Week 9 - NLP

Question 1:

Which of the following is/are true?

1. Topic modelling discovers the hidden themes that pervade the collection
2. Topic modelling is a generative model
3. Dirichlet hyperparameter Beta used to represent document-topic Density?
4. None of the above

Multi-choice multi correct.
3rd statement is false.
1st statement is true.
2nd statement is true.
Options 1) and 2) are correct.

Question 2:

Which of the following is/are true?

1. The Dirichlet is an exponential family distribution on the simplex positive and negative vectors sum to one
2. Correlated Topic Model (CTM) predicts better via correlated topics
3. LDA provides better fit than CTM
4. CTM draws topic distributions from a logistic normal

Multi-choice multi correct
Statement 1 is false.
Statement 2 is true.
Statement 3 is false.
Statement 4 is True
Options 2 and 4

Question 3:

You have a topic model with the parameters $\alpha = 0.89$ and $\beta = 0.04$. 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

Sparser ⇒ [0.8 0.1 0.1]
Denser ⇒ [0.33 0.33 0.34]
Req: sparser dist. Over words and denser dist. Over topics.
Alpha => topics, Beta ⇒ words.
Alpha value should be increased and Beta value should be decreased.
Option 4 should be correct.

Question 4:

Which of the following is/are false about LDA assumption?

1. LDA assumes that the order of documents matter
2. LDA is not appropriate for corpora that spans hundreds of years
3. LDA assumes that documents are a mixture of topics and topics are a mixture of words
4. LDA can decide on the number of topics by itself.

Page 76: topics are collection of words => Bag-of-word analogy where order of words doesn't matter.
and documents are collection of these topics.
Statement 2, 3 is true
Statement 1,4 is false.
Option 1, 4 should be correct.

Question 5:

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

1. Unsupervised.
2. Supervised.
3. Semi-automated.
4. None of the above.

Option: unsupervised. Option 1) is true.

Question 6:

Which of the following is/are False about Gibbs Sampling?

1. Gibbs sampling is a form of Markov chain Monte Carlo (MCMC)
2. Sampling is done sequentially and proceeds until the sampled values approximate the target distribution
3. It can not estimate the posterior distribution directly
4. Gibbs sampling falls under the category of variational methods

Multi-choice multi correct:
Statement 1 is true.
Statement 2 is true.
Statement 3 is false.
Statement 4 is false.

Option 3) and 4)

Question 7:

For question 8 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 8 calculate the value upto 4 decimal points and choose your answer

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

Option 1) is correct.

Question 8:

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

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

Option 2) is correct.