Quiz: https://www.menti.com/alfbi3uchk5k

1. Stemming of words
    a. All steps should be included
    b. Should not use any online stemming website (such as https://trevorfox.com/tools/stemmer/, https://textanalysisonline.com/nltk-porter-stemmer, etc.)
2. All the distance metrics such as Levenshtein, Hamming, etc.
3. Calculation of TF-IDF, count vectoriser, bag of words, term-document matrix, cosine similarity, jaccard similarity, etc.
4. Also advantages and disadvantages of these models, use-cases
5. Compute inverted index, merge algorithm etc.
6. N-grams, maximum likelihood, probability of the sentences
7. Understanding of language: -
    a. P(english|want) = .0011
    b. P(chinese|want) = .0065
    c. P(to|want) = .66
    d. P(eat | to) = .28
    e. P(food | to) = 0
    f. P(want | spend) = 0
8. Apply Smoothing technique on bi-grams and solve the following
    a. Corpus Sentences
        i. win money now
        ii. cheap products available
        iii. meeting scheduled for tomorrow
        iv. your report is due tomorrow
    b. Test Sentence
        i. cheap money report
9. Compute perplexity
10. Neural network models - AND, OR, XOR, etc.
11. Calculation of Computation graphs, Backward passes,
12. Word2vec

a. Conceptual questions on word Embeddings, for example,
    i. Generate document/paragraph embeddings from pretrained word2vec embeddings, what would be the dimensions of the embeddings, etc.
    ii. Identifying synonyms using word2vec
    iii. Evolution of meaning of words over the years
    iv. Identifying positive and negative reviews of movies
    v. Polysemy handling
b. Focus on understanding the steps of the model - why it is done what it is done
c. Advantages of each step
d. How negative sampling work? What is the trade-off in selecting the optimal number of negative samples? How might you modify the negative sampling process to ensure a better representation of rare words?
e. Comparison between word2vec, GloVe, FastText (details of GloVe and FastText not required, just the concept is enough)

13. BERT:
    a. Same as before, understanding conceptually is important
    b. Main difference between BERT and word2vec
    c. Cross-attention, multi-head attention, scaled dot product etc.
    d. Transformers
14. Graph analytics
    a. Computation of all the definitions
15. Graph embeddings
    a. Impact of negative samples in these models. How are they generated
    b. Loss functions
    c. Usage of p and q in node2vec
    d. Importance of using random walks in node2vec

Question 1. 16 Pts

(1.1) [4 Pts] Are the following sentences True or False? Justify your answer as precisely as possible. (max. 80 words)

(a) Stemming increases the size of the vocabulary. - **False**

(b) Stemming increases the recall in a boolean retrieval system. - **True**

(1.2) [8 Pts] Stem the following words using the ***Porter Stemming Algorithm*** [1] and show all the intermediate steps and the corresponding rules. The words are:

(a) University -**Univers**

(b) Caress -**Caress**

(c) Universe - **univers**

(d) Cars -**car**

(1.3) [2 Pts] Some of the aforementioned words in Question 1.2 would be reduced to the same stem word(s). In your opinion, are there any pairs of words that share the same stem word(s) that should not be treated as equivalent? Provide your reasoning. (max. 50 words) — **university and universe are not equivalent but will have the same stemmed word, hence stemming in such scenarios is not recommended**

(1.4) [2 Pts] After stemming using the Porter Stemming Algorithm the word "BOSS" remains unchanged. Which rule in the stemming algorithm prevents any changes in the word "BOSS". – $SS- > SS$

Question 2. 10 Pts

(2.1) [4 Pts] Compute Inverted Index representation of the following documents. Show all steps.

(a) **Document 1**: To be, or not to be, that is the question.

(b) **Document 2**: This above all: to thine own self be true.

(2.2) [4 Pts] Use the ***Merge Algorithm*** on the following posting lists and determine the final merged list. (Show all the intermediate steps)

(a) $A = [1, 3, 7, 9, 15]$

(b) $B = [2, 3, 8, 9, 14, 20]$

(2.3) [2 Pts] Consider the ***Merge Algorithm*** that combines two sorted posting lists, A and B, of lengths n and m respectively, into a single merged list C. Analyze the time complexity of the algorithm if every element of A is same as every element of B.

Question 3. 24 Pts

Given the following **training** sentences:

(a) The quick brown fox runs

(b) The lazy dog sleeps

(c) The fox sleeps on the grass

(d) The dog jumps over the fence

The **test sentence**: *The fox jumps.*
Compute the following considering the beginning and the end of sentence tokens:

(3.1) [8 Pts] Compute the Maximum Likelihood Estimates (MLE) for all the **bigrams** in the training set.

(3.2) [8 Pts] Compute Laplace smoothing for all the bigrams in the training set.

(3.3) [2 Pts] Using the MLE, compute the probability of the test sentence.

(3.4) [2 Pts]Using the Laplace smoothing, compute the probability of the test sentence.

(3.5) [2 Pts] What is the impact of Laplace smoothing on the probability of the test sentence?

(3.6) [2 Pts] Compute the MLE of the bigram (***the, runs***) and explain what this probability indicates about the relationship between the words ***the*** and ***runs***.

Question 5. 17 Pts

(5.1) You are working on a text analysis project involving document similarity for a dataset containing 30 documents. Each document consists of multiple sentences. You decide to generate document embeddings using pre-trained Word2Vec embeddings with the following specifications:

(a) **Word2Vec Model**: The pre-trained Word2Vec model (say, Google News) is used which provides word embeddings with a dimension of 300.

(b) **Dataset**: Each document in the dataset consists of 50 sentences, with each sentence containing approximately 10 words.

(c) **Embedding Methodology**: To compute the document embeddings, the mean of the word embeddings for all words in the document is calculated.

(5.1.1) [2 Pts] What will be the dimensions of the generated document embeddings when using the Mean Pooling approach?

(5.1.2) [3 Pts] After generating the document embeddings, you plan to measure the similarity between documents. Discuss the advantages and disadvantages of using Cosine Similarity versus Euclidean Distance for this task. (max. 80 words)

(5.1.3) [3 Pts] If a document contains words not present in the pre-trained Word2Vec vocabulary i.e. OOV words, how would this affect the resulting document embeddings? Propose a solution to handle out-of-vocabulary words. (max. 80 words)

(5.2) [3 Pts] What is one major drawback of the word2vec model besides OOV words and how can it be improved? Explain with an example (max. 80 words)

(5.3) [3 Pts] BERT has a maximum input length of 512 tokens. How would you handle documents longer than this limit in your search engine? Discuss one strategy. (max. 80 words)

(5.4) [3 Pts] Why is the dot product of Query and Key matrices scaled by the square root of their dimension in the attention mechanism? What would happen if this scaling factor were omitted? (max. 80 words)

Did not include Question 6, because different embedding models were in the syllabus last year!!!!