CS4248 Final Assessment

Your Name: _____ **Your ID:** _____
# of Questions: 25
Date and Time of Exam Creation: Mon, May 01, 2023 @ 21:19:01
Total Exam Points: 120.00

Note: All MCQ/MRQ/TF questions omitted

Question #: 19

[Calculation and Essay; 3 parts; **9 marks** in total] Measuring similarity between sentences, documents, and domains is an important task in NLP. In this problem, we want to calculate the similarity between two sentences.

	everything	is	hey	hi	how	going	up	what
d1								
d2								

Consider the following sentences from our sample data set:
d1 = "Hi, how is everything going?"
d2 = “Hey, what is up?”
Suppose after case-folding (to lowercase) and removing all punctuation marks, we have the vocabulary V ={everything, is, hey, hi, how, going, up, what}.
a) Fill the above Bag-Of-Words table with the words' frequency
(follow the format "*d1 | everything =0"* in answering this question). (**3 marks**)
b) Calculate the cosine similarity between d1 and d2. (**4 marks**)
c) Accoding to the cosine similarity score, are the two sentences similar or dissimilar? Justify your answer. A short 1-2 sentences or phrases is sufficient. (**2 marks**)

Your answer [*Do try to answer all three parts of the question; demarcate each part's answer clearly*]:

Item Weight: 9.0

Question #: 20

[Essay; 2 parts; **8 marks** in total] A popular model used for sentiment classification is an LSTM model: This model inputs word vectors to the LSTM model at each time step and uses the last hidden state vector to predict the sentiment label (y). Recall that we also used a simple “bag-of-vectors” model for sentiment classification: we used the average of all the word vectors in a sentence to predict the sentiment label.
a) Name at least one benefit that the LSTM model has over the bag-of-vectors model. (**3 marks**)
b) If we choose to update our word vectors when training an LSTM model on sentiment classification data, how would these word vectors differ from ones not updated during training? Explain with an example. Assume that the word vectors of the LSTM model were initialized using GloVe or word2vec. (**5 marks**)

Your answer [*Do try to answer all two parts of the question; demarcate each part's answer clearly*]:

Item Weight: 8.0

Question #: 21

[Essay; **9 marks**] We can use Hidden Markov Models (HMMs) to address Part-of-Speech Tagging, and the Viterbi Algorithm to reduce the cost of decoding.

Suppose the length of a text sequence is T, the number of possible tags is N. Then, the complexity of the Viterbi Algorithm is $O(TN^2)$.

a) Propose a way to reduce the complexity of the Viterbi Algorithm to $O(T \cdot k \cdot N)$, where $k < N$. [Hint: your method may return a suboptimal result; 2-3 sentences is sufficient for a description] (**6 marks**)

b) Can we find an optimal sequence of tagging for the input text, after reducing the complexity to $O(T \cdot k \cdot N)$? Why? (**3 marks**)

Your answer [*Do try to answer all two parts of the question; demarcate each part's answer clearly*]:

Item Weight: 9.0

Question #: 22

[Calculation; **8 marks**] Given a function f(x1, x2) =x_1^2 +x_2 +1, the initial value (x_1, x_2) =(1, 0).
a) State the derivative of the function; (**2 marks**)
b) Use the method of gradient descent, and calculate the value of (x_1, x_2) after two steps. The learning rate is 0.01. Show all work. (**6 marks**)

Your answer [*Do try to answer all two parts of the question; demarcate each part's answer clearly*]:

Item Weight: 8.0

Question #: 23

[Calculation and Essay; 4 parts; **18 marks** in total]
Greedy search in a generative decoder expands the path beyond the token with the highest conditional probability from the vocabulary V.
The beam search algorithm is an improved version of greedy search, parameterised with a beam size.
The beam size, b, allows beam search to expand the paths of the b tokens with the highest conditional probabilities at each time step.
a) What is the main drawback of greedy decoding? (an answer of 1-2 relevant sentences or phrases is sufficient) (**3 marks**)
b) What is the main drawback of beam search based decoding? (an answer of 1-2 relevant sentences or phrases is sufficient) (**3 marks**)

Consider the table below. We consider the probability between tokens.

y_{t-1}			y_t	
last	night	global	peace	<s >
0.3	0.02	0.1	0.001	last
0.01	0.3	0.1	0.02	night
0.001	0.1	0.03	0.02	global
0.001	0.003	0.001	0.5	peace

c) Given the empty start sequence <s >, decode a sequence of length 3 (without counting the <s >tokens) using greedy search. Show all calculations and probabilities. (**4 marks**)
d) Given the empty start sequence <s >, using a beam search of size 2, what is the decoded sequence of length 3 (again, without counting the <s >tokens)? What is the maximum probability of the sequence decoded using the strategy? Show all calculations and probabilities [You can indicate a probability of a sequence like P(w1, w2, w3) =XX]. (**8 marks**)

Your answer [*Do try to answer all four parts of the question; demarcate each part's answer clearly*]:

Item Weight: 18.0

Question #: 24

[Essay; **10 marks**] A technology vendor, Open Elgoog, is creating an NLP system to score job applicant's résumés for pre-screening for their global clients' human resources (HR) departments.
For privacy's sake, their clients do not give any training data of their historical hires to Open Elgoog. They have also heard about other cases where using historical hiring to train system perpetuates improper practices.
Open Elgoog thus decides crawl the internet to train word embeddings unrelated to job applicants to be unbiased. For the sake of time, they train a small word embedding model instead from only wine review sites, as such review sites have words related to sentiment. In deployment, they take a job applicant's résumé to calculate an average embedding and score it relative to the distance between seed words synonymous with "poor" and ones synonymous with "excellent".
Identify any issues with Open Elgoog's system. Explain your answer in detail.

Answer:

Item Weight: 10.0

Question #: 25

[Essay; **7 marks**]
The transformer architecture uses Multi-Headed Attention as a key feature of its architecture.
Describe what would happen if a transfomer implementation sets them number of attention heads to 1; i.e., single-attention head.
For full credit, give a detailed explanation.

Answer:

Item Weight: 7.0