Brac University
Department of Computer Science and Engineering
CSE431: Natural Language Processing
Examination: Final, Duration: 3 hours
No. of Questions: 9, Semester: Fall 2020
Full Marks: 50, No. of Pages: 1

Answer all questions. Figures in the right margin indicate respective marks. Please keep the answers short.

ID: _____

Name in UPPERCASE: _____

Email address for viva notification: _____

Section A

Question 1) Discuss how naive Bayes algorithm handles **unknown words** and **stop words**? [10 points]

Question 2) Mention 5 usages of text categorization except sentiment analysis [5 points]

Using **Laplace smoothing**, the following example demonstrates training and testing naive Bayes. Imagine a sentiment analysis domain with the two classes positive (+) and negative (-), and take the following mini training and test documents.

	Category	Document	Words	Words in documents in the same category	Documents in the same category
Training	-	just plain boring	3	14	3
	-	entirely predictable ***and*** lacks energy	5		
	-	**no surprises** and ***very*** **few laughs**	6 (5 new)		
	+	very **powerful**	2 (1 new)	9	2
	+	***the*** **most fun film of** the **summer**	7 (6 new)		
Test	?	predictable with no fun	3 (not 4)		

N_c = number of documents in training data belonging to the class c

N_{doc} = total number of training documents

Using P(c) = $\frac{N_c}{N_{doc}}$, we get prior probability of positive & negative sentiment classes as $P(-) = \frac{3}{5}$ $P(+) = \frac{2}{5}$

Vocabulary size, |V| = 3+5+5+1+6 = 20 words (just, plain, boring, entirely, predictable, and, lacks, energy, no, surprises, very, few, laughs, powerful, the, most, fun, film, of, summer)

Using $\hat{P}(w_i|c) = \frac{count(w_i,c)+1}{\sum_{w\in V}(count(w,c)+1)} = \frac{count(w_i,c)+1}{(\sum_{w\in V} count(w,c))+|V|}$ the likelihoods of the three words "predictable", "no", and "fun" of being positive/negative are:

$$P(\text{"predictable"}|-) = \frac{1+1}{14+20} \quad P(\text{"predictable"}|+) = \frac{0+1}{9+20}$$
$$P(\text{"no"}|-) = \frac{1+1}{14+20} \quad P(\text{"no"}|+) = \frac{0+1}{9+20}$$
$$P(\text{"fun"}|-) = \frac{0+1}{14+20} \quad P(\text{"fun"}|+) = \frac{1+1}{9+20}$$

For the test sentence,

Denominator for negative classes = 14+20 = 34

Denominator for positive classes = 9 + 20 = 29

$$P(-)P(S|-) = \frac{3}{5} \times \frac{2\times2\times1}{34^3} = 6.1\times10^{-5}$$
$$P(+)P(S|+) = \frac{2}{5} \times \frac{1\times1\times2}{29^3} = 3.2\times10^{-5}$$

As 6.1 > 3.2, the model predicts the class "negative" for the test sentence, S = "predictable with no fun".

Question 3) Design 2 training documents and 1 test document each with 3 to 5 words. [5 points]

Based on your answer of Question 3, answer Questions 4 to 6.

Question 4) Calculate prior probabilities. [3 points]

Question 5) Calculate vocabulary size. [2 points]

Question 6) Predict sentiment class for the test document. [10 points]

Section B

Question 7) Design 3 documents about Chatbots & Dialogue Systems each with 3 to 5 words. [5 points]

Question 8) Calculate Term-Document Matrix for 2 words based on your answer of question 7 [5 points]

Question 9) Show the spatial visualization of the document vectors for your designed documents in Question 7 showing two of the dimensions, corresponding to any two words. [5 points]