Topic Models

Lecture 1
Data Analysis
E0 259

Data Usage Today

- 2.5 quintillion bytes generated everyday
- Average knowledge worked inundated with several Gb of data per day
- Cognitive capacity: 2-60 bps for attention, decision-making, perception, motion, and language
- 10^6 bps for sensory processing

Topic Modeling

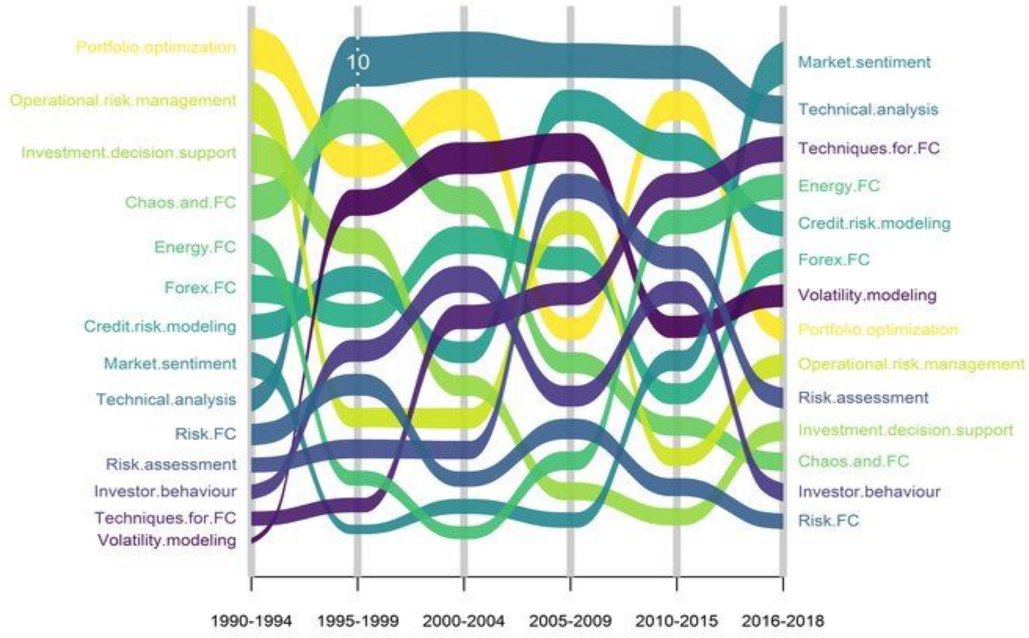
- A way to auto organize data based on topics/themes in document
- Helps with summarizing, search and auto categorization etc

Applications - Discover Themes

- 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

Model Topic Evolution



Machine Learning in Finance: A Topic Modeling Approach - Aziz et. al 2019

Organize and Search Large Document Collection

Topic list coverage checklist [edit]

Each of the following links needs to be checked. Where there's an article rather than a redirect, a merge into the corresponding outline should be considered. Each outline includes a link to the corresponding index.

Art and culture [edit]

Culture • Classical studies • Cooking • Critical theory • Hobbies • Literature

Art and **Entertainment** • Fiction • Game • Poetry • Sports

Performing arts • Dance • Film • Music • Opera • Theatre

Visual arts • Architecture • Crafts • Drawing • Film • Painting • Photography • Sculpture •Typography

Geography and places [edit]

Geography

Health and fitness [edit]

Health • Exercise • Health science • Nutrition

History and events [edit]

History • Classical antiquity • Medieval history (Middle Ages) • Renaissance

Mathematics and abstractions [edit]

Mathematics • Arithmetic • Algebra • Calculus • Discrete mathematics • Geometry • Trigonometry • Logic • Statistics

Natural sciences and nature [edit]

Main list: List of science topics

Biology • Animals • Biochemistry • Botany • Ecology • Zoology

Physical science • Astronomy • Chemistry • Earth sciences • Physics Fractions

People and self [edit]

People and **Self** • Biology • Psychology • Relationships

Philosophy and thinking [edit]

Philosophy • Philosophical theories • Humanism • Logic • Thinking • Transhumanism

Religion and spirituality [edit]

Religion

Social sciences and society [edit]

Main list: List of science topics

Social sciences • Archaeology • Critical theory • Economics • Geography • History • Linguistics • Law • Political science • Psychology • Sociology • Relationships

Topics from US Speeches

Topic # 0: government people central nicaragua mondale military america freedom el country men security salvador today states president force economic political united

Topic # 1: coal war miners men day miner mine board country great world sons peace workers mining mines fellow responsibilities died end

Topic # 2: people children today years day work president world time good american church live mondale faith freedom great make place life

Topic # 3: states government united congress great country public made people state citizens year present time power war part foreign law treaty

Topic # 4: nuclear world men people power soviet great america test american religious continue peace government work states good communist danger president

Topic # 5: chinese legations peking imperial china legation yamen foreigners government boxers antiforeign tsungli blows demanded provinces movement primaries exhibits 92nd

Topic # 6: world peace people nations united war nation states american america freedom years great time government today free country security soviet

Topic # 7: president space national iran ive treaty united people im years states nafta policy great world security weve staff administration board

Topic # 8: today great people men world life man country nation time years day university america society government free educated honor americans

Topic # 9: government american slavery states federal united world attack defense nazi constitution german question war control ships people affirmative congress time

Topic # 10: statute law purpose men union combinations capital states army company companies made business great people combination united tobacco antitrust corporations

Topic # 11: president people states question time year congress bill state made united slavery decision house constitution problem prices today point make

Topic # 12: people congress business great government national world years american men make nation country law work labor year federal time economic

Topic # 13: beloved rescue people cherokees good men iran give nation operation made states united agent lands indian man advice nations great

Topic # 14: president vietnam people made south time country united government states congress american general secretary make good military action hope war

Topic # 15: people watergate government national made political present facts year war matter time great american make house case president greece america

Topic # 16: iraq america people nation men terrorists american iraqi freedom free life great women forces government day terror democracy country world

Topic # 17: tax day president great relief americans american today john remember kennedy thanksgiving months years house god time conservatives good fellow

Topic # 18: war forces american enemy men fighting people south japanese vietnam north united americans world end vietnamese victory troops peace great

Topic # 19: president senator question kennedy states years people united nixon america man uh administration republican party good country time made im

Topic # 20: people government america years work tax american year congress americans make time health care president children jobs budget federal economy

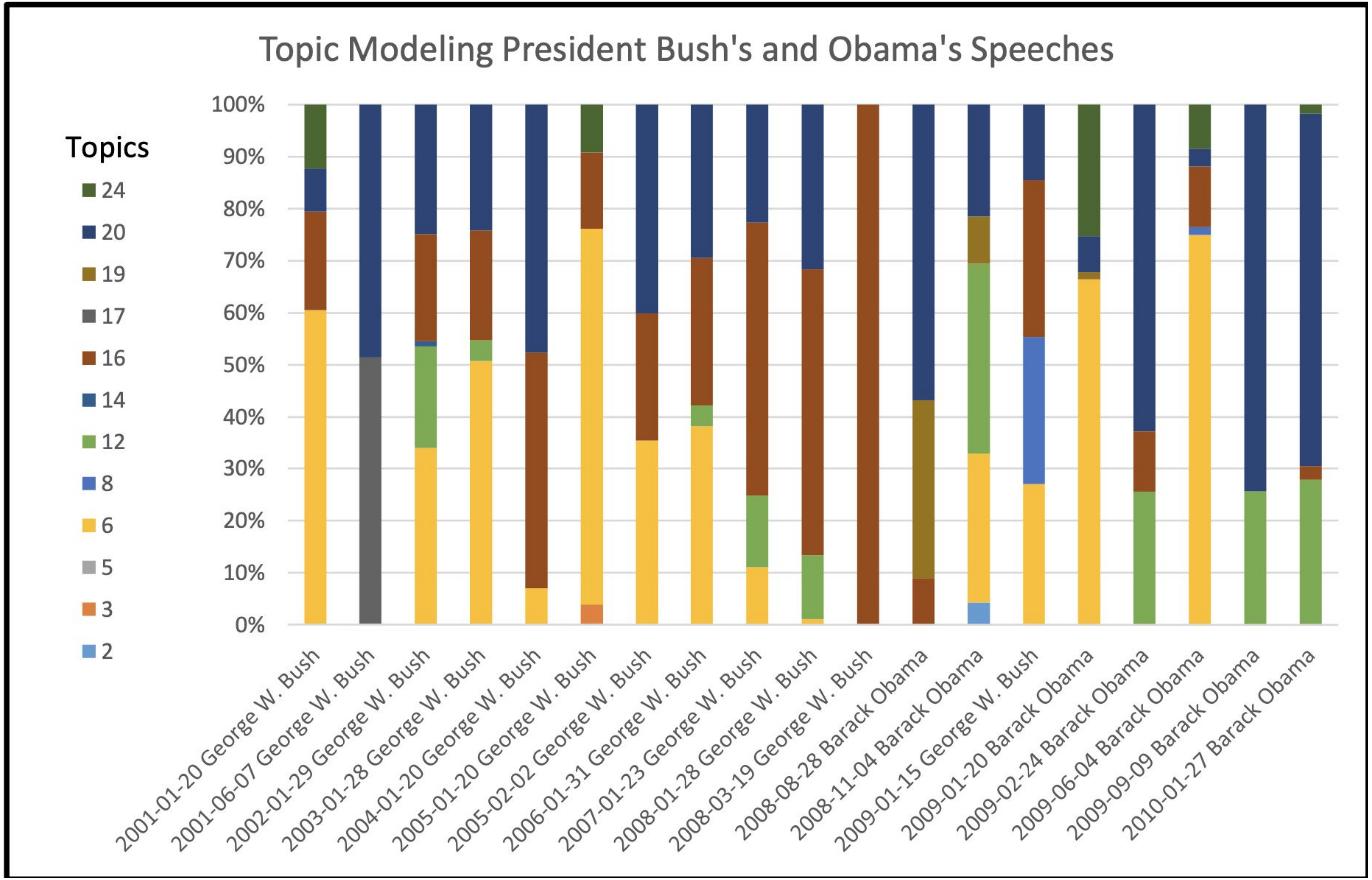
Topic # 21: world united states freedom people peace country nations great policy years time american countries america today power men free president

Topic # 22: soviet nuclear union arms missiles weapons soviets treaty world gorbachev berlin freedom people secretary europe united strategic states peace president

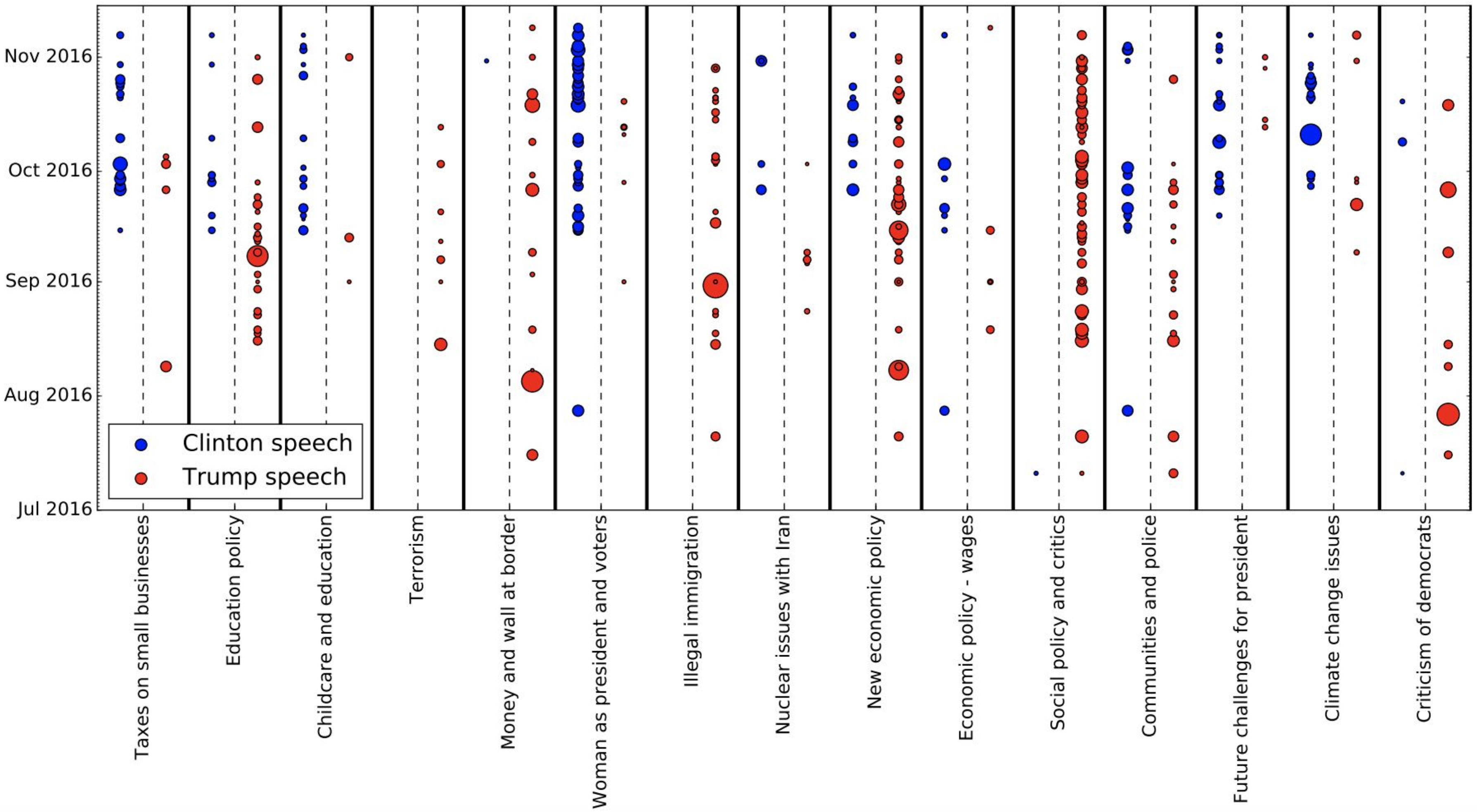
Topic # 23: people rights constitution party government union time states great law national polish years state victims democratic american republican platform country

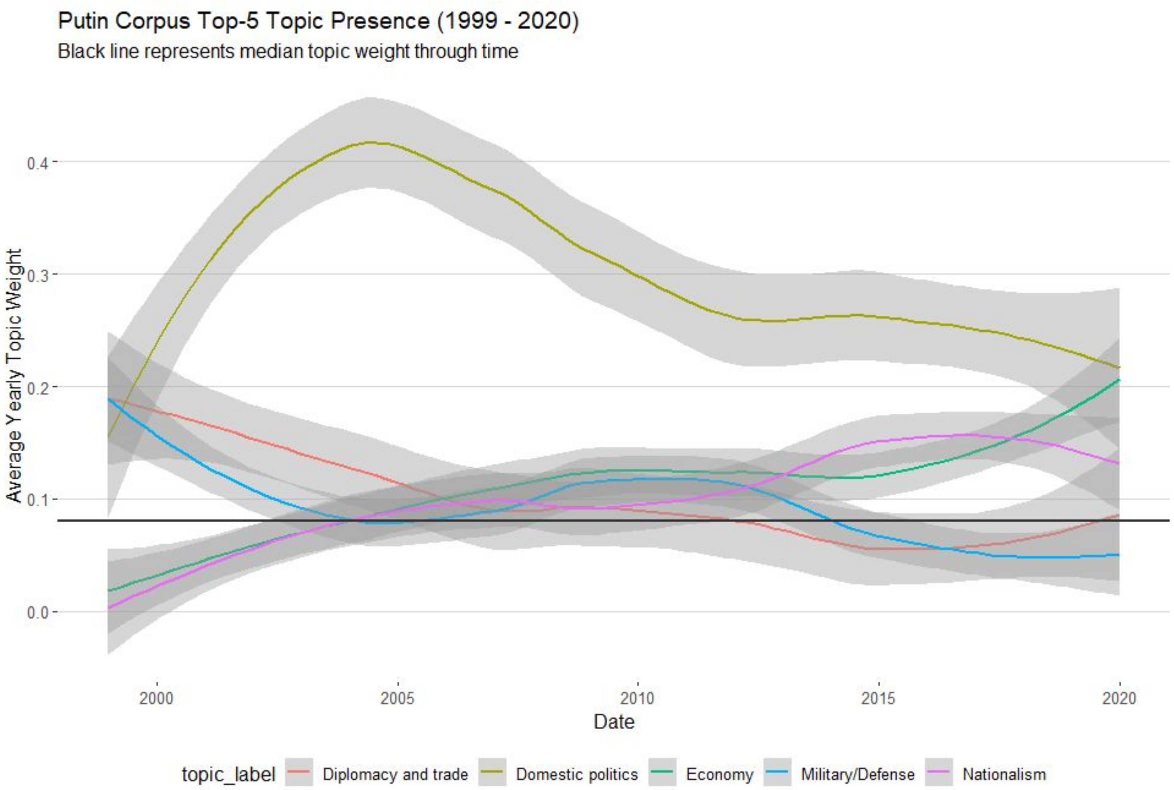
Topic # 24: energy oil congress people years future american year world government time federal program percent great make foreign states united america

Villadsen, Ole (2016): Analyzing Presidential Speeches with Topic Modeling. figshare. Dataset.
https://doi.org/10.6084/m9.figshare.2060724.v1



Villadsen, Ole (2016): Analyzing Presidential Speeches with Topic Modeling. figshare. Dataset.
https://doi.org/10.6084/m9.figshare.2060724.v1


Nov 2016
Oct 2016
Sep 2016
Aug 2016
Jul 2016
Clinton speech
Trump speech
Taxes on small businesses
Education policy
Childcare and education
Terrorism
Money and wall at border
Woman as president and voters
Illegal immigration
Nuclear issues with Iran
New economic policy
Economic policy - wages
Social policy and critics
Communities and police
Future challenges for president
Climate change issues
Criticism of democrats

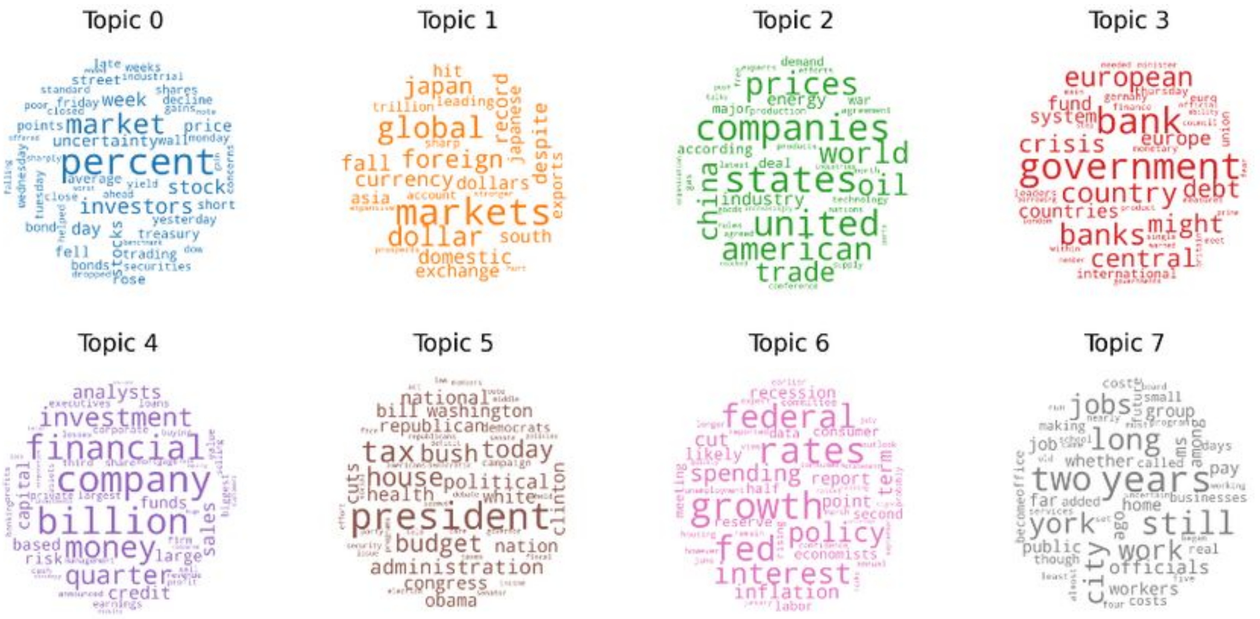


https://medium.com/the-die-is-forecast/topic-modeling-as-osint-exploring-russian-presidential-speech-topics-over-time-ad6018286d37

Topic Modeling vs Document Classification

- Topic Modeling is unsupervised Learning.
- Apriori no labeled data on which documents belong to which topics
- Apriori no information on what topics are even present in documents
- Each document could belong to a mixture of topics
- Document classification is supervised learning

Topics are Word Distributions



Word embeddings for topic modeling: an application to the estimation of the economic policy uncertainty index - Belmonte et. al 2021

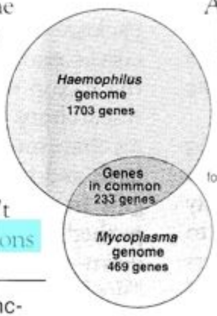
Each Document is a Mixture of Topics

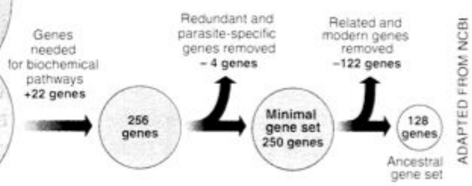
- Each document is a mixture of topics
- Each colour is a different topic

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.

Latent Semantic Analysis

- Formulate a term document matrix
- Each entry could be term frequency, tf-idf score etc.
- How do we solve?
- SVD

$$
\begin{array}{cc}
 & \begin{array}{ccccc} & & \mathbf{d}_j & & \\ & & \downarrow & & \end{array} \\
\mathbf{t}_i^T \rightarrow & \begin{bmatrix}
x_{1,1} & \dots & x_{1,j} & \dots & x_{1,n} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
x_{i,1} & \dots & x_{i,j} & \dots & x_{i,n} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
x_{m,1} & \dots & x_{m,j} & \dots & x_{m,n}
\end{bmatrix}
\end{array}
$$

Term Document Matrix - src: wikipedia

Limitations of LSA

- Storage and compute overhead
- Doesn’t capture polysemy - multiple meanings of a word
    - He was booked into the hotel vs he was booked by the referee
    - I was walking along the river bank vs I withdrew money from the bank
- Hard to interpret results

Probabilistic Models

- Assume that some well defined probabilistic model with certain parameters, generates each document
- We know the model, but not the parameters.
- All we can observe are the documents and the words.
- How do we figure out the parameters?

Parameter Estimation

- Maximum Likelihood Estimation
- Maximum A Posterior
- Expectation Maximization

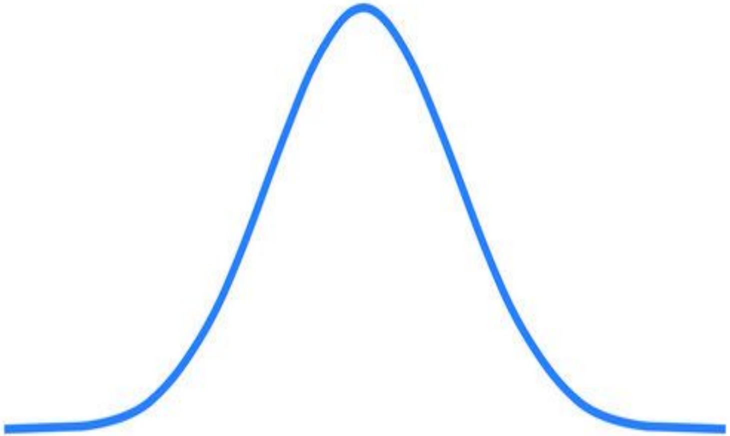
Maximum Likelihood Estimation

- Assume you know that some samples you are observing is from a Gaussian distribution.
- How do you estimate the mean μ and variance σ^2

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \mu)^2$$

Why is this correct?



MLE for Gaussian Distribution

- $p(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$
- $\mu, \sigma = \arg\max_{\mu,\sigma} p(x|\mu, \sigma)$
- $\mu, \sigma = \arg\max_{\mu,\sigma} \prod_{i=1}^{N} p(x_i|\mu, \sigma)$

Assuming i.i.d random variables

- $\mu, \sigma = \arg\min_{\mu,\sigma} \sum_{i=1}^{N} \ln(\sigma) + \frac{(x_i-\mu)^2}{2\sigma^2}$

Taking negative log

Differentiate w.r.t μ and σ and we see that intuition is exactly MLE!!!

Maximum A Posteriori

- India scored 250 in T20
- Candidate explanations
  - Kohli scored a century
  - Bumrah scored a century
  - Pakistan gave 200 in extras
- P(India scored 250 | explanations)
  - P(India scored 250 | Kohli scored a century) = 0.8
  - P(India scored 250 | Bumrah scored a century) = 0.0001
  - P(India scored 250 | Pakistan gave 200 in extras) = 0.00000001
- This is based on some “prior” assumptions

Maximum A Posteriori

$$p(\theta|X) \propto p(X|\theta)p(\theta)$$

prior

$$-\ln(p(\theta|X)) = -\ln(p(X|\theta)) - \ln p(\theta) + c$$

$$\arg\min_\theta -\ln(p(X|\theta)) - \ln p(\theta)$$

MLE with penalty for prior

Expectation Maximization

- Assume Gaussian mixture model
- Data is coming from one of the Gaussian distributions
- You know how many Gaussian distributions there are in mixture model
- You don't know their parameters
- How do you estimate parameters from data?

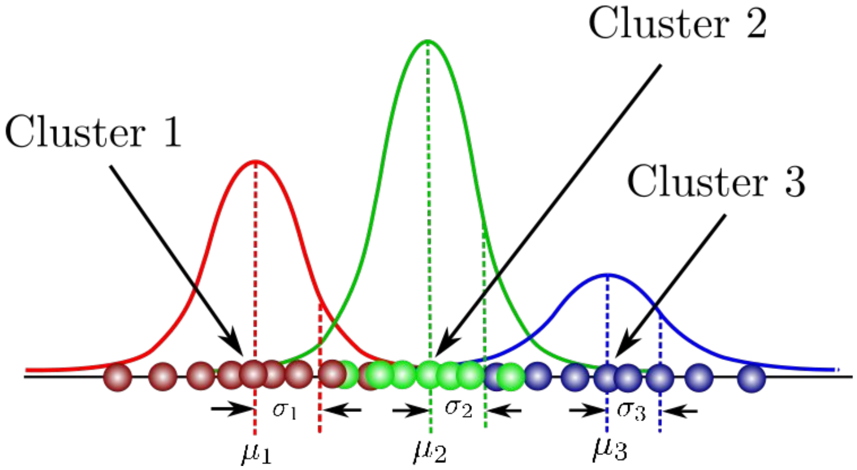
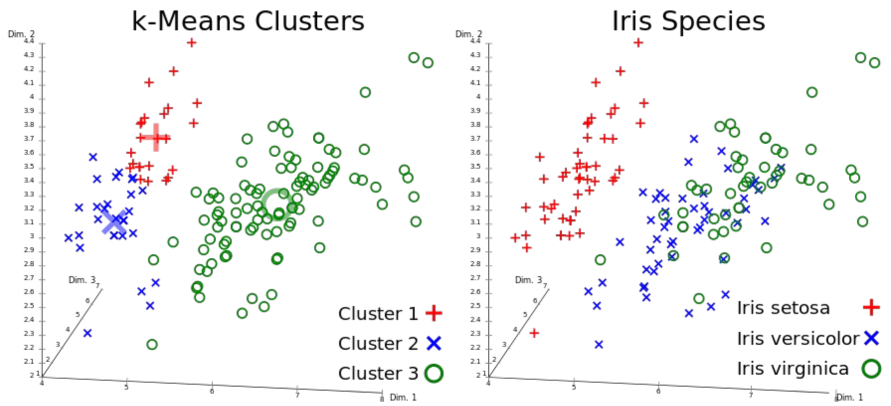


Image source:
https://medium.com/@yara.ahmed.amin/gaussian-mixture-model-4c71342b67d3

Recall k-means clustering

- Assume there are k clusters
- Pick k means at random
- Assign each point to the nearest cluster point
- Recompute the k-means, iterate
- EM algorithm - soft means instead of hard means.



Expectation Maximization Algorithm

- Assume there are *k* clusters
- Each cluster is from some unknown Gaussian model
- Start by assuming some random parameters (μ_i, σ^2_i) for each cluster.
- E-step: estimate probability that a point comes from a particular cluster
- M-step: do MLE on mean and variance (μ_i, σ^2_i) for each cluster.
- Repeat until (μ_i, σ^2_i) converge

E-M Algorithm Computations

$$p(c|x_j) = \frac{p(x_j|c)p(c)}{\sum_{i=1}^{k} p(x_j|i)p(i)}$$

E-Step

$$\mu_c = \frac{\sum_{j=1}^{N} p(c|x_j)x_j}{\sum_{j=1}^{N} p(c|x_j)}$$

$$\sigma_c^2 = \frac{\sum_{j=1}^{N} p(c|x_j)(x_j-\mu_c)^2}{\sum_{j=1}^{N} p(c|x_j)}$$

$$p(c) = \frac{\sum_{j=1}^{N} p(c|x_j)}{N}$$

M-Step

prior

Probabilistic Models

- Assume that some well defined probabilistic model with certain parameters, generates each document
- We know the model, but not the parameters.
- All we can observe are the documents and the words.
- How do we figure out the parameters?

Simple Model: Unigram Language Model

- Assume there is only one topic
- Assume you know probability distribution of words in the topic p(w|t)
- Assume you need to generate a document with this topic t
- Just pick each word from this distribution independently
- $p(w_1, w_2, \ldots w_n) = p(w_1).p(w_2)...p(w_n)$
- Will most likely get gibberish but ok first cut

How do we estimate parameters here?

- Observation, words in a document and their counts.
- Assume each word w, occurs c_w times
- Assume there are N words in total
- Best estimate of p(w)?

Quantum	20
Computing	30
Bell	50
Inequality	50
Schoredinger	20
Einstein	20
Podolsky	20
Rosen	20
Spin	30
Wave	60
Collapse	20
Measurement	15
Uncertainty	15
The	80
But	70

How do we estimate the Probabilities from Document

- Assume each word in document occurs c_w times.
- Assume probability of word w occurring is p_w
- Then for a given document d, $p(d|p_w) = \prod_{w=1}^{N} p_w^{c_w}$
- $\arg\max_{p_w} \prod_{w=1}^{N} p_w^{c_w}, s.t. \sum_{w=1}^{N} p_w = 1$

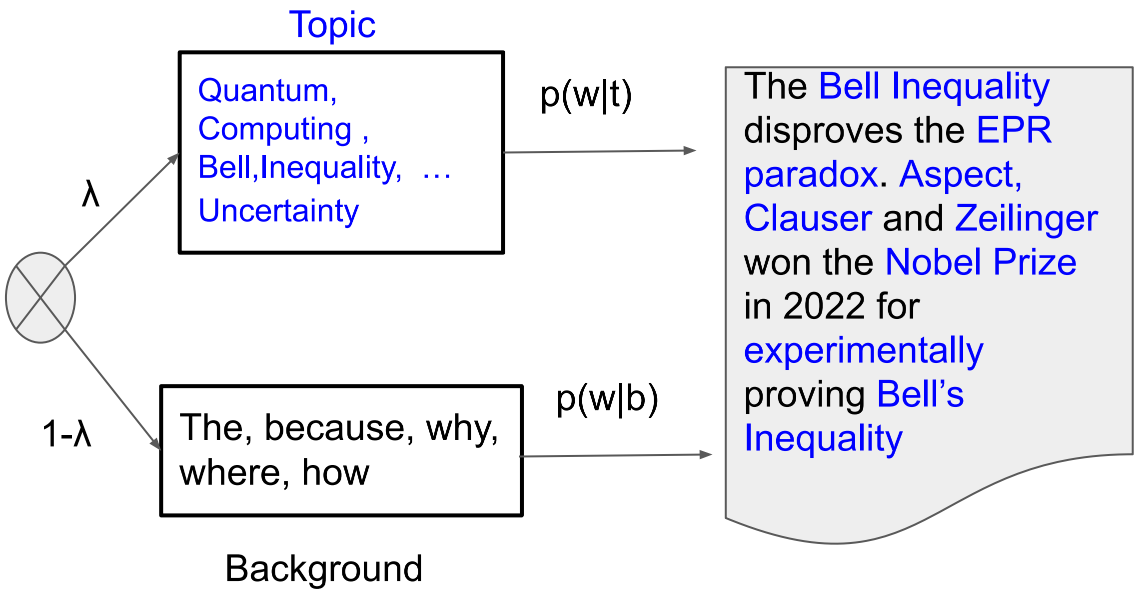
Taking log

- $\arg\max_{p_w} log(\prod_{w=1}^{N} p_w^{c_w}, s.t. \sum_{w=1}^{N} p_w = 1)$
- $\arg\max_{p_w} \sum_{w=1}^{N} c_w log(p_w), s.t. \sum_{w=1}^{N} p_w = 1)$
- $\arg\max_{p_w} \sum_{w=1}^{N} c_w log(p_w), +\lambda(p_w - 1)$

- differentiate w.r.t p_w and set to 0
- $\frac{c_w}{p_w} + \lambda = 0$
- $\implies p_w = -\frac{c_w}{\lambda}$
- $\sum_{w=1}^{N} -\frac{c_w}{\lambda} \ = 1$
- $\implies \lambda = -\sum_{w=1}^{N} c_w$
- $\implies p_w = \frac{c_w}{\sum_{w=1}^{N} c_w}$!!

But Some Words are very Common!!!

- We could eliminate stop words etc., alternatively
- Assume there is a background model and a topic model



MLE Estimate of Parameters

- $p(w) = \lambda p(w|t) + (1-\lambda)p(w|b)$
- $p(D|\Lambda) = \arg\max_{\lambda, p(w|t), p(w|d)} \prod_{w \in D} [\lambda p(w|t) + (1-\lambda)p(w|b)]$
- s.t. $\sum_{w \in D} p(w|t) = 1, \sum_{w \in D} p(w|d) = 1$

Simple Example

- $\lambda = 0.5$
- Only 2 words in document - *the, Bell*
- For background model, we know $p(Bell|b) = 0.1$ and $p(the|b) = 0.9$
- We now need to estimate $p(Bell|t)$ and $p(the|t)$
- i.e. $\arg\max_{p(w|t)} [0.5p(Bell|t) + 0.5 * 0.1] * [0.5p(the|t) + 0.5 * 0.9]$
- s.t. $p(Bell|t) + p(the|t) = 1$

What is the MLE?

- maximum is attained when
- $0.5p(Bell|t) + 0.5 * 0.1 = 0.5p(the|t) + 0.5 * 0.9$
- $p(Bell|t) = 0.9, p(the|t) = 0.1$!!

Gives Bell much higher probability for topic automatically!

What if Some Words Occur more Frequently?

- $\lambda = 0.5$
- 5 words in document - *the, the, the, the, Bell*
- $[0.5p(Bell|t) + 0.5 * 0.1] * [0.5p(the|t) + 0.5 * 0.9]^4$

Q: Will p(the | t) increase or decrease?

A: High frequency words will always have higher probability for a given topic

Q: What happens if we decrease λ?

A: Probability of background set ⬆ probability of high frequency word ⬇

How do we Estimate p(w | t)?

The Bell Inequality disproves the EPR paradox. Aspect, Clauser and Zeilinger won the Nobel Prize in 2022 for experimentally proving Bell's Inequality

- Use Expectation Maximization algorithm, assume λ and $p(w|b)$ known
- Define new hidden variable Z
- $Z = 0$ if w from topic t, else $Z = 1$
- $p^n(Z = 0|w) = \frac{\lambda p^n(w|t)}{\lambda p^n(w|t)+(1-\lambda)p^n(w|b)}$ E-Step
- $p^{n+1}(w|t) = \frac{c_w p^n(Z=0|w)}{\sum_{v \in V} c_v P^n(Z=0|v)}$ M-Step