LING 410X: Language as Data

Semester: Spring '18

Instructor: Sowmya Vajjala

Iowa State University, USA

20 March 2018

Class Outline

- Assignment 4 - brief discussion
- Introduction to topic modeling
- Assignment 5 description
- This week and next week: discussion about topic modeling, and how to make our own models

Announcement

- Thursday: No class, as I am at a conference and our talk is scheduled at the same time. But, there is an optional exercise, based on Chapter 13 in the textbook.

Rest of the semester

Macro Analysis of Texts:

1. Classification (Chapter 12) - This is what we discussed before the break
2. Topic modeling (Chapter 13) - this week and next
3. Clustering (Chapter 11) - When talking about visualization, after topic modeling

Assignment 4 Grading

- First question: frequent words in the training and testing csv files - straight forward steps (everyone did this part well)

Assignment 4 Grading

- First question: frequent words in the training and testing csv files - straight forward steps (everyone did this part well)
- Second question:
  - More like a exploratory assignment. The tutorial works, I wanted you to try changing somethings in that procedure and see what happens.
  - My take on your submissions: I liked the ones which documented their experiences -e.g., what worked, what failed? did they manage to get better results? etc.
  - The errors some of you saw when you tried to change the SVM algorithm to something else: Are they only in Windows? (some of you did not see this). One other possibility: memory limits.
  - I did not deduct any points for these errors - they are a part of the assignment!
- Where to go from there: Look for some publicly available classification dataset (there are many, I gave some links last time) and try to do classification.

Introduction to Topic Modeling

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: Lot of math behind it (but we do not have to understand the math to apply topic models)

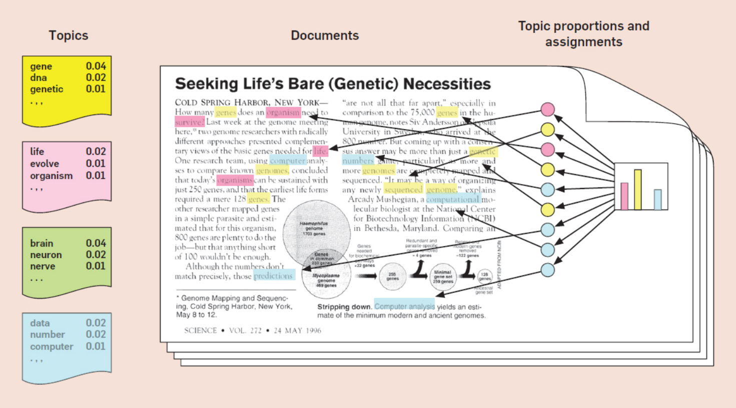
What is topic modeling?

- 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: Lot of math behind it (but we do not have to understand the math to apply topic models)
- One of the most popular methods of analyzing unstructured text data.

Latent Dirichlet Allocation (LDA)

- LDA is the simplest 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



Topics
Documents
Topic proportions and assignments
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
ADAPTED FROM NCBI

gene	0.04
dna	0.02
genetic	0.01
. . ,	

life	0.02
evolve	0.01
organism	0.01
. . ,	

brain	0.04
neuron	0.02
nerve	0.01
. . .	

data	0.02
number	0.02
computer	0.01
. . ,	

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

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

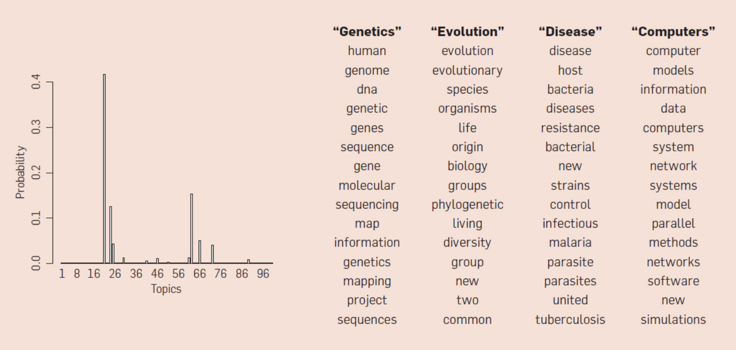
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

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

What does a Topic Model do? -2

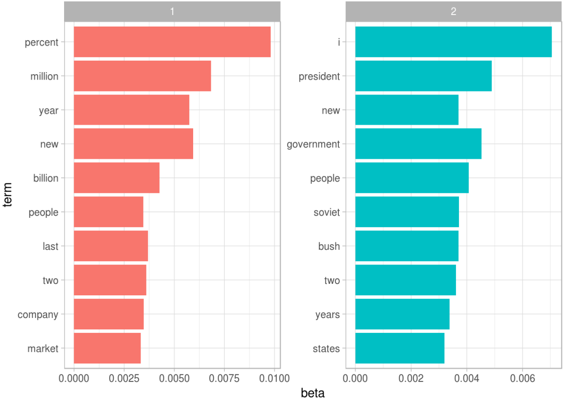
Real inference with LDA - topic model built using 17000 articles from Science journal.



“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`

What does a Topic Model do? -3



source: `http://tidytextmining.com/topicmodeling.html`

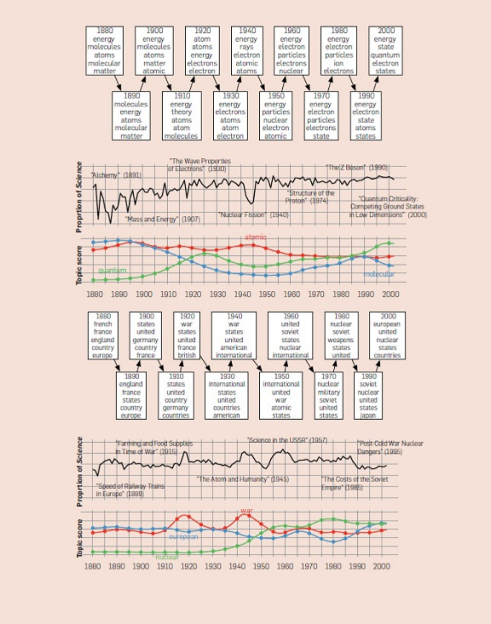
How are topic models useful? -1

The bestseller code

- One application comes from the textbook author.
- He and his team analyzed best selling novels using topic models and concluded that best seller novels focus on a small number of topics instead of discussing 1000 things in one story :)
- Okay, it is more detailed than that, I am just telling a micro-summary of what they concluded.

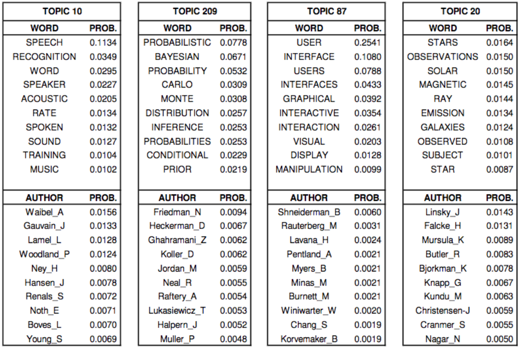
How are topic models useful? -2

Analysing topics over time



How are topic models useful? -3

Analyzing topics by author



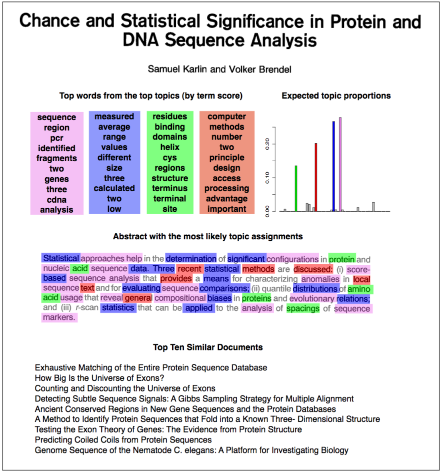
TOPIC 10		TOPIC 209		TOPIC 87		TOPIC 20	
WORD	PROB.	WORD	PROB.	WORD	PROB.	WORD	PROB.
SPEECH	0.1134	PROBABILISTIC	0.0778	USER	0.2541	STARS	0.0164
RECOGNITION	0.0349	BAYESIAN	0.0671	INTERFACE	0.1080	OBSERVATIONS	0.0150
WORD	0.0295	PROBABILITY	0.0532	USERS	0.0788	SOLAR	0.0150
SPEAKER	0.0227	CARLO	0.0309	INTERFACES	0.0433	MAGNETIC	0.0145
ACOUSTIC	0.0205	MONTE	0.0308	GRAPHICAL	0.0392	RAY	0.0144
RATE	0.0134	DISTRIBUTION	0.0257	INTERACTIVE	0.0354	EMISSION	0.0134
SPOKEN	0.0132	INFERENCE	0.0253	INTERACTION	0.0261	GALAXIES	0.0124
SOUND	0.0127	PROBABILITIES	0.0253	VISUAL	0.0203	OBSERVED	0.0108
TRAINING	0.0104	CONDITIONAL	0.0229	DISPLAY	0.0128	SUBJECT	0.0101
MUSIC	0.0102	PRIOR	0.0219	MANIPULATION	0.0099	STAR	0.0087
AUTHOR	PROB.	AUTHOR	PROB.	AUTHOR	PROB.	AUTHOR	PROB.
Waibel_A	0.0156	Friedman_N	0.0094	Shneiderman_B	0.0060	Linsky_J	0.0143
Gauvain_J	0.0133	Heckerman_D	0.0067	Rauterberg_M	0.0031	Falcke_H	0.0131
Lamel_L	0.0128	Ghahramani_Z	0.0062	Lavana_H	0.0024	Mursula_K	0.0089
Woodland_P	0.0124	Koller_D	0.0062	Pentland_A	0.0021	Butler_R	0.0083
Ney_H	0.0080	Jordan_M	0.0059	Myers_B	0.0021	Bjorkman_K	0.0078
Hansen_J	0.0078	Neal_R	0.0055	Minas_M	0.0021	Knapp_G	0.0067
Renals_S	0.0072	Raftery_A	0.0054	Burnett_M	0.0021	Kundu_M	0.0063
Noth_E	0.0071	Lukasiewicz_T	0.0053	Winiwarter_W	0.0020	Christensen-J	0.0059
Boves_L	0.0070	Halpern_J	0.0052	Chang_S	0.0019	Cranmer_S	0.0055
Young_S	0.0069	Muller_P	0.0048	Korvemaker_B	0.0019	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 region pcr identified fragments two genes three cdna analysis	measured average range values different size three calculated two low	residues binding domains helix cys regions structure terminus terminal site	computer methods number two principle design access processing advantage 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

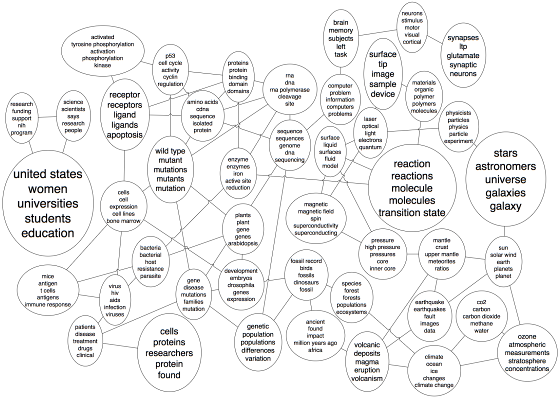
Exhaustive Matching of the Entire Protein Sequence Database
How Big Is the Universe of Exons?
Counting and Discounting the Universe of Exons
Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment
Ancient Conserved Regions in New Gene Sequences and the Protein Databases
A 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`

How are topic models useful? -5

Topic Graphs


activated tyrosine phosphorylation activation phosphorylation kinase
p53 cell cycle activity cyclin regulation
proteins protein binding domain domains
rna dna rna polymerase cleavage site
brain memory subjects left task
neurons stimulus motor visual cortical
synapses ltp glutamate synaptic neurons
surface tip image sample device
computer problem information computers problems
materials organic polymer polymers molecules
research funding support nih program
science scientists says research people
receptor receptors ligand ligands apoptosis
amino acids cdna sequence isolated protein
sequence sequences genome dna sequencing
laser optical light electrons quantum
physicists particles physics particle experiment
surface liquid surfaces fluid model
wild type mutant mutations mutants mutation
enzyme enzymes iron active site reduction
reaction reactions molecule molecules transition state
stars astronomers universe galaxies galaxy
united states women universities students education
cells cell expression cell lines bone marrow
plants plant gene genes arabidopsis
magnetic magnetic field spin superconductivity superconducting
bacteria bacterial host resistance parasite
pressure high pressure pressures core inner core
mantle crust upper mantle meteorites ratios
sun solar wind earth planets planet
mice antigen t cells antigens immune response
virus hiv aids infection viruses
gene disease mutations families mutation
development embryos drosophila genes expression
fossil record birds fossils dinosaurs fossil
species forest forests populations ecosystems
earthquake earthquakes fault images data
co2 carbon carbon dioxide methane water
patients disease treatment drugs clinical
cells proteins researchers protein found
genetic population populations differences variation
ancient found impact million years ago africa
volcanic deposits magma eruption volcanism
climate ocean ice changes climate change
ozone atmospheric measurements stratosphere concentrations

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

Fun with topic models

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

A small exercise

- Think for 5-10 minutes and try to list some (up to 5) applications of topic modeling in your discipline/topics of your interest
- Now, discuss with our neighbor and compare your and their ideas
- After about 15 minutes, let us discuss what you think are potential applications of topic models.

Another 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? Think and discuss with your classmates for 5 minutes and we will later collect all answers.

- 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

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?

Topic Models in R

- different libraries: mallet, topicmodels, LDA etc.
- Textbook follows mallet
- Your Assignment 5 will use tm and topicmodels.

Assignment 5 description

- Deadline: 31st March
- grade: 15%
- Num. questions: 1
- What to do?: Build a topic model with the given data, following given instructions, and answer questions about what you did.
- Difficulty level: moderate, but the program takes a few minutes to complete running.
- R libraries needed: tm, topicmodels

Topic Models - The Textbook Way

- The author used mallet library to develop topic models for a corpus of novels and authors (same one he used in Chapters 11-12).
- I will follow a different method (using tm, which we used before), but I recommend you to also go through this.
- Note: You won't be an expert in text mining with one undergrad course. It is okay if you don't have a 100% understanding of this.
- The goal is to introduce you to different possible ways, give some ideas, and make you think.

For Thursday

- I uploaded a Zip file, start with that. It contains all you need to follow Chapter 13's example.
- Attendance question: Try to follow the textbook example (materials provided in a zip file), and write your notes in the forum for 22nd March.

Next Week

- Read this before coming to class: `https://goo.gl/L8MFfG`
- Or this: `http://www.scottbot.net/HIAL/index.html@p=19113.html`
- Optional, additional reading (for next week): `https://goo.gl/azc7Gc` (Has some math)
- Other references: `http://tidytextmining.com/topicmodeling.html`
- Attendance question for today: What are four things you need to build a topic model? - Answer can be found by reading first url.