Cultural Analytics

ENGL 64.05

Fall 2019

Prof. James E. Dobson

October 21, 2019

Key Method: TF-IDF

- Term Frequency-Inverse Document Frequency is a widely used method to weight different documents for comparing term (word) frequencies.
- Used in topic modeling (many methods (although not LDA) convert raw token count into tf-idf weights before modeling) to compare documents of different length.
- Heavily used in indexing and search applications.
- First described by Karen Spärck Jones in 1972.

Karen Spärck Jones
(1935 – 2007)

Key Concept: Supervised / Unsupervised

- Supervised methods involve the labeling of data and/or the presence of operator feedback.
  - Classification with training data are supervised.
- Unsupervised methods (in theory) operate without the presence or hands of an operator.
  - Extraction of collocated terms and topic modeling would be considered two unsupervised methods.
- The notation suggests objective / subjective mapping but this is not exactly correct.

Topic Models

- David M. Blei's "Probabilistic Topic Models" (2012) introduces the concept to a wider audience.
- Documents are comparable text objects. Collections are groups of similar documents sharing some pre-existing attributes.
- Topic modeling, in his description, is associated with search and information retrieval: "rather than finding documents through keyword search alone, we might first find the theme that we are interested in, and then examine the documents related to that theme" (77).

David M. Blei

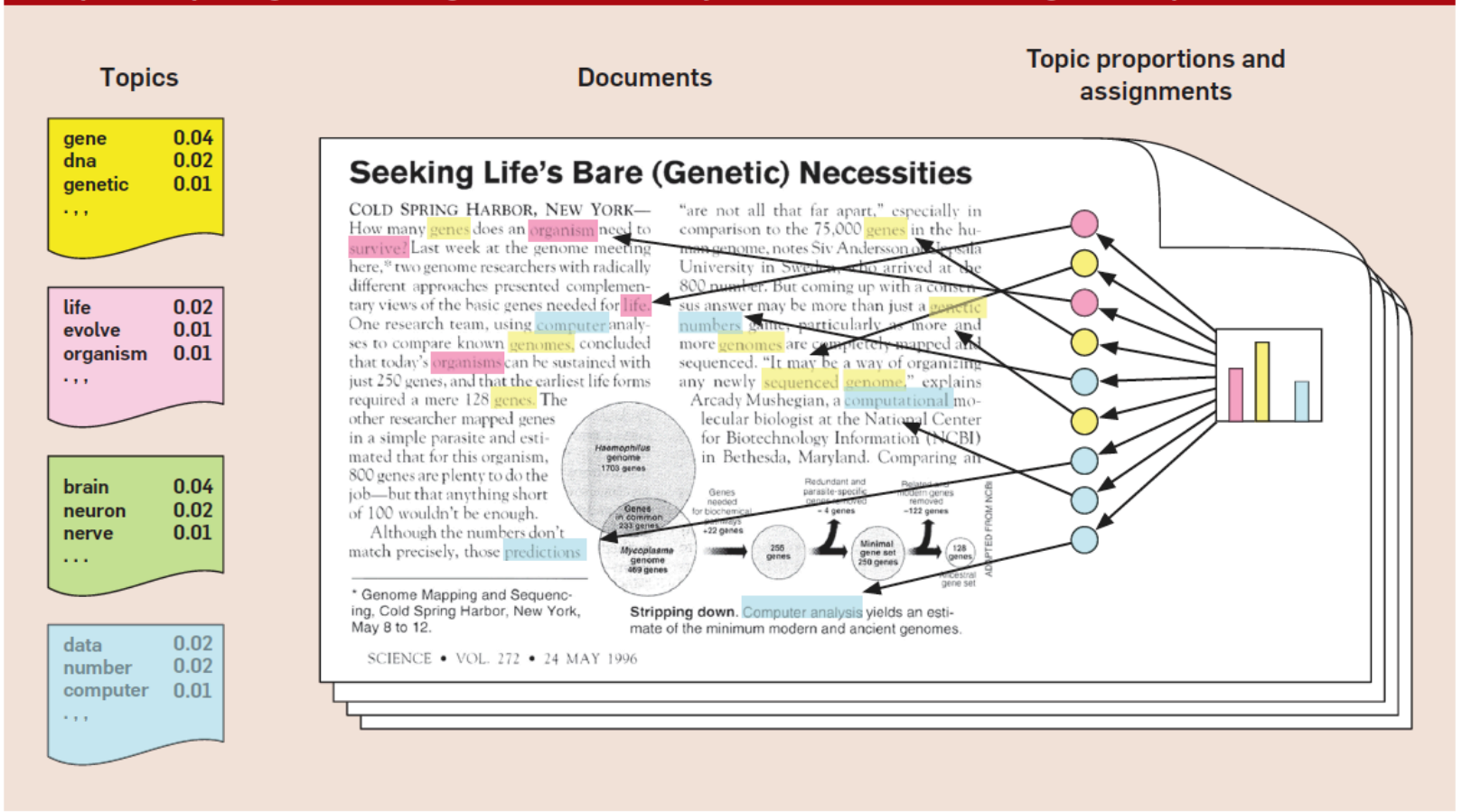
Definition: Topic

"We formally define a *topic* to be distribution over a fixed vocabulary. For example, the *genetics* topic has words about genetics with high probability and the evolutionary biology topic as words about *evolution biology* with high probability. We assume that these topics are specified before any data has been generated" (78)

LDA: latent Dirichlet allocation

- Major assumption: "all the documents in the collection share the same set of topics but each document exhibits those topics in different proportion" (79).
- LDA presents the "hidden structure" of the document by inferring the hidden topic structure from observed documents.
- The topic model is the set of frequent words that make up the frequent topics.
- It is, as Blei writes, retrospective: "the topics that emerge from the inference algorithm are interpretable for almost any collection that is analyzed. The fact that these look like topics has to do with the statistical structure of observed language and how it interacts with the specific probabilistic assumptions of LDA" (n. c, 79).
- These topics are used to automatically "annotate" each document in the collection.

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



Limitations of the Basic Topic Model

- Blei points out that there are some assumptions that present problems for certain tasks:
  - *bag of words* (BoW) model assumes that ordering does not matter. The presence and frequency of these words in these documents is all that matters. No context.
  - *unordered documents*: the basic model assumes a static set of documents sharing the same language model.
  - The number of topics is assumed to be known and *fixed* across the documents.

Other Limitations of Topic Models?

- What other assumptions and limitations might we find in this approach?
- How applicable are information science techniques to cultural objects?

Matthew Jockers

- This chapter of *Macroanalysis* uses topic models (LDA) to examine the presence of themes as linked to gender, nationality, and genre.
- His ongoing concern is with large-scale, data-driven methods rather than "the anecdotal type of analysis to which we are still accustomed, which is to say a close reading" (119).
- He argues that "unlike KWIC [key word in context] and collocate lists, which require careful human interrogation in order to parse out one word sense from another, topic modeling works in an unsupervised way, inferring information about individual word senses based on their repeated appearance in similar contextual situations" (124).

Theme, Topic, or Something Else?

"In this chapter, I use the terms theme, topic, and motif as synonyms for the same general concept: namely, a type of literary material, that is, "subject matter," that recurs with some degree of frequency throughout and across a corpus. This material functions as a central and unifying unit of a text or texts. Despite a long history of studying theme and motif in literature and even more extensively in folklore, these terms do remain ambiguous, terms of convenience. I believe that the word clusters discussed here are self-evidently thematic and that even while the matter of what constitutes a theme or motif is a broad area in which some things are black, some white, and some gray, most readers will recognize in the word distributions the larger thematic category to which these words belong. Although handbooks such as *Themes and Motifs in Western Literature* (Daemmrich and Daemmrich 1987) may help us to understand theme, even these scholarly compilations are open to the charge of being arbitrary—they define by example, not by concise definition. Daemmrich offers the theme of "Death," for example, and I am comfortable accepting "Death" as a theme. But Daemmrich also records a theme called "Eye." To my mind, "Eye" would be more appropriately chronicled in a dictionary of symbols than in a handbook of themes" (123).

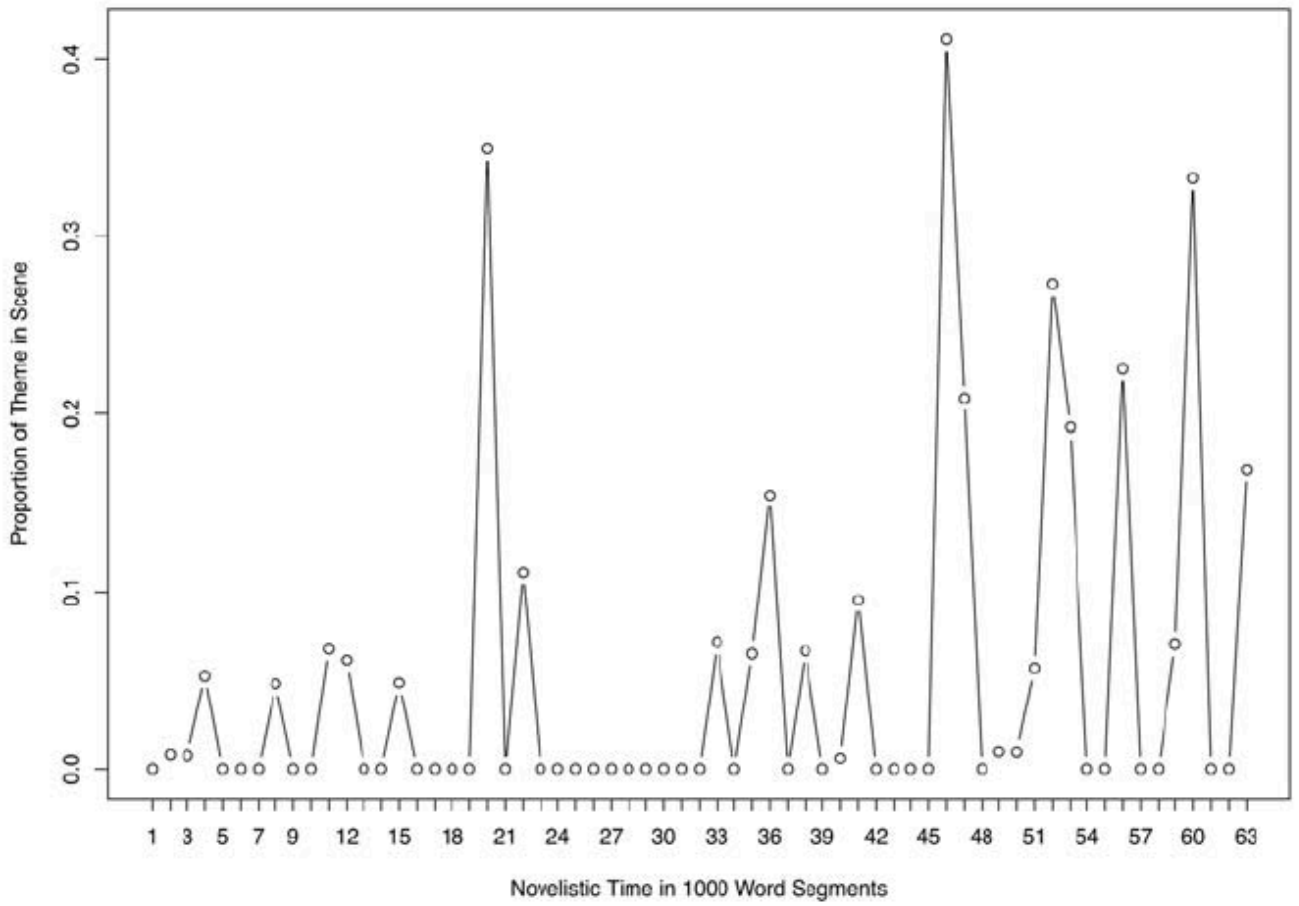
Preprocessing

- Because all documents in the same language will share some features (e.g., punctuation), we'll want to remove these.
- The same might apply to high frequency function words (stop words).
- Jockers uses Named Entity Recognition (NER) to remove all proper nouns (character names, etc).
- He only includes nouns through the use of a Part of Speech (POS) tagger.
- He then constructs his collection of documents from 1,000 word segments.

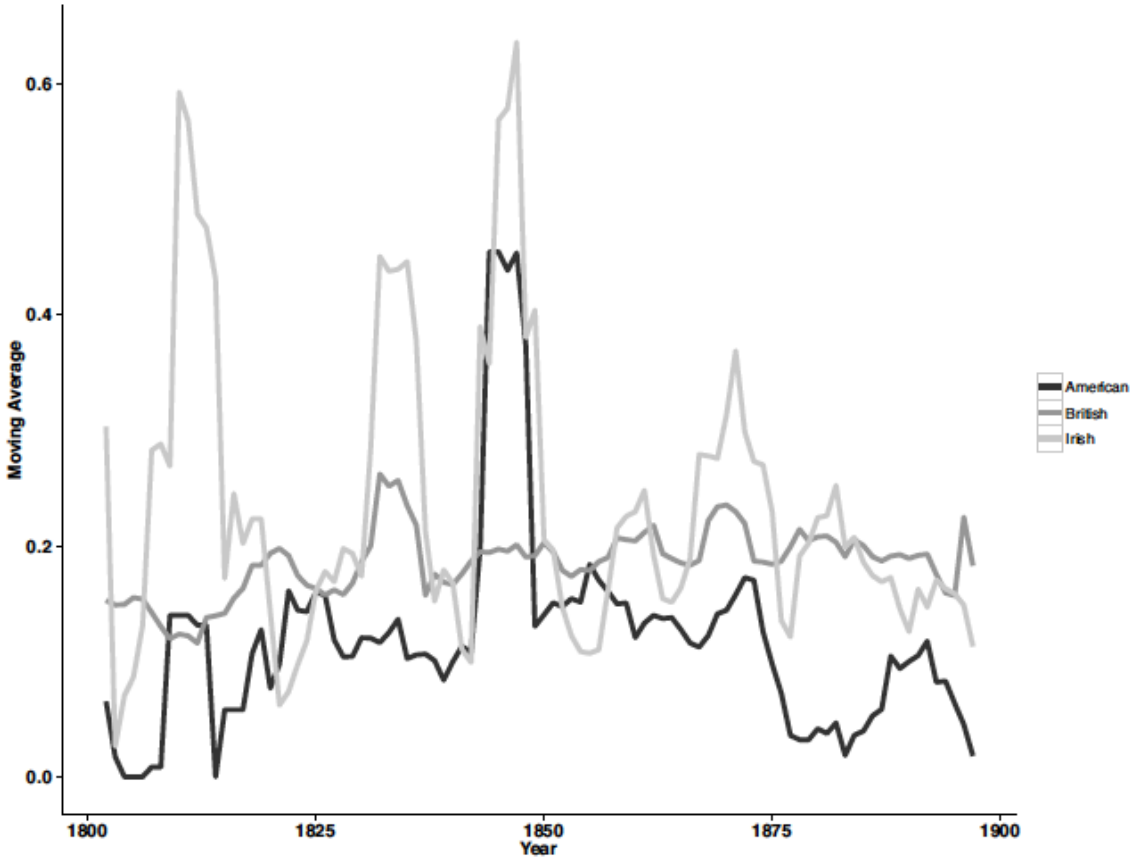
Cultural Analytics

Visualizing Topics

Visualizing Topics



Visualizing Topics



Interpretability

- How do we assess the degree to which a topic is interpretable?
- Literary scholars want interpretable data that can be socially anchored (to use Underwood's expression from our discussion).
- Computer and information scientists are not as interested in the anchoring of these topics: uninterpretable topics are interesting in their own right.