LIVE SESSION 9 (NOC24_CS39)

Shubhi Bansal
PMRF, IIT Indore

Unsupervised algo **

LDA : Latent Dirichlet Allocation

hidden / ~~not known~~ apriori (not known beforehand)

topics are unknown / hidden in the data

The document ∈ topics   topics are believed to be

present ∴ text is generated based on topics

Dirichlet : distribution of distributions

Document $\equiv$ topics $\equiv$ words

document

topic

words

Dirichlet is distribution of topics in doc.

~~topic~~

& distribution of words in a topic

<u>Allocation</u>: once we have Dirichlet, we will allocate topics to documents & words of the

document to <u>topics</u>

LDA: each word in each doc. comes from a topic & the topic is selected from a per-document distribution of topics

$\theta_{t,d}$

$\alpha_{t,d} = P(t|d) =$ Prob. distribution of topics in a document.

$\phi_{w,t}$

$\beta_{w,t} = P(w|t) =$ Prob. distribution of words in a topic

$$\text{LDA} = \sum_{t \in T} P(w|t) \cdot P(t|d)$$

$P(w|t)$

LDA :

$$= \sum_{t \in T} P(w|t,d) * P(t/d)$$

Assume we have conditional independence

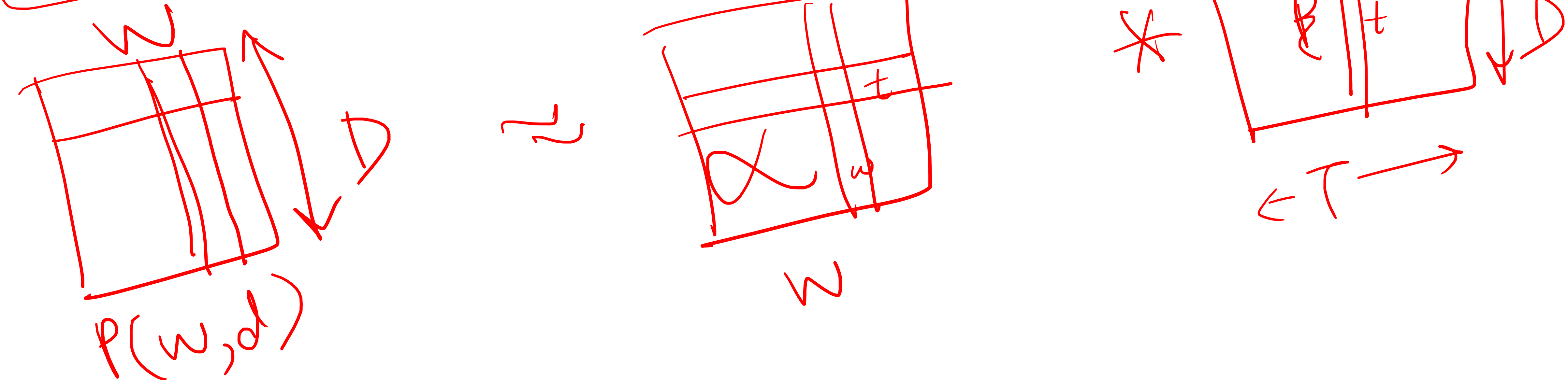
$$\sum_{t \in T} P(w|t) * P(t/d)$$

T : # topics

W (V) : # words in the entire collection of documents

vocabulary

$$P(w|d) = \sum_{t=1}^{T} P(w|t) \cdot P(t|d)$$



$\theta \quad \phi$

Dirichlet parameters $\underbrace{(\alpha, \xi)}$

↳ control if all words have same probability in a topic

OR

will that topic have extreme bias towards some words dataset : all news articles of ~~country~~ France from 2018

I want to make use of LDA to find out topics

eg:} France won 2018 world cup

Document has words (V) = { football, world cup 2018 winners france } $|V| = 5$

~~Given~~ :

Step Random assignment of topics to words in a doc

	T3	T2	T1	T3	T1
Doc i	football	world cup	2018	winners	france

Given K = No. of topics = 3

	T1	T2	T3
Doc i	2	~~1~~ 0	2

Plus, you also have a count how many times a word is associated with a given topic

	T1	T2	T3
football	1	0	35
world cup	10	~~8~~ 7	1
2016	42	1	0
winners	0	0	20
france	50	0	1

Idea: After random initialization, you want to converge at a point where words signify the topic that you're trying to figure out so " we can go on reassigning the topic of each word at every pass".

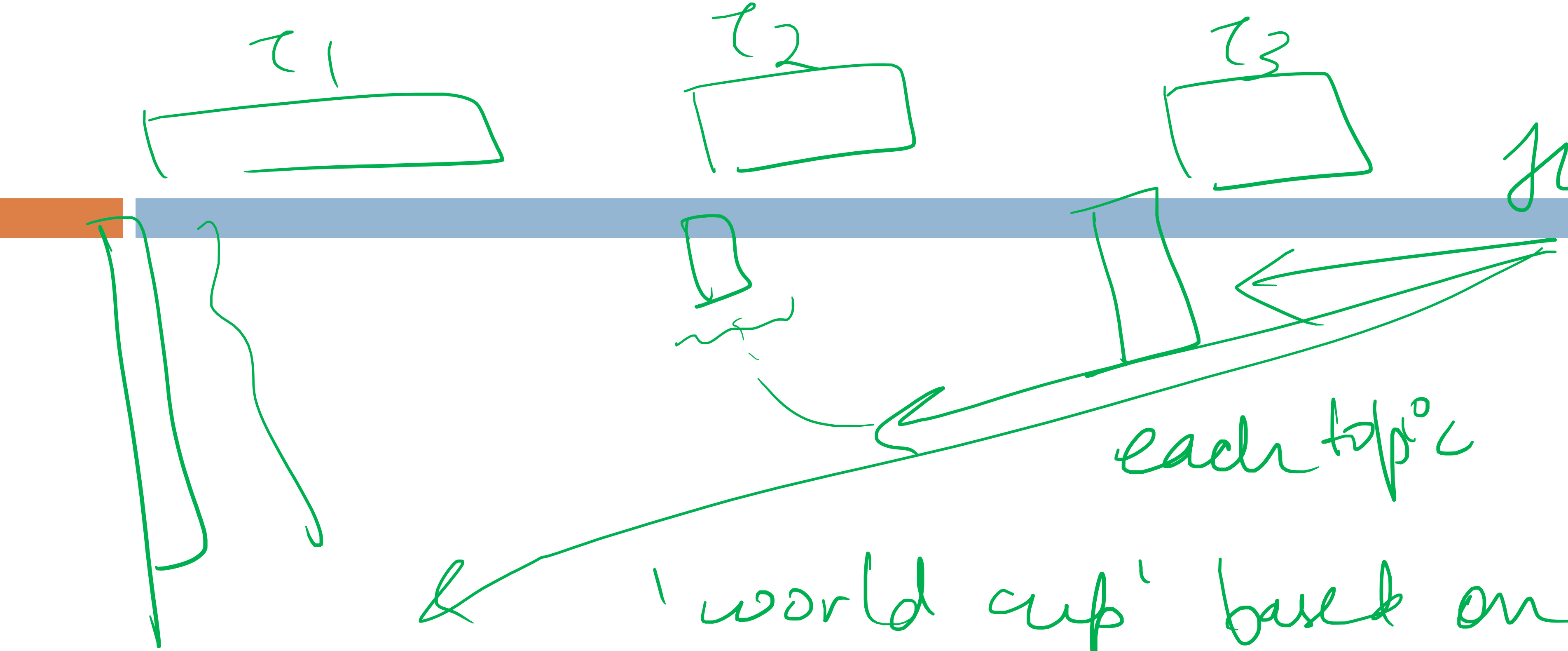
I reduce the count of world cup $\in \tau 2$ from 1 to 0
& remove the topic assigned to it so
then the count changes

Reassign the topic based on probability distribution

Topic 1 Topic 2 Topic 3

	T1	T2	T3
Doc i	2	0	2

How much doc likes each topic based on other assignments in the doc



How much each topic likes the word 'world cup' based on ~~other~~ assignments in other documents

	τ_1	τ_2	τ_3
world cup	10	7	1

How much doc likes each topic & How much topic likes a word.

Repeat for all words in corpus in one pass & depending on how many passes you're in LDA setup; this process after "reassigning" topics for all words at every pass & that; a stage will come when whole convergence would happen]

Question 1

In Topic modeling which hyperparameters tuning used for represents document-topic Density? 'distribution

a) Dirichlet hyperparameter Beta

b) Dirichlet hyperparameter alpha

c) Number of Topics (K)

d) None of them

Answer - b) Dirichlet hyperparameter alpha

Question 2

n Topic modeling which hyper parameters tuning used for represents Word-Topic Density?

a) Alpha parameter

b) Number of Topics (K)

c) Beta parameter

d) None of them

Ans: c

Question 3

Classically, topic models are introduced in the text analysis community for__________________ topic discovery in a corpus of documents.

a) Unsupervised.

b) Supervised.

c) Semi-automated.

d) None of the above.

Answer - a) Unsupervised

- LDA is an unsupervised learning algorithm, meaning it doesn't use labeled data to guide its learning process. It operates solely on the assumption that documents are generated from a fixed number of topics, and the goal is to uncover these topics. Without external guidance or criteria, it cannot determine the optimal number of topics.

spam ham

3 topics

$|V| = 5$

w_1 T_1 T_2 T_3

email 1

email 2

email 3

Question 4

Topic model techniques is/are __________ .

a) Latent semantic indexing (LSI).

b) Probabilistic latent semantic analysis (PLSA).

c) Latent Dirichlet allocation (LDA).

d) All of the above.

Answer - d) All of the above

- Latent semantic indexing (LSI), probabilistic latent semantic analysis (PLSA), and latent Dirichlet allocation (LDA) are all **topic model techniques.**
- **LSI** and **PLSA** are considered forerunners to LDA. They identify underlying themes in documents.
- **LDA** is the most popular technique and builds upon the foundation of LSI and PLSA.

topic modeling techniques

LSI PLSA

ancestors

LDA

Question 5

____________ is a scoring of how rare the word is across documents.

a) Inverse Document frequency.

b) Term frequency.

c) File frequency.

d) None of the above.

Answer - a) Inverse Document frequency

raw count

spots

$$TF\text{-}IDF = \underbrace{TF(t, d)}_{\text{raw count}} * IDF(t)$$

?

$$= \left(\frac{\text{No. of times } t \text{ appears in } d}{\text{No. of terms in } d} \right) * \left(\frac{\text{No. of docs in corpus}}{\text{No. of } \cancel{\text{times}} \text{ docs that contain the term } t} \right)$$

all terms

[gives higher weightage to terms that occur with less freq.

Question 6

Latent Dirichlet Allocation (LDA) and Latent Semantic Allocation (LSA) are based on ________________________________ assumptions.

a) Distributional hypothesis.

b) Statistical mixture hypothesis.

c) Both of the above.

d) Not any from (a) and (b).

Answer - c) Both of the above

Dirichlet

$\rightarrow P(t|d) * P(w|t)$

pitch trophy

world cup ∈ sports

cricket

- **Distributional hypothesis:** This assumption applies to both LDA and LSA. It states that words with similar meanings tend to appear in similar contexts within documents. This allows the models to identify relationships between words based on how often they co-occur.
- **Statistical mixture hypothesis:** This assumption is particularly important for LDA. It suggests that documents are a mixture of latent topics, and each topic is characterized by a probability distribution over words. This allows LDA to not only identify topics but also represent the proportion of each topic within a document.

$\boxed{P(w|t) * P(t|d)}$

- LSA leverages the distributional hypothesis to uncover semantic relationships, while LDA builds upon that foundation with the statistical mixture hypothesis to model documents as a combination of latent topics.

Question 7

One of the basic assumptions of LDA and LSA as a distributional hypothesis which means ________________________.

a) Similar topics make use of similar words.

b) Different topics make use of similar words.

c) Similar topics make use of different words.

d) None of the above.

Question 8

D.H. S.H.

Latent semantic analysis

One of the basic assumptions of LDA and LSA as a statistical mixture hypothesis which means ___________.

a) Documents talk about several topics.

b) Similar topics make use of similar words. Distributional hypothesis

c) Documents talk about prefixed topics.

d) None of the above.

You just specify K(# of topics)

Question 9

- **LSA, LDA also ignores syntactic information and treats documents as bags of words. True/False**

TRUE

* D.H. Lovely, beautiful ∈ aesthetics

sunrise, sunset ∈ nature | photography J.H. / D.H.

- Both LSA and LDA focus on word co-occurrence and treat documents as bags of words, ignoring the syntactic structure and word order within the documents. This simplification allows them to handle large amounts of text data efficiently but can miss out on capturing the nuances of language conveyed through sentence structure and grammar.

** S.H. = $P(w|t) * P(t|d)$

POS

Dependency tree

V, Imp

Question 10

α, β

Choose the correct statement from below –

I. A low value of alpha will assign fewer topics to each document whereas a high value of alpha will have the opposite effect.
II. A low value of beta will use fewer words to model a topic whereas a high value will use more words, thus making topics more similar between them.
III. LDA cannot decide on the number of topics by itself. (correct)

specified by user

programmer

a) (I).

b) (II).

c) III).

d) All of the above.

- Latent Dirichlet Allocation (LDA) is a powerful tool for uncovering hidden thematic structures in text data, but it has one key limitation: it can't determine the optimal number of topics on its own. Here's why:
  - **Trade-off between granularity and coherence:** Imagine a collection of documents. With a very high number of topics, LDA might identify very specific themes, like "baking cookies" or "fixing a leaky faucet." While these are technically topics, they may not be very informative. On the other hand, with too few topics, LDA might lump together unrelated concepts under a broad umbrella like "food" or "home improvement." There's a sweet spot where the topics are both specific and meaningful.
  - **The model doesn't understand meaning:** LDA is a statistical model that works with word probabilities, not semantics. It doesn't inherently "know" what a good topic is. It can only find clusters of words that frequently co-occur. The number of clusters it finds might not directly correspond to the number of meaningful themes in your data.

Question 11

- For question 8 use the following information. Suppose you are using Gibbs sampling to estimate the distributions, ~~&~~ α and ß for topic models. The underlying corpus has 3 documents and 5 words, {machine, learning, language, nature, vision} and the number of topics is 2. At certain point, the structure of the documents looks like the following Doc1: nature(1) language(1) vision(1) language(1) nature (1) nature (1) language(1) vision(1) Doc 2: nature(1) language(1) language(2) machine(2) vision (1) learning (2) language(1) nature(1) Doc3: machine (2) language (2) learning (2) language(2) machine(2) machine(2) learning(2) language(2) (number) -number inside the brackets denote the topic no. 1 and 2 denote whether the word is currently assigned to topics t1 and t2 respectively. η = 0.3 and a = 0.3 For question 8 calculate the value upto 4 decimal points and choose your answer 8) Using the above structure the estimated value of ß(2) nature at this point is

Doc2 : 1+1+1 = 3

Doc3 : 8

- 3 documents (D)
- 5 words ($|V| = 5$) {m/c, learning, lang, nature, vision}
- 2 topics (K)
- $\eta = 0.3$
- $\alpha = 0.3$

nature $\in$ T2

$\beta(2) = ?$ "nature" : $\not\alpha$.

Doc 1 : 0 times $\in$ T2.

Doc2 : 0 times $\in$ T2

Doc3 : 0 times $\in$ T2

$z(2) = $ nature

To calculate $z(2)$; the word 'nature' at this pt. in Gibbs sampling; we need to see how many times the word 'nature' is assigned to topic 2 and then add smoothing term.

$(\eta = 0.3)$

$$\xi(2) = \left(\frac{\text{count of nature assigned to topic 2} + \eta}{\text{count of all words assigned to topic 2} + \eta * \text{vocabul size}} \right)$$

(~~words~~) (occurrences)

~~Lo~~ duplicacy

$$= \left(\frac{0 + 0.3}{11 + 5 * 0.3} \right) = \left(\frac{0.3}{12.5} \right) = 0.0240$$

(doc1, doc2, doc3)

count of nature assigned to topic2 = 0

count of all words - - - - - - - =