NLP For Economists

Topic Modeling

Sowmya Vajjala

Munich Graduate School of Economics - LMU Munich

Guest Course, October 2020

Goals

- Give a general overview of topic models and their applications

note: since I don't have much experience with topic models other than classroom instruction, I am giving you relevant resources where possible.
useful resources:

- `https://mimno.infosci.cornell.edu/`
- `http://www.cs.columbia.edu/~blei/topicmodeling.html`

What is topic modeling?

- Topic Models are a group of algorithms which attempt to discover latent themes in large collections of documents.
- They use statistical methods to analyze word usage in the texts to discover what "themes" run through them, how these themes connect to each other etc.

What is topic modeling?

- Topic Models are a group of algorithms which attempt to discover latent themes in large collections of documents.
- They use statistical methods to analyze word usage in the texts to discover what "themes" run through them, how these themes connect to each other etc.
- Good thing about them: they do not expect us to provide any prior annotations/categories for texts. Topics will "emerge" from the analysis.
- Bad thing: "Topics" need not necessarily be meaningful, unless you know how to tweak the models

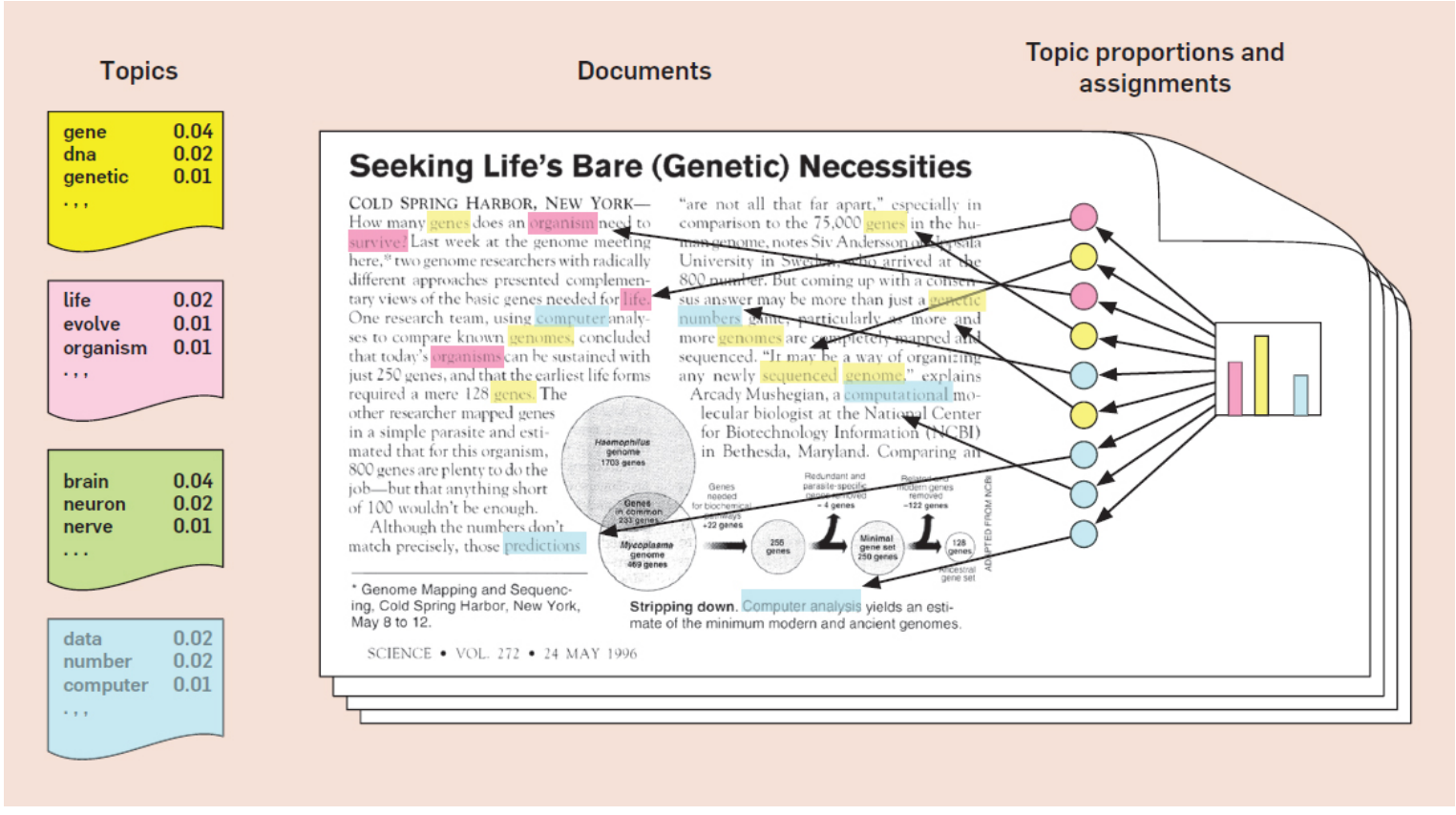
What is topic modeling?

- Topic Models are a group of algorithms which attempt to discover latent themes in large collections of documents.
- They use statistical methods to analyze word usage in the texts to discover what "themes" run through them, how these themes connect to each other etc.
- Good thing about them: they do not expect us to provide any prior annotations/categories for texts. Topics will "emerge" from the analysis.
- Bad thing: "Topics" need not necessarily be meaningful, unless you know how to tweak the models
- They seem to be the most popular method for analyzing unstructured text data in Economics

Latent Dirichlet Allocation (LDA)

- LDA is the most known topic modeling algorithm
- Intuitions:
  - each document is a mixture of multiple topics
  - each topic can be characterized by some set of keywords related to that topic.
  - a keyword can exist in multiple topics with different degrees of importance.

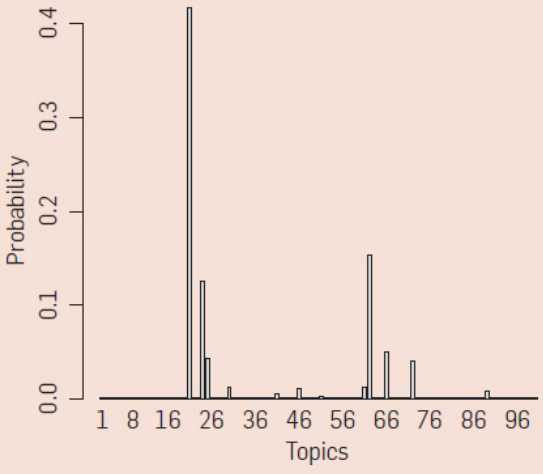
What does a Topic Model do?-1



source: `https://goo.gl/azc7Gc`

What does a Topic Model do? -2

Real inference with LDA - topic model built using 17000 articles from Science journal.

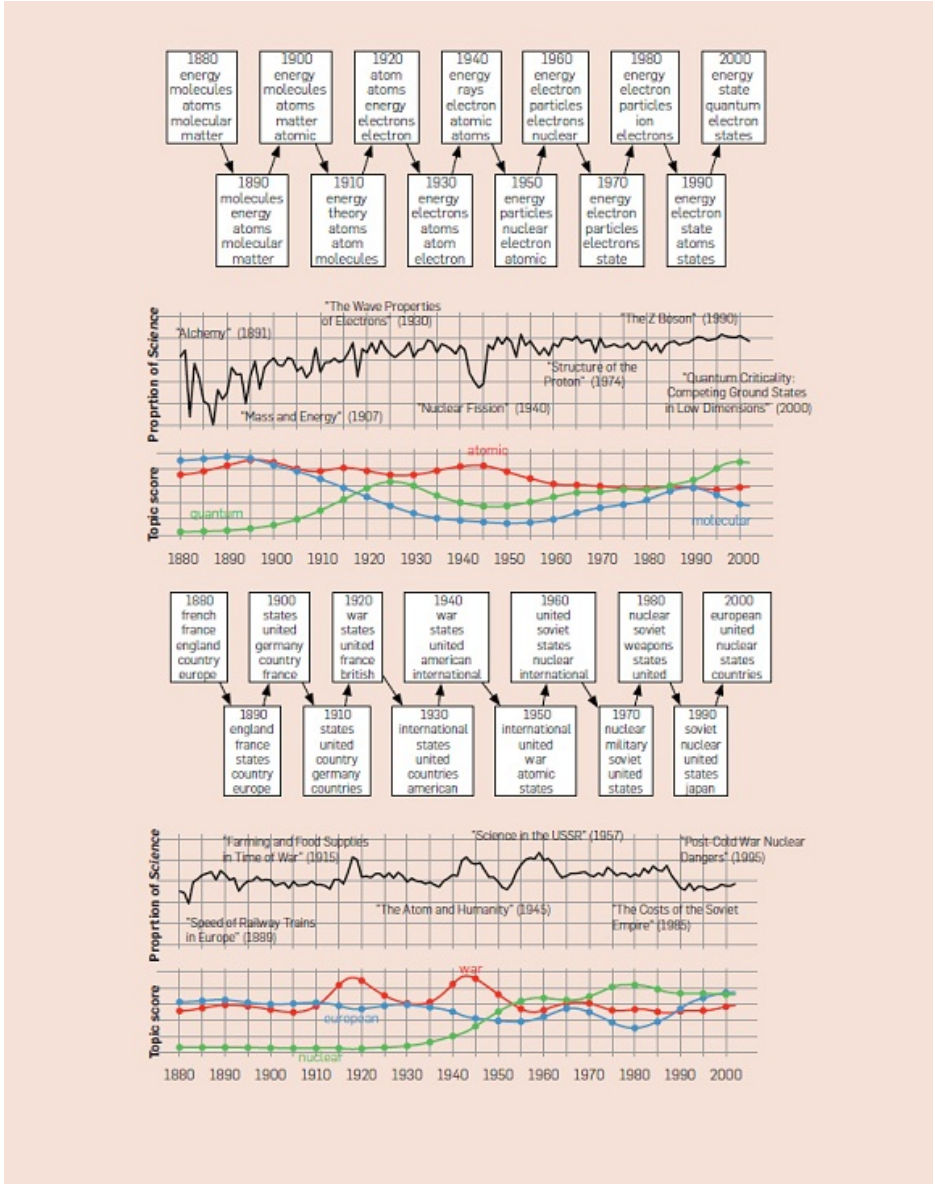


"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

source: `https://goo.gl/azc7Gc`

How are topic models useful? -2

Analysing topics over time



source: `https://goo.gl/azc7Gc`

How are topic models useful? -3

Analyzing topics by author

TOPIC 10	
WORD	**PROB.**
SPEECH	0.1134
RECOGNITION	0.0349
WORD	0.0295
SPEAKER	0.0227
ACOUSTIC	0.0205
RATE	0.0134
SPOKEN	0.0132
SOUND	0.0127
TRAINING	0.0104
MUSIC	0.0102
AUTHOR	**PROB.**
Waibel_A	0.0156
Gauvain_J	0.0133
Lamel_L	0.0128
Woodland_P	0.0124
Ney_H	0.0080
Hansen_J	0.0078
Renals_S	0.0072
Noth_E	0.0071
Boves_L	0.0070
Young_S	0.0069

TOPIC 209	
WORD	**PROB.**
PROBABILISTIC	0.0778
BAYESIAN	0.0671
PROBABILITY	0.0532
CARLO	0.0309
MONTE	0.0308
DISTRIBUTION	0.0257
INFERENCE	0.0253
PROBABILITIES	0.0253
CONDITIONAL	0.0229
PRIOR	0.0219
AUTHOR	**PROB.**
Friedman_N	0.0094
Heckerman_D	0.0067
Ghahramani_Z	0.0062
Koller_D	0.0062
Jordan_M	0.0059
Neal_R	0.0055
Raftery_A	0.0054
Lukasiewicz_T	0.0053
Halpern_J	0.0052
Muller_P	0.0048

TOPIC 87	
WORD	**PROB.**
USER	0.2541
INTERFACE	0.1080
USERS	0.0788
INTERFACES	0.0433
GRAPHICAL	0.0392
INTERACTIVE	0.0354
INTERACTION	0.0261
VISUAL	0.0203
DISPLAY	0.0128
MANIPULATION	0.0099
AUTHOR	**PROB.**
Shneiderman_B	0.0060
Rauterberg_M	0.0031
Lavana_H	0.0024
Pentland_A	0.0021
Myers_B	0.0021
Minas_M	0.0021
Burnett_M	0.0021
Winiwarter_W	0.0020
Chang_S	0.0019
Korvemaker_B	0.0019

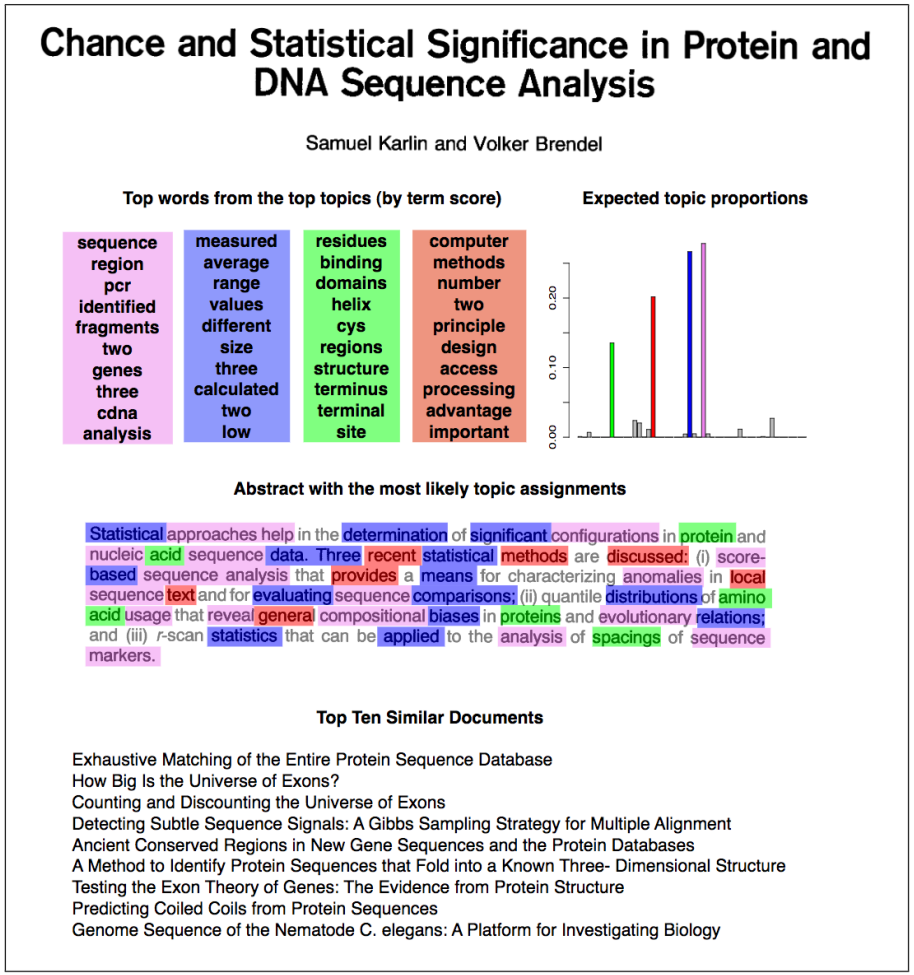
TOPIC 20	
WORD	**PROB.**
STARS	0.0164
OBSERVATIONS	0.0150
SOLAR	0.0150
MAGNETIC	0.0145
RAY	0.0144
EMISSION	0.0134
GALAXIES	0.0124
OBSERVED	0.0108
SUBJECT	0.0101
STAR	0.0087
AUTHOR	**PROB.**
Linsky_J	0.0143
Falcke_H	0.0131
Mursula_K	0.0089
Butler_R	0.0083
Bjorkman_K	0.0078
Knapp_G	0.0067
Kundu_M	0.0063
Christensen-J	0.0059
Cranmer_S	0.0055
Nagar_N	0.0050

Figure 3: An illustration of 4 topics from a 300-topic solution for the CiteSeer collection. Each topic is shown with the 10 words and authors that have the highest probability conditioned on that topic.

source:
`https://mimno.infosci.cornell.edu/info6150/readings/398.pdf`

How are topic models useful? -4

Picking up similar documents



Chance and Statistical Significance in Protein and DNA Sequence Analysis

Samuel Karlin and Volker Brendel

Top words from the top topics (by term score)

sequence	measured	residues	computer
region	average	binding	methods
pcr	range	domains	number
identified	values	helix	two
fragments	different	cys	principle
two	size	regions	design
genes	three	structure	access
three	calculated	terminus	processing
cdna	two	terminal	advantage
analysis	low	site	important

Expected topic proportions

Abstract with the most likely topic assignments

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Three recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence text and for evaluating sequence comparisons; (ii) quantile distributions of amino acid usage that reveal general compositional biases in proteins and evolutionary relations; and (iii) *r*-scan statistics that can be applied to the analysis of spacings of sequence markers.

Top Ten Similar Documents

Exhaustive Matching of the Entire Protein Sequence Database
How Big Is the Universe of Exons?
Counting and Discounting the Universe of Exons
Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment
Ancient Conserved Regions in New Gene Sequences and the Protein Databases
A Method to Identify Protein Sequences that Fold into a Known Three- Dimensional Structure
Testing the Exon Theory of Genes: The Evidence from Protein Structure
Predicting Coiled Coils from Protein Sequences
Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology

FIGURE 4. The analysis of a document from *Science*. Document similarity was computed using Eq. (4); topic words were computed using Eq. (3).

source:
`http://www.cs.columbia.edu/~blei/papers/BleiLafferty2009.pdf`

A small exercise

Interpreting Topic Models

What do you think of these topics (and their 5 most frequent keywords)? If you are asked to evaluate this topic model now, what will you look for?

- Topic 1 : Onion, Cream, Black pepper, Milk, Cinnamon
- Topic 2: Cumin, Coriander, Turmeric, Fenugreek, Lemongrass
- Topic 3: Vanilla, Cream, Almond, Coconut, Oat
- Topic 4: Olive oil, tomato, parmesan cheese, lemon juice, garlic
- Topic 5: soy sauce, scallion, sesame oil, cane molasses, roasted sesame seed
- Topic 6: Milk, pepper, yeast, potato, lemon juice
- Topic 7: Scallion, garlic, ginger, soy bean, pepper
- Topic 8: Pepper, vinegar, onion, tomato, milk

source: `https://gist.github.com/inkhorn/9044779/#file-recipe-analysis-r`

Some questions to ponder on:

- Coherence among the keywords for a topic (Is some word looking out of place?)
- Are there two topics that perhaps should be one?
- Can we name the topics with what we think is the group?
- Do you think the topic model learnt something about ingredients in this example?

Building Topic models

- gensim is a popular library for topic models in python. `https://radimrehurek.com/gensim/`
- sklearn also has an implementation of LDA
- some links if you want to explore in Python:
  - `https://medium.com/analytics-vidhya/topic-modeling-using-gensim-lda-in-python-48eaa2344920`
  - `https://github.com/practical-nlp/practical-nlp/blob/master/Ch7/02_TopicModelling.ipynb`
- I was about to write a demo code, and discovered this in browser topic model builder!: `https://mimno.infosci.cornell.edu/jsLDA/`
- Try it out (if you want, you can submit your analysis of this tool instead of the actual Assignment 3!)

Concluding Remarks

- Topic models are a good tool to use when we have a large corpus of texts, and no other related annotation.
- Although they are complex mathematical models, they are relatively easier to implement for a not so experienced person.
- However, training/tuning model performance can take time, it can be hard to understand what is a good number of topics, get topics with coherent keywords, etc.