explanation, and communication** in research.
- a model's utility is **not in being a perfect copy, but in being a useful simplification** that highlights important features
- importance of the observer's role in modeling – models must be tailored to the questions and needs of the researcher

What makes models useful (toy car metaphor)

What are some characteristics of models?

- Simple / Simplification (cf. Okham's razor: scope vs simplicity)
- Computationally tractable: importance of models being able to be computed or simulated
- Consistent: models should reliably produce the same results under the same conditions
- Manipulable: a model should be adjustable to test different scenarios or hypotheses

Can we distinguish between good and bad models?

- "All models are wrong but some are useful" (George Box) models can answer questions related to the attributes they simulate or represent models cannot answer questions outside their scope or design parameters

- Functional (>< essentialist) perspective; cf. Minsky

I prefer thinking from a functional (what it does) rather than an essentialist (what it is) perspective

How do we model text?

This brings me to the question How do we model text – especially if we want that model of text to be a good representation of the topics that are discussed in a large corpus of text such as the PPA?

One approach is the Bag of Words (BOW) method. This approach simplifies text by treating it as a collection of individual words, regardless of their order. It's a common technique in natural language processing and information retrieval. However, it often yields a sparse model with many zeros, due to the absence of most words in individual documents.

In thematic analyses, such as topic modelling, you want to give more attention words that seem to have a more specific meaning. For this reason, we will apply a very common scaling or weighting procedure to our corpus, namely TF-IDF (term frequency-inverse document frequency). This weighting scheme will modify the weight of words: if a a term occurs in fewer documents, it will receive a higher IDF-score.

Intuitively what this does:

- Common words across the corpus are penalized by assigning them lower weights.
- Rare words across the corpus are emphasized with higher weights, underlining their potential specificity to certain chunks (i.e. thgeir relevance to particular topics).

topic modeling algorithms can automatically identify and group words that frequently occur together in documents

- helps in discerning the underlying themes or 'topics' that run through a corpus without reading each document

Popular topic modeling techniques:

- Latent Dirichlet Allocation (LDA): Describes documents as mixtures of topics that spit out words with certain probabilities.
- Non-Negative Matrix Factorization (NMF): Factorizes the document-word matrix into two lower-dimensional matrices, revealing the latent topic structure. Example of Topics:

cfr. example: lists of words associated with two different topics, as derived from topic modeling, along with the probability scores next to each word, which indicate the likelihood of the word being used in a particular topic

There are various techniques to achieve topic modeling. Ultimately, they share two core characteristics:

First, they are unsupervised. This means they can operate with minimal human intervention, such as labeling or categorizing the data. However, the down side is their appetite for data – these methods thrive on large corpora. Secondly, at their core, these methods summarize a Bag of Words (BOW) into two distinct but interrelated tables. The 'Document-Topic table and a 'Topic-Word table

Cooking metaphor

reconstruct the jars (topics), te jar's contents (the individual words), and the cooking process (how many words from wich jars were taken to create a dish).

e.g. **paella vs. goulash** → they will have some spices/topics in common, but different distributions and amounts of these spices. and then some of the spices they will have not in common.

NMF

We will use an established method, called Non-Negative Matrix Factorization (NMF). NMF is an unsupervised machine-learning algorithm which is still used frequently in distributional semantics for building document-level topic representations.

Conceptually, what the model is trying to do, is come up with a list of topic scores for each document (e.g. 300), on the basis of the actual words which the document contains. These topic scores are then used to try and reconstruct which words were originally present in the document. This reconstruction is the objective of the topic modelling procedure. Of course, this reconstruction will never be perfect and, therefore, it is said to "violate" the objective to some extent.

As you could glean from the scores printed above, the model is trying to push its reconstruction error down, through formulating a better "topic summary" of the document in the 30 topic scores which it can use for each document. At some point, the model will either hit the maximum number of iterations which we specified (max_iter) or it will notice that it cannot reduce the error any further. This point in training is what we call convergence and the model typically stop training at this stage -- because it notices that it doesn't learn anything anymore.

Next, by examining both the topic words and the document-topic distributions, we can start to draw conclusions about the latent topical structures that organize the corpus, such as common terminology or discourses that span multiple texts.

Interpreting or evaluating the results of a topic model can be difficult. In a well-known paper by Chang et al. (2009) the interpretative phase of building a topic models has even been likened to reading tea leaves. A first and very useful step is to visualize which words are most significant for specific topics.

Another common and very intuitive approach to visualizing topics is through so-called word clouds. Here the relative size of the words in the topic clouds reflects how central they are to the topic.

Topic Modeling the PPA

Once we have identified topics within the corpus, we can examine how they have evolved over time. We can achieve this by plotting the prominence of each topic across different time periods, represented by the publication dates of the documents (or chunks) in the corpus. We calculate the mean topic weights for each year. The use of a rolling average also smooths out the line, preventing any single document or short-term change from overly influencing the trend appearance.

These are cherries that I have picked !