

Quiz 1:

Section A:

1. Why is tokenization in Japanese difficult?
  - a. It has no spaces to separate out different words. Also multiple variants of the language can be in the same sentence.
  - b. Because it has a lot of morphology.
  - c. Because the word order is very different from English
  - d. None of the above

Ans: A

2. If I tried to do max match segmentation with vocab as follows {ABC, ABCDE, DEF, GH, FGH, AB, CDEF}, what will be the output for the string "ABCDEFGH"?
  - a. ABC, DEF, GH
  - b. AB, CDEF, GH
  - c. ABC, DE, FGH
  - d. ABCDE, FGH

Ans: D

3. Stemming is not good for which of these applications?
  - a. Web search
  - b. Machine translation
  - c. It is not good for both
  - d. It is good for both

Ans: B

4. Porter's algorithm is what kind of algorithm?
  - a. Machine learning based
  - b. Deep learning based
  - c. Rule-based
  - d. Dynamic programming

Ans: C

5. What kind of queries cannot be handled by inverted index?
  - a. Queries with entities, e.g., Gandhi age
  - b. Queries with presence and absence of words like "brutus" and "Calpurnia" but not "caesar"
  - c. Phrase queries like "Indian School of Business"
  - d. Inverted indexes can handle all of these types of queries.

Ans: C

6. Term-document incidence matrix has
  - a. Terms as rows, Documents as columns, each cell is 0 if term is not present in document, else 1.
  - b. Terms as rows, Topics as columns, each cell is 0 if term is not related to topic, else 1.
  - c. Topics as rows, Documents as columns, each cell is 0 if document is not related to topic, else 1.
  - d. Terms as rows, Documents as columns, each cell is TFIDF value given the term and the document.

Ans: A

7. If a word occurs in 1% of documents, what will be its IDF?

- a. 1
- b. 2
- c. 3
- d. 4

Ans: B

8. Consider a collection of three documents: ABACDE, ABCDBEF, AGGHI. What is the IDF for each word?

- a. A: 0, B:  $\log 2/3$ , C:  $\log 2/3$ , D:  $\log 2/3$ , E:  $\log 2/3$ , F:  $\log 1/3$ , G:  $\log 1/3$ , H:  $\log 1/3$ , I:  $\log 1/3$
- b. A: 0, B:  $3/2$ , C:  $3/2$ , D:  $3/2$ , E:  $3/2$ , F: 3, G: 3, H: 3, I: 3
- c. A: 0, B:  $\log 3/2$ , C:  $\log 3/2$ , D:  $\log 3/2$ , E:  $\log 3/2$ , F:  $\log 3$ , G:  $\log 3$ , H:  $\log 3$ , I:  $\log 3$
- d. A: 0, B:  $\log(3) * \log(3/2)$ , C:  $\log(2) * \log(3/2)$ , D:  $\log(2) * \log(3/2)$ , E: 0, F: 0, G:  $\log(2) * \log(3)$ , H: 0, I: 0

Ans: C

9. IDF is a function of

- a. Term only
- b. Term and collection
- c. Document only
- d. Term and document

Ans: B

10. For cosine similarity based relevance ranking, what vectors are used to represent each query and document?

- a. TFIDF
- b. TF
- c. IDF
- d. Word frequencies

Ans: A

## 2

Q-1: What is not an application of topic models?

- a. Discovering important factors discussed about a product on [Twitter](#)
- b. Discovering top types of complaints submitted to a customer care center.
- c. Studying evolution of aspects for a news story.
- d. Classifying an email as spam or not.

Ans: D

Q-2: When constructing a document, what distribution are words assumed to follow?

- a. Gaussian
- b. Binomial
- c. Multinomial
- d. Bernoulli

Ans: C

Q-1: What is not an application of topic models?

- a. Discovering important factors discussed about a product on Twitter
- b. Discovering top types of complaints submitted to a customer care center.
- c. Studying evolution of aspects for a news story.
- d. Classifying an email as spam or not.

Ans: D

Q-2: When constructing a document, what distribution are words assumed to follow?

- a. Gaussian
- b. Binomial
- c. Multinomial
- d. Bernoulli

Ans: C

Q-3: For the unigram model, how is the topic for each document decided?

- a. Unigram model allows a mixture of topics for every word in a document

Q-3: For the PLSA model, how is each word generated?

- a. Each document is generated by first choosing a topic  $z$  and then generating  $N$  words independently from the conditional multinomial
- b. The words of every document are drawn independently from a single multinomial distribution
- c. Select a doc with probability  $P(d)$ . Pick a latent topic  $z$  with probability  $P(z|d)$ . Generate a word  $w$  with probability  $P(w|z)$ .
- d. Select a doc with probability  $P(d)$ . Pick a latent topic  $z$  with probability  $P(z)$ . Generate a word  $w$  with probability  $P(w|d)$ .

Ans: C

Q-4: If  $V$ =#unique words,  $D$ =#documents,  $K$ =#topics, what is the number of parameters in the LDA model (with only alpha Dirichlet priors and no beta Dirichlet priors)?

- a.  $K+KV$
- b.  $D$
- c.  $K+V$
- d.  $KV$

Ans: A

Q-5: How many parameters does a Dirichlet distribution take?

- a. Same as the size of the sample
- b. Same as number of topics
- c. Same as number of documents
- d. Same as number of words

Ans: A

Q-6: SVD is the mathematical foundation for which of these models?

- a. LSI
- b. PLSA
- c. LDA
- d. Mixture of unigrams I

Ans: A

Q-7: Which distribution is also called as Discrete distribution?

- a. Gaussian
- b. Multinomial
- c. Binomial
- d. Dirichlet

Ans: A

Q-8: If  $V$ =#unique words,  $D$ =#documents,  $K$ =#topics, what are the parameters in the smoothed LDA model?

- a. Alpha and beta Dirichlet parameters
- b. Alpha Multinomial and Beta Dirichlet parameters
- c. Word-topic parameters, document-topic parameters
- d. Word-topic parameters, document-topic parameters, Alpha and beta Dirichlet parameters

Ans: A

Answer to 7 is B, not A

Q-9: If  $V$ =#unique words,  $D$ =#documents,  $K$ =#topics, in a non-smoothed LDA model, how many times do you sample from the Dirichlet distribution and what is the size of each sample?

- a.  $V, K$
- b.  $D, V$
- c.  $D, K$
- d.  $V, D$

Ans: C

Q-10: Which of these best describes the generative process for smoothed LDA model?

- a. Choose  $\theta \sim \text{Dirichlet}(\alpha)$ . For each of the  $N$  words  $w_n$ : (1) Choose a topic  $z_n \sim \text{Multinomial}(\theta)$  (2) Choose a word  $w_n$  from  $p(w_n|z_n, \beta)$ , a multinomial probability conditioned on the topic  $z_n$
- b. Choose  $\theta \sim \text{Dirichlet}(\beta)$ . For each of the  $N$  words  $w_n$ : (1) Choose a topic  $z_n \sim \text{Multinomial}(\theta)$  (2) Choose a word  $w_n$  from  $p(w_n|\beta)$ , a global multinomial probability.
- c. Choose  $\theta \sim \text{Multinomial}(\alpha)$ . For each of the  $N$  words  $w_n$ : Choose a word  $w_n$  from  $p(w_n|z_n)$ , a multinomial probability conditioned on the topic  $z_n$
- d. For each of the  $N$  words  $w_n$ : (1) Choose a topic  $z_n \sim P(d|z_n)$  (2) Choose a word  $w_n$  from  $p(w_n|z_n, \beta)$ , a multinomial probability conditioned on the topic  $z_n$

Ans: A

## Section A

Q-1: According to Bayes rule,  $c$  is the class label and  $d$  is the document  $P(c|d) \propto P(d|c)P(c)$

what is  $P(c|d)$  called?

- a. Prior probability
- b. Posterior probability
- c. Data likelihood
- d. None of these

Ans: B

Q-2: What is the conditional independence assumption in Naïve Bayes?

- a. Words are independent of each other
- b. Classes are independent of each other
- c. Words are independent of each other given the class label
- d. Words and classes are independent of each other

Ans: C

Q-3: Which of the following is not an example of a context pattern

- a. [ORGANIZATION]'s [JOBTITLE] [PERSON]
- b. sale of [ORGANIZATION]
- c. [PERSON] joined [ORGANIZATION]
- d. Cap. Word + {Street, Boulevard, Avenue, Crescent, Road}

Ans: D

Q-4: Which of the following are machine learning based models for named entity recognition?

- a. HMMs and CRFs
- b. List lookup based approach
- c. Internal evidence based method
- d. Patterns based method

Ans: A

Q-5: Which of these is not a supervised learning problem?

- a. Spam classification
- b. NER
- c. Topic modelling using LDA
- d. Classifying a period as End-of-sentence or not

Ans: C

Q-6: How is precision for a particular class defined?

- a. Correctly predicted in the class/total actual instances in the class
- b. Correctly predicted in the class/total predicted instances across all classes
- c. Correctly predicted in the class/total instances in the dataset
- d. Correctly predicted in the class/total predicted in the class

Ans: D

Q-7: 3 patients are selected for chemotherapy ; Rest are declared healthy! 1 year later ... 1 of them did not actually have cancer! What is the precision of the system?

- a. 0.67
- b. 0.40
- c. 0.96
- d. 0.33

Ans: A

Q-8: A system which needs to identify cancer-risk patients. Recall not 100% will imply that some patients will die of cancer.

- a. Precision
- b. Recall
- c. Either precision or recall
- d. None of these

Ans: B

Q-9: What do you learn as part of Naïve Bayes classifier training?

- a. Class Prior probabilities
- b. Class Prior probabilities and Class Posterior Probabilities
- c. Class Prior probabilities and conditional probabilities
- d. Joint probability of (word, class) pairs.

Ans: C

Q-10: Given a training dataset with equal number of 1000 positive and 2000 negative instances, let a test instance have 4 words. The probabilities of those 4 words are 0.5, 0.6, 0.7, 0.8 in positive class, and 0.8, 0.7, 0.6, 0.4 in negative class. What class will be predicted for that instance?

- a. Positive
- b. Negative
- c. Both positive and negative with equal probability
- d. Can't say

Ans: B