Natural Language Processing

Assignment- 9

TYPE OF QUESTION: MCQ

Number of questions: 7 **Total mark: 10 [4*1 + 3*2] (Q5, Q6, Q7 carries two marks each)**

Question 1.

Which of the following is false?

1. Dirchlet distribution is an exponential family distribution
2. LDA is a generative model
3. Dirchlet distribution is taken over the simplex i.e positive vectors that sum to one
4. A higher value of alpha will assign fewer topics to each document whereas a high value of alpha will have the opposite effect.

Answer : 4

Solution : Refer to Lecture 43 of Week 9

Question 2:

In Topic modeling which hyperparameters tuning used to represent document-topic Density?

1. Dirichlet hyperparameter Beta word topic intensity
2. Dirichlet hyperparameter alpha document topic intensity
3. Number of Topics (K)
4. None of them

Answer: 2

Solution:

alpha is used to represent document-topic intensity

Question 3:

You have a topic model with the parameters α = 0.8 and β = 0.03. Now, if you want to have sparser distribution over words and denser distribution over topics, what should be the values for α and β?

1. Both α and β values should be decreased
2. Both α and β values should be increased
3. α should be decreased, but β should be increased
4. α should be increased, but β should be decreased

Answer: 4

Solution:

α : topic distribution

β : word distribution

Question 4 :

In Gibbs sampling choose the correct option from below

1. It can not directly estimate the posterior distribution over *z*
2. It is a form of Markov chain Monte Carlo
3. Here sampling is done in parallel
4. Sampling is stopped before sampled values approximate the target distribution

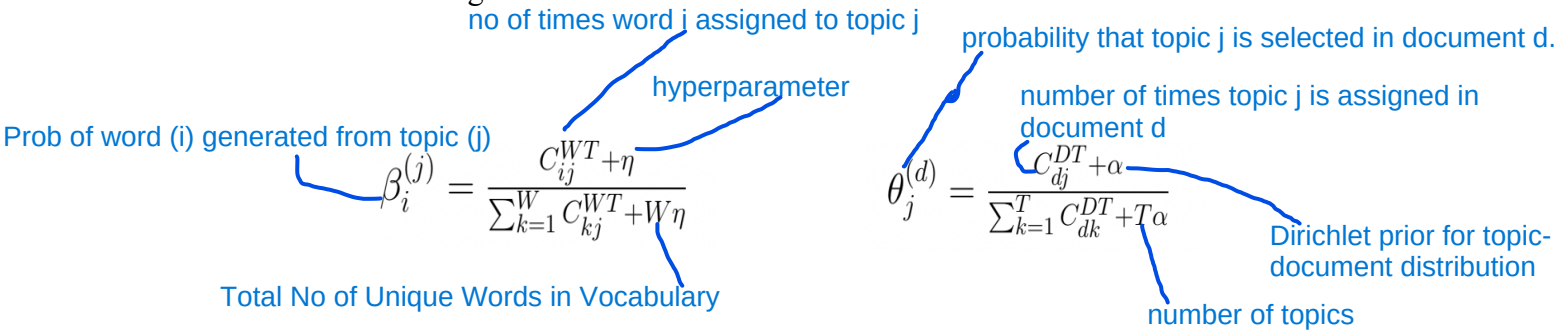
Answer: 2

Solution:

In gibbs sampling, we do sequential sampling until the sampled values approximate the target distribution. This also can directly estimate the posterior distribution over *z*

For question 5 , 6 and 7 use the following information.

Suppose you are using Gibbs sampling to estimate the distributions, θ and β 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)
Doc2: 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. $\eta = 0.3$ and $\alpha = 0.3$

For question 5,6,7 calculate the value upto 4 decimal points and choose your answer

Question 5 :

Using the above structure the estimated value of $\beta(2)_{nature}$ at this point is

1. 0.0240
2. 0.02459
3. 0.0260
4. 0.0234

Sigma[k=1 to 5](Cwt) = (4+0+4+0+3=11) ---> Sum of 2 Topic Columns

C WT nature,2 = 0

W = 5 (unique word counts)

Answer: 1

beta2nature = (0 + 0.3) / (11 + (5*0.3)) = 0.3/12.5 = 0.024

Solution:

	t1	t2
machine	0	4
nature	5	0
language	5	4
vision	3	0
learning	0	3

beta measures how likely a word is under a topic.

Topic 2 has zero occurrences of "nature".

But Dirichlet prior eta = 0.3 smooths it, avoiding zero probability.

Total topic-2 word assignments = 11.

Final smoothed probability = 0.3 / (11 + 1.5) = 0.024.

$\beta(2)_{nature}$ = (0+0.3)/(11+5*0.3) = 0.3/12.5 = 0.024

Question 6 :

Using the above structure the estimated value of θ_{t1}^{doc2}

1. 0.6562
2. 0.6162
3. 0.6385
4. 0.50000

alpha=0.3

T = 2

C[DT doc2,t1] = 5
C[DT doc2,t2] = 3
Total topic assignment in doc2 = 8

Answer: 2

theta[t1, doc2] = [5+0.3] / [(5+3) + 2*0.3] = 5.3/8.6 = 0.6162

Solution:

	t1	t2
doc1	8	0
doc2	5	3
doc3	0	8

θ_{t1}^{doc2} = (5+0.3)/(8+2*0.3) = 5.3/ 8.6 = 0.6162

Question 7 :

Using the above structure the estimated value of θ_{t2}^{doc2}

1. 0.6562
2. 0.3975
3. 0.3837
4. 0.3707

Answer: 3

Solution:
Use the same formulae mentioned in Question 9 solution

Topic 1 Topic 2
Doc1 8 0
Doc2 5 3
Doc3 0 8

theta[t2,doc2] = (3+0.3) / (8 + (2*0.3)) = 3.3/8.6