Text Analytics & Business Application

Topic Modeling

Qinglai He

Department of Operations and Information Management

Wisconsin School of Business

WISCONSIN SCHOOL OF BUSINESS

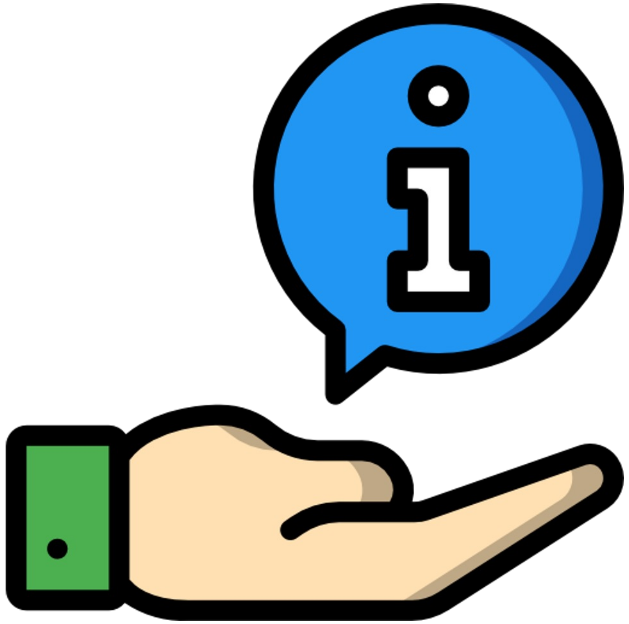
Outline of Today’s Class

- Intro to Text Modeling
- Latent Dirichlet Allocation
  - LDA Concepts
  - How does LDA work?
  - Pros & Cons
- Performance Measures
  - Perplexity
  - Coherence score

Intro to Topic Modeling

The Problem with Information

- Needle in a haystack: as more information becomes available, it is harder and harder to find what we are looking for.
- Need new tools to help us organize, search and understand information.

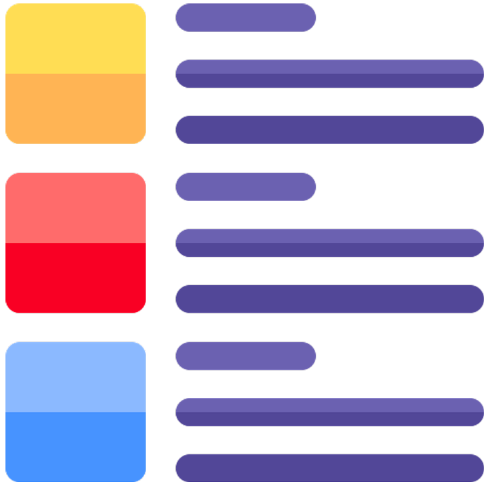


What is “Topic”?

Topics can be defined as “a repeating pattern of co-occurring terms in a corpus”.

Examples:

- “health”, “doctor”, “patient”, “hospital” – **Healthcare**
- “farm”, “crops”, “wheat” – **Farming**
- “planet”, “asteroid”, “orbit” – **Astronomy**
- “kill”, “attack”, “soldier”, “death” – **War**
- “player”, “round”, “win”, “goal” – **Game**



Topic Modeling Concepts

- Given a large collection of documents, it's not easy to make sense of them manually.
- We can break the documents into words and phrases. Group these words and phrases together.
  - Text clustering
  - Topic modeling
- Topic modeling is a collection of unsupervised statistical learning methods used to discover latent topics in large text documents.
- It is useful for analyzing various forms of text
  - News articles, tweets, etc.

```
Topic #0: 0.013*"jacky" + 0.006*"dahlia" + 0.005*"novel" + 0.005*"one" + 0.004*"story" + 0.004*"also" + 0.004*"book" +
0.004*"team" + 0.004*"narrator" + 0.003*"jeremy“

Topic #1: 0.010*"book" + 0.009*"war" + 0.006*"in" + 0.006*"world" + 0.005*"novel" + 0.005*"states" + 0.004*"also" +
0.004*"new" + 0.004*"chapter" + 0.004*"story“
```

Topic Modeling Techniques

- Popular topic modeling algorithms include:
  - Latent Dirichlet Allocation (LDA)
  - Latent Semantic Analysis (LSA)
  - Probabilistic Latent Semantic Analysis (PLSA).
- LDA is the technique most commonly used in practice.

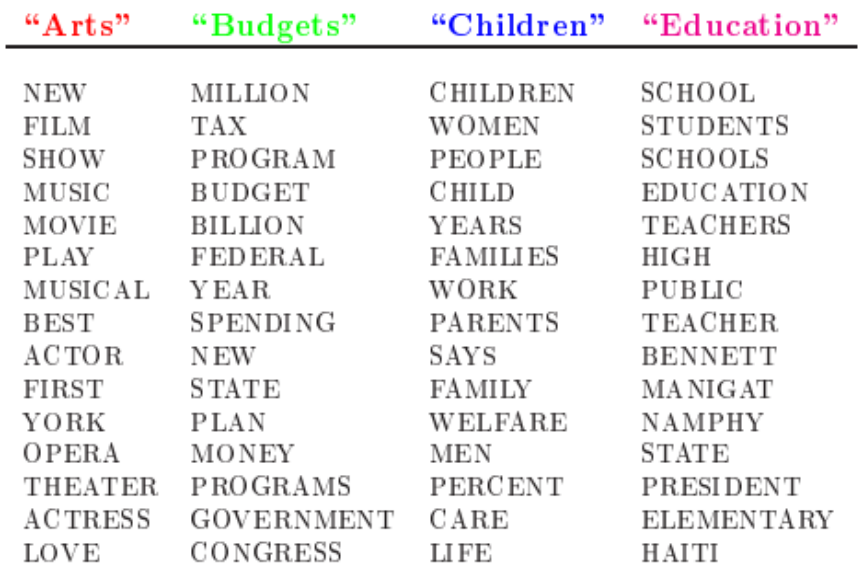
LDA (Latent Dirichlet Allocation)

What Is This Paragraph About?

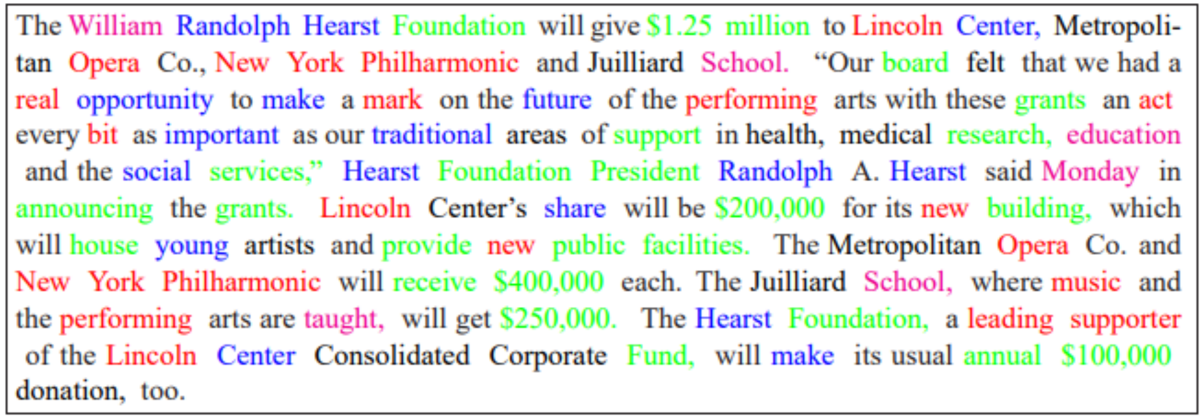
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that w had a real opportunity to make a mark on the future of the performing art with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services.” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $200,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

LDA Concepts

- Latent refers to hidden variables.
- A Dirichlet distribution is a probability distribution over other probability distributions.
- Allocation means that some values are allocated based on the two distributions.



"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI



The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

References

https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf https://en.wikipedia.org/wiki/Dirichlet_distribution

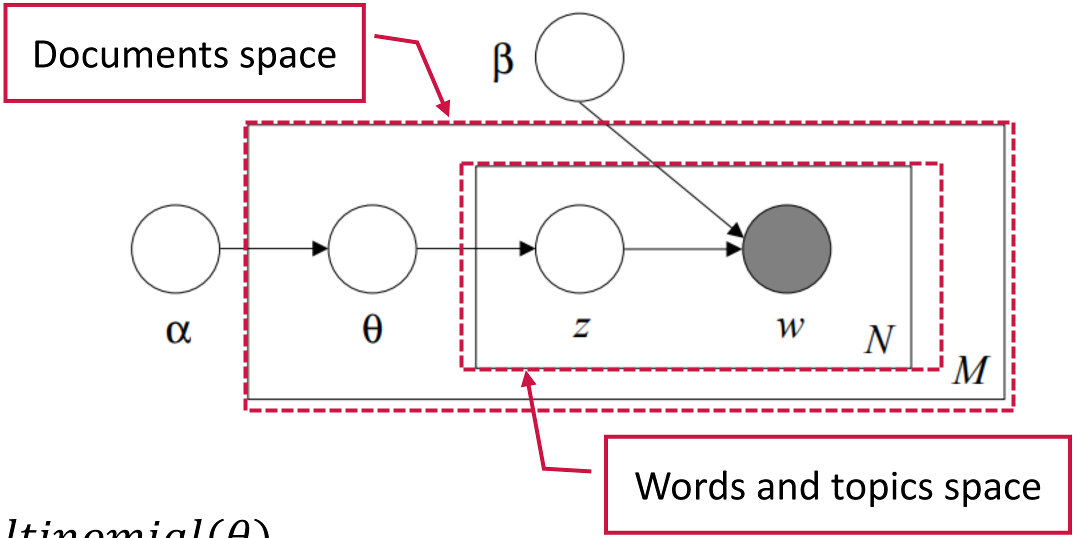
Assumption of LDA

LDA assumes:

1. Documents are a mixture of topics
   - The documents under consideration are produced from a mixture of topics
2. We have a list of topics with a probability distribution
3. Topics are a mixture of words
   - For every topic, there's an associated list of words with a probability distribution
4. We sample k topics from topic distribution
   - K, the number of topics, is a hyperparameter. The optimal value for k is found by trail and error.
5. For each of the k topics selected, we sample words from the corresponding distribution

Human interpretation is required to understand and name the identified topics

Mathematically…



- Each document contains K topics $(z = 1, \dots, K)$
  - The topic z follows Multinomial distribution, i.e., $z \sim Multinomial(\theta)$
  - θ is a hyper parameter follow the Dirichlet distribution with given parameter α, i.e., $\theta \sim Dirichlet(\alpha)$
- Words in the document are generated from those topics.
  - Word w follows Multinomial distribution according to topic z , i.e., $w \sim Multinomial(\phi_z)$
  - ϕ follows the Dirichlet distribution with given parameter β, i.e., $\phi \sim Dirichlet(\beta)$
- All documents contain a particular set of topics, but the proportion of each topic in each document is different.
- **The goal is to infer the multinomial parameters $\boldsymbol{\theta}$ for each document, and $\boldsymbol{\phi_z}$ for each topic.**

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \dots \theta_k^{\alpha_k - 1}$$

$$\theta_i \geq 0, \qquad \sum_{i=1}^{k} \theta_i = 1$$

https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Inference

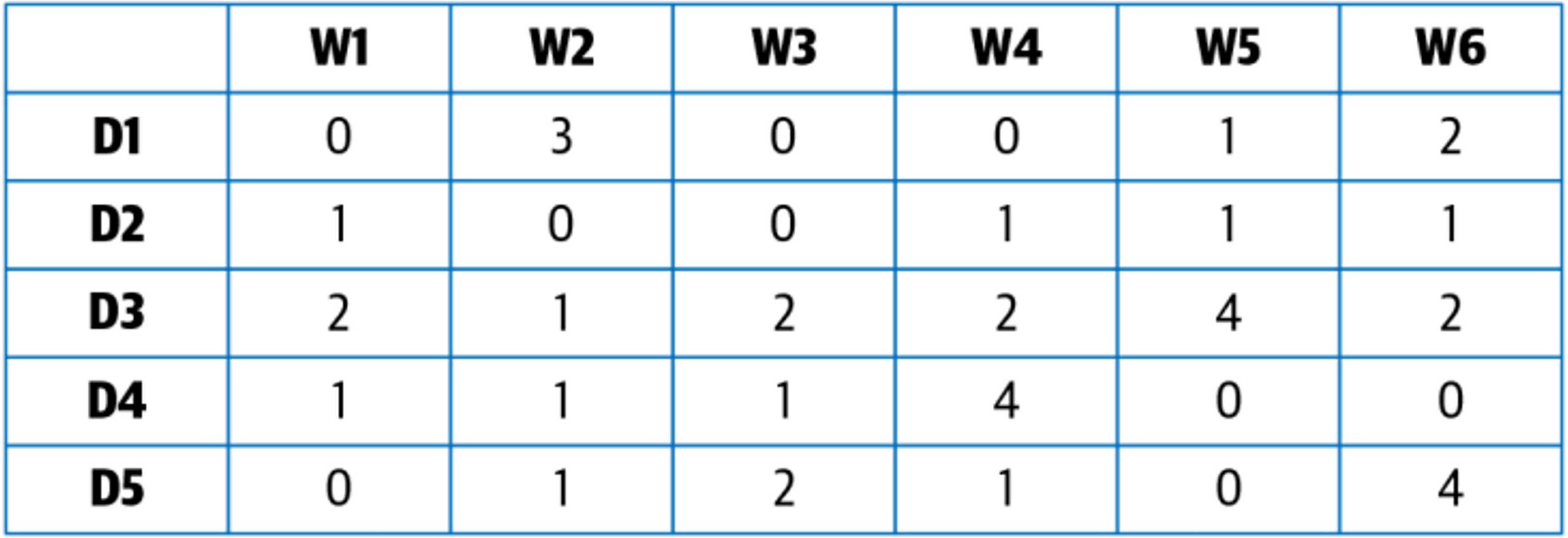
- *(Not the focus of this course)* How to uncover the hidden topic and word distributions in a corpus?
  - Approximation methods: search over the topic structure
    - **Sampling–based algorithm** attempt to collect samples from the posterior to approximate it with an empirical distribution.
    - **Variational algorithms** posit a parameterized family of distributions over the hidden structure and then find the member of that family that is closest to the posterior.

Another Way to Explain the Goal

- Given a set of documents, LDA tries to backtrack the generation process and figure out what topics would generate these documents in the first place.
- How does LDA do this backtracking?
  - It does so by factorizing a document-term matrix (M) that keeps count of words across all documents.
  - It has all the m documents D1 , D2 , D3 … Dm arranged along the rows and all the n words W1 ,W2 , ..,Wn in the corpus vocabulary arranged as columns.
  - M[i,j] is the frequency count of word Wj in Document Di

Backtracking with LDA

Document–term matrix (M):



	W1	W2	W3	W4	W5	W6
D1	0	3	0	0	1	2
D2	1	0	0	1	1	1
D3	2	1	2	2	4	2
D4	1	1	1	4	0	0
D5	0	1	2	1	0	4

Backtracking with LDA

- Note that if each word in the vocabulary represents a unique dimension and the total vocabulary is of size n, then the ith row of this matrix is a vector that represents the ith document in this n-dimensional space.
- LDA factorizes M into two submatrices: M1 and M2. M1 is a document–topics matrix and M2 is a topic–terms matrix, with dimensions (M, K) and (K, N), respectively.

Backtracking with LDA

- LDA converts this document-word matrix into two other matrices: document-topics matrix and topic-terms matrix as shown below:

	W1	W2	W3	W4	W5	W6
D1	0	3	0	0	1	2
D2	1	0	0	1	1	1
D3	2	1	2	2	4	2
D4	1	1	1	4	0	0
D5	0	1	2	1	0	4

	K1	K2	K3	K4
D1	1	0	0	1
D2	1	1	0	0
D3	1	0	0	1
D4	1	0	1	0
D5	0	1	1	1

	W1	W2	W3	W4	W5	W6
K1	1	0	0	1	0	0
K2	0	1	1	0	1	1
K3	1	1	0	1	1	0
K4	1	0	0	0	1	0

How the Results Look Like?

- A topic modeling results using Yelp data.
- We generated 5 topics

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
0.030	love	0.052	back	0.088	place	0.067	food	0.023	friendly
0.016	always	0.038	go	0.028	amazing	0.065	good	0.023	restaurant
0.014	recommend	0.024	wait	0.022	think	0.047	service	0.023	taste
0.013	way	0.020	get	0.022	star	0.036	time	0.017	meal
0.013	menu	0.020	come	0.013	awesome	0.032	great	0.016	staff
0.012	cold	0.018	eat	0.012	pretty	0.018	really	0.016	want
0.011	breakfast	0.017	definitely	0.011	well	0.017	make	0.015	dish
0.011	friend	0.016	server	0.011	give	0.016	bad	0.015	fry
0.011	customer	0.013	try	0.010	burger	0.016	go	0.015	order
0.010	rude	0.012	take	0.010	vegas	0.016	also	0.015	fresh

LDA Pros and Cons

Pros

- LDA can assign a topic probability to a new document thanks to the document-topic Dirichlet distribution.
- It can be applied to both short and long documents.
- Topics are open to human interpretation.
- As a probabilistic module, LDA can be embedded in more complex models or extended. Following the original work by Blei et al. (2013), many work extended LDA and addressed some original limitations

LDA Pros and Cons

Cons

- The number of topics must be known beforehand.
- The bag-of-words approach disregards the semantic representation of words in a corpus.
- It requires an extensive pre-processing phase to obtain a significant representation from the textual input data.
- Topics are distributions over words; therefore, interpreting topics semantics from these distributions becomes important. LDA lacks guidance on how to properly interpret topics.
- Studies report LDA may yield *too general* (Rizvi et al., 2019) or *irrelevant*(Alnusyan et al., 2020) topics. Results may also be inconsistent across different executions (Egger et al., 2021).

Other Topic Modeling Methods

- LSA
- PLSA
- Many other LDA variants
- Word embedding-based topic modeling (Das et al. 2015)
- …

Das, R., Zaheer, M., & Dyer, C. (2015). Gaussian LDA for Topic Models with Word Embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Vol. 1, pp. 795–804).

Performance Measures

Metric 1: Perplexity

- Perplexity is the most typical evaluation of LDA models (Bao & Datta, 2014; Blei et al., 2003).
  - Perplexity measures the modeling power by calculating the inverse log-likelihood of unobserved documents (a decreasing function).
  - There are many variations for calculating the perplexity scores. Below is just one version of it.

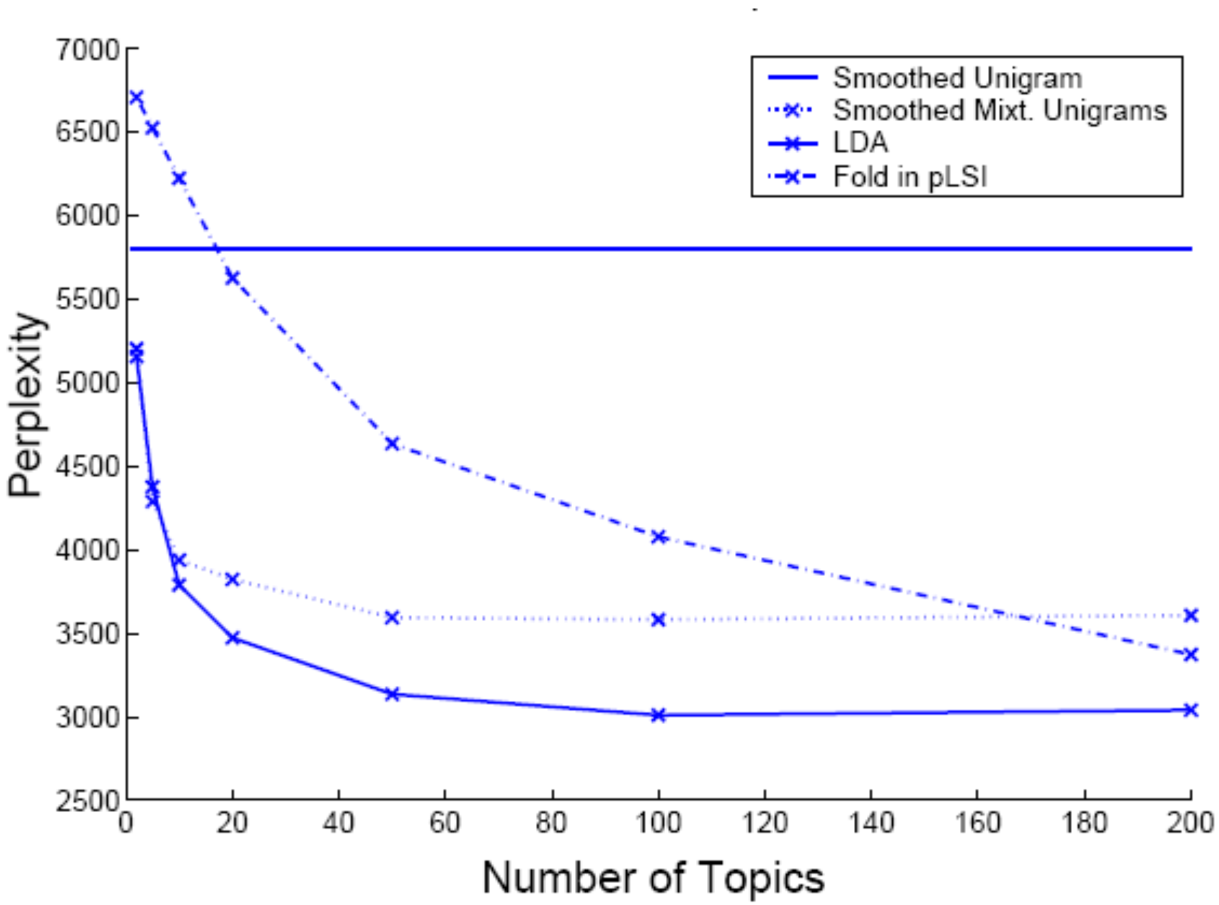
$$perplexity(D_{test}) = \exp\left\{ - \overbrace{\frac{\sum_{d=1}^{M} \boxed{\log p(\mathbf{w}_d)}}{\sum_{d=1}^{M} N_d}}^{\text{Average log-likelihood of all unobserved document}} \right\}$$

Log-likelihood of each unobserved document

$\mathbf{W}_d$: words in document d;
N_d: Length of document d

References: Weifeng Li & Hsinchun Chen, An overview of topic modeling.

Metric 1: Perplexity



The figure compares LDA with other topic modeling approaches. The LDA model is consistently better than all other benchmark approaches. Moreover, as the number of topics go up, the LDA model becomes better (i.e., the perplexity decreases.)

Metric 1: Perplexity

- Perplexity is commonly used to compare the performance of different language models or to evaluate the effectiveness of different model training approaches.
- Better models have **lower perplexity**, suggesting less uncertainties about the unobserved document.
- Limitations
  - Perplexity is not perfect. More topics will lead to lower perplexity.
  - Researchers found that perplexity is not strongly correlated to human judgment. Sometimes, it is even anti-correlated.

References: Weifeng Li & Hsinchun Chen, An overview of topic modeling.

Metric 2: Coherence Score

- We sort each topic from highest to lowest term weight and then select the first n terms for each topic. We then measure the coherence for terms in each topic, which essentially measure how similar these words are to one another.
- Coherence score in topic modeling to measure how **interpretable** the topics are to humans.
  - CV Coherence Score
    - the default metric in the Gensim topic coherence pipeline module.
  - UMass Coherence Score
  - UCI Coherence Score
  - Word2vec Coherence Score

References: Weifeng Li & Hsinchun Chen, An overview of topic modeling.

Metric 2: Coherence Score

- Topic coherence evaluates the semantic nature of the learned topics.
  - Specifically, it measures the semantic similarity among the top keywords of a topic. Topic coherence has shown to be correlated with human evaluations of topic quality.
- Topic coherence a topic $\boldsymbol{\beta}_k$ is calculated by:

$$\text{coherence}(\boldsymbol{\beta}_k) = \sum_{(w_i,w_j)\in V_n} \text{score}(w_i, w_j)$$

  - where V_n is the top n keywords of the topic $\boldsymbol{\beta}_k$
- There are two commonly used score metrics:

The Extrinsic UCI Metric (Newman et al. 2010):

$$\text{score}(w_i, w_j) = \log p(w_i, w_j)/p(w_i)p(w_j)$$

where $p(w_i, w_j)$ is the word co-occurrence probability of word pair w_i, w_j estimated from an *external* corpus (e.g., Wikipedia) and $p(w_i)$ is the probability of word w_i in the external corpus.

The Intrinsic UMass Metric (Mimno et al. 2011):

$$\text{score}(w_i, w_j) = \log\left(\mathrm{D}(w_i, w_j) + 1\right)/\mathrm{D}(w_j)$$

where $\mathrm{D}(w_i, w_j)$ counts the number of documents word pair w_i, w_j co-occurred and $\mathrm{D}(w_j)$ counts the number of documents containing w_j.

References: Weifeng Li & Hsinchun Chen, An overview of topic modeling.

How Many Topics to Choose?

- The author of LDA suggests to select the number of topics from 50 to 150 (Blei 2012); however, the optimal number usually depends on the size of the dataset.
- Cross validation on perplexity/coherence score is often used for selecting the number of topics.
  - Specifically, we propose possible numbers of topics first, evaluate the average perplexity using cross validation, and pick the number of topics that has the lowest perplexity/highest coherence score.

References: Weifeng Li & Hsinchun Chen, An overview of topic modeling.

Choosing the Best Coherence Score

- There is no one way to determine whether the coherence score is good or bad. The score and its value depend on the data that it’s calculated from.
    - For instance, in one case, the score of 0.5 might be good enough but in another case not acceptable. The only rule is that we want to **maximize this score**.
    - Elbow curve can be applied to help select the optimal number of topics.

References: Weifeng Li & Hsinchun Chen, An overview of topic modeling.

!
?
QUIZ

Q1. Topic modeling uses a/an _________ machine learning technique.

A. Supervised
B. Unsupervised

Answer: B

Q2. LDA model will explicitly indicate the the identified topics.

A. True

B. False

Answer: B

QUIZ

Q3. Which of the following statement is the most accurate about coherence score?

A. The coherence score is the correlation of words between topics
B. The coherence score will decrease with the increase in the number of topics
C. The coherence score in topic modeling measures the interpretability of topics for humans.
D. We can determine whether the coherence score in a model is good or bad.

Answer: C

QUIZ

Q4. The higher the perplexity, the more accurate the model is.

A. True
B. False

Answer: B

Take 10 minutes break…

Python - LDA Workflow

- Clean data
  - Remove URLs, HTML tags, emails, and non-alpha characters etc.
  - Text preprocessing, such as removing stop words and normalizing words
- Make the dictionary and corpus
- Test both corpus using the LDA model
  - Evaluate model with Perplexity or Coherence Score
- Visualize and analysis

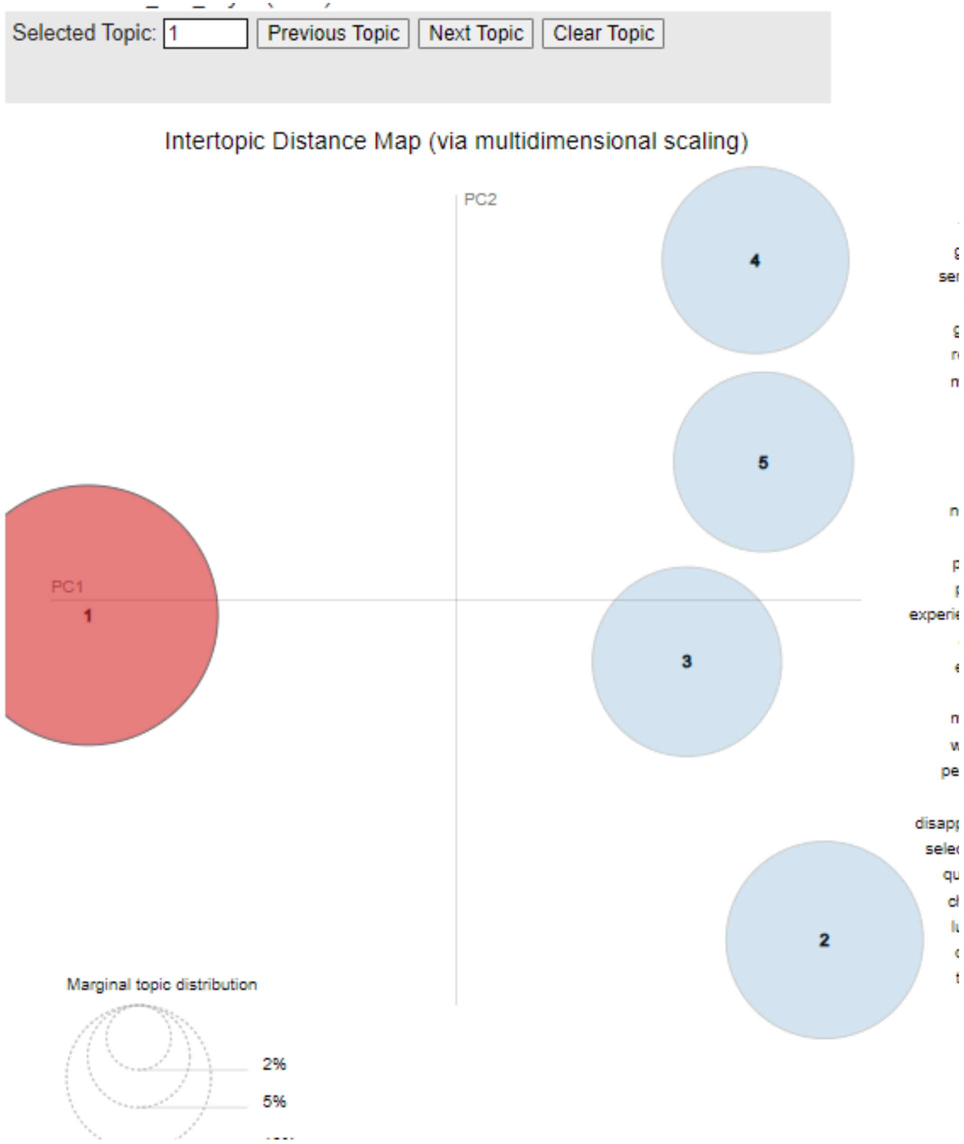
Topic Modeling and Latent Dirichlet Allocation (LDA) | DataScience+ (datascienceplus.com)

LDA Interpretation

- In the coding example, we are going to use the yelp review data.
- We generated 5 topics

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
0.030	love	0.052	back	0.088	place	0.067	food	0.023	friendly
0.016	always	0.038	go	0.028	amazing	0.065	good	0.023	restaurant
0.014	recommend	0.024	wait	0.022	think	0.047	service	0.023	taste
0.013	way	0.020	get	0.022	star	0.036	time	0.017	meal
0.013	menu	0.020	come	0.013	awesome	0.032	great	0.016	staff
0.012	cold	0.018	eat	0.012	pretty	0.018	really	0.016	want
0.011	breakfast	0.017	definitely	0.011	well	0.017	make	0.015	dish
0.011	friend	0.016	server	0.011	give	0.016	bad	0.015	fry
0.011	customer	0.013	try	0.010	burger	0.016	go	0.015	order
0.010	rude	0.012	take	0.010	vegas	0.016	also	0.015	fresh

Visualization of LDA Results



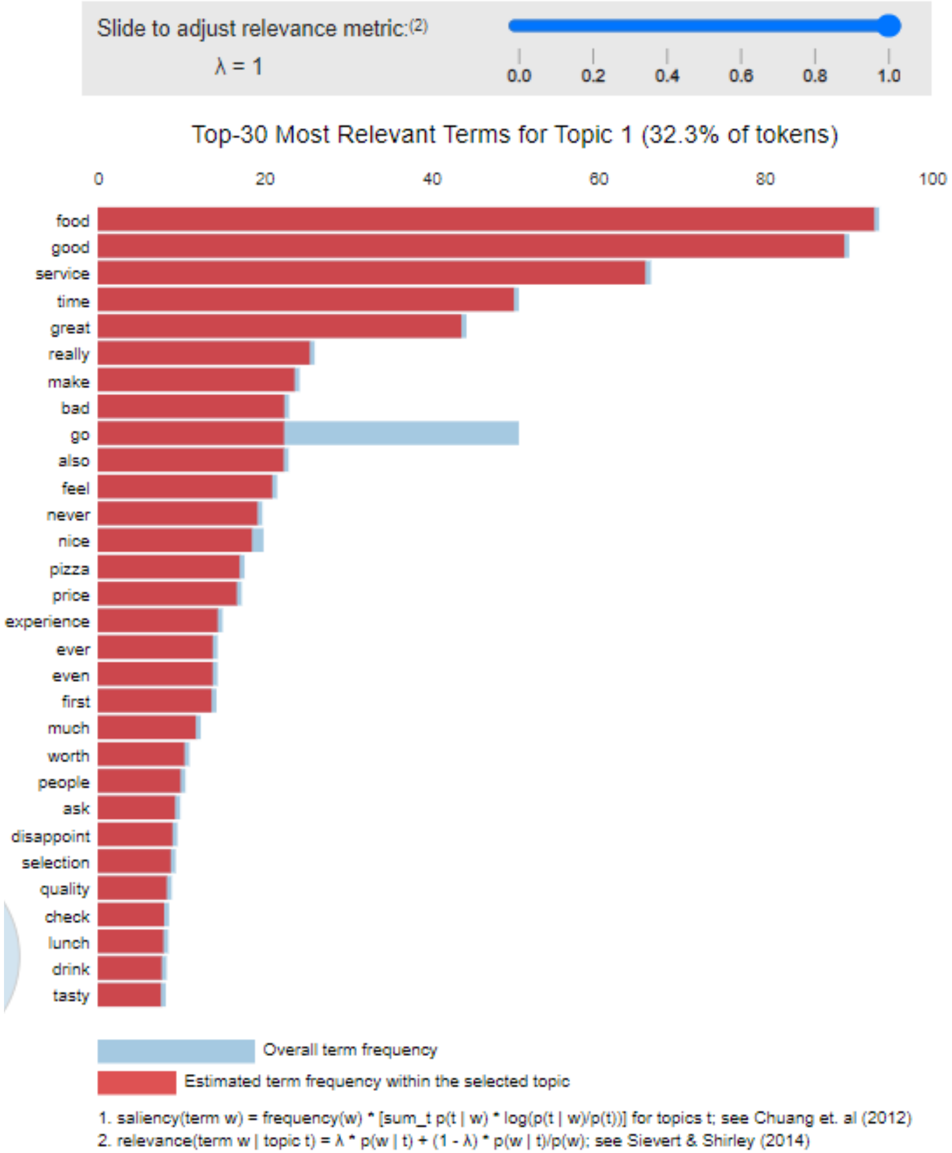
Selected Topic: 1

Intertopic Distance Map (via multidimensional scaling)

- Each bubble represents a topic.
- The larger the bubble, the higher percentage of the number of documents in the corpus is about that topic.
- The further the bubbles are away from each other, the more different they are.

Topic Modeling and Latent Dirichlet Allocation (LDA) | DataScience+ (datascienceplus.com)

LDA Visualization Example



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Topic Modeling and Latent Dirichlet Allocation (LDA) | DataScience+ (datascienceplus.com)

- Blue bars represent the overall frequency of each word in the corpus.
- If no topic is selected, the blue bars of the most frequently used words will be displayed.
- Red bars give the estimated number of times a given term was generated by a given topic.
- The word with the longest red bar is the word that is used the most by the document belonging to that topic.

Find out more about LDA with Python

- Topic Modeling and Latent Dirichlet Allocation (LDA)
- Using LDA Topic Models as a Classification Model Input

Exercises using Google Colab