M1522.006800 Introduction to Natural Language Processing

Fall 2022 Midterm Practice Exam

Question 1

Mark the following statements as true or false.

a) N-gram models do not properly represent long-range dependency.

b) LSTMs improve credit assignment compared to standard RNNs.

c) Better language models exhibit higher perplexity.

d) A high accuracy can be seen when we evaluate a classifier whose output is constant.

e) Euclidean distance captures semantic similarity between documents well.

f) Exact match queries in web search often produce either too few or too many results.

Question 2

Given the five sentences below, calculate the following bigram probabilities. Assume that <s> is prepended to the beginning of each sentence, and </s> is appended to the end of each sentence.

"you do not like them"
"so you say"
"try them"
"try them"
"and you may"

a) $P(\text{say} \mid \text{you})$

b) $P(\text{and} \mid \text{<s>})$

c) $P(\text{them} \mid \text{try})$

d) $P(\text{try} \mid \text{<s>})$

e) $P(\text{</s>} \mid \text{them})$

Question 3

Below are tables of bigram and unigram counts.

		w_n						
		I	love	Italian	food	my	favorite	hobby
w_{n-1}	I	41	2452	0	0	0	0	0
	love	4	9	40	21	80	0	1
	Italian	0	0	0	55	0	0	0
	food	25	0	0	0	2	0	0
	my	1	12	0	7	0	839	12
	favorite	7	0	0	132	4	0	5
	hobby	5	0	0	0	0	0	0

I	love	Italian	food	my	favorite	hobby
32498	2943	73	486	5117	1310	105

a) Fill in the bigram probability table below. (Round to 4 decimal places.)

		w_n						
		I	love	Italian	food	my	favorite	hobby
w_{n-1}	I					0	0	0
	love					0.0272	0	0.0003
	Italian					0	0	0
	food					0.0041	0	0
	my	0.0002	0.0023	0	0.0014	0	0.1640	0.0023
	favorite	0.0053	0	0	0.1008	0.0031	0	0.0038
	hobby	0.0476	0	0	0	0	0	0

b) Fill in the bigram probability table below with add-1 smoothing. Assume that the size of the vocabulary is 8453. (Round to 4 decimal places.)

		w_n						
		I	love	Italian	food	my	favorite	hobby
w_{n-1}	I					0.0000	0.0000	0.0000
	love					0.0071	0.0001	0.0002
	Italian					0.0001	0.0001	0.0001
	food					0.0003	0.0001	0.0001
	my	0.0001	0.0010	0.0001	0.0006	0.0001	0.0619	0.0010
	favorite	0.0008	0.0001	0.0001	0.0136	0.0005	0.0001	0.0006
	hobby	0.0007	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

Question 4

Answer the following questions.

a) Briefly explain the training objectives (tasks) of CBOW and skip-gram, respectively.

b) What is the role of softmax in these methods?

Question 5

We have a dataset with three class labels, namely Urgent, Normal, and Spam. The following is the confusion matrix for these classes.

	True Class			
Predicted Class		Urgent	Normal	Spam
	Urgent	7	8	9
	Normal	1	2	3
	Spam	3	2	1

a) Complete the tables below.

Class: Urgent		
	True Urgent	True not
System Urgent		
System not		

Class: Normal		
	True Normal	True not
System Normal		
System not		

Class: Spam		
	True Spam	True not
System Spam		
System not		

b) Compute the precision, recall and F1-score for each class and complete the table below.

Class	Precision	Recall	F1-score
Dog			
Cat			
Squirrel			