# Natural Language Processing

## Assignment- 9

### TYPE OF QUESTION: MCQ

**Number of questions**: 10 **Total mark: 10 X 1 = 10**

---

**Question 1. Vikram has lots of documents and he wants to model the content as well as connections.**

**Which topic modelling technique will be suitable for it?**

1. Correlated Topic Model

2. Relational Topic Model

3. Dynamic Topic Model

4. Supervised Latent Dirichlet Allocation

**Answer: 2**

**Solution:**

---

**Question 2: In Topic modeling which hyperparameters tuning used for represents document-topic Density?**

1. Dirichlet hyperparameter Beta
2. Dirichlet hyperparameter alpha
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 = 0.9 and β = 0.05. 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:  How does Correlated Topic Model create relations among topics?**

1. By having lots of general words inside the topics
2. By Removing stop-words
3. By using logistic normal distribution
4. None of the above

**Answer: 3**
Solution:

---

**Question 5: 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.

1. (I).
2. (II).
3. III).
4. All of the above.

**Answer - 4.**

**Solution:**
All of the above

**Question 6 :**

**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 $z$

**Question 7 : Which of the following is/ are true ?**

1. Dirichlet distribution is a family of exponential distribution
2. LDA is impacted by the order of documents
3. In LDA the number of latent clusters are identified automatically
4. All of the above are true

**Answer: 1**

**Solution:**
The order of documents does not matter in LDA, we need to identify the number of latent clusters in advance in the LDA topic model.

**For question 8 , 9 and 10 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. η = 0.3 and α = 0.3

$$\beta_i^{(j)} = \frac{C_{ij}^{WT}+\eta}{\sum_{k=1}^{W} C_{kj}^{WT}+W\eta} \qquad \theta_j^{(d)} = \frac{C_{dj}^{DT}+\alpha}{\sum_{k=1}^{T} C_{dk}^{DT}+T\alpha}$$

For question 8,9,10 calculate the value upto 4 decimal points and choose your answer

**Question 8 : Using the above structure the estimated value of β(2)nature at this point is**

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

**Answer: 1**

**Solution:**

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

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

**Question 9 : Using the above structure the estimated value of $\theta_{t1}^{doc2}$**

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

**Answer: 2**

**Solution:**

| | t1 | t2 |
|---|---|---|
| doc1 | 8 | 0 |
| doc2 | 5 | 3 |
| doc3 | 0 | 8 |

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

**Question 10 : Using the above structure the estimated value of $\theta_{t2}^{doc2}$**

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

**Answer: 3**

**Solution:**
Use the same formulae mentioned in Question 9 solution