Department of Data Science
IIT Palakkad
CS5624 : Natural Language Processing

0800-0850 **Quiz 1 (21 Feb 2022)** **Marks : 15**

Instructions

1. Write your answers neatly in Blue/ Black ink. Do not use pencil / Red ink. If your answer is not legible, you will not get any marks for that.

2. Doubts and questions will not be answered during the exam. If you have to make any assumption about unspecified things, write the assumption clearly with justification.

3. Answer all parts of a question together. If the parts of a single question are not together, then only the first part will be evaluated. Other parts will not get any marks.

4. Write your name and ID number at the top of the answer sheet. Save the pdf with the following naming convention: [roll_no]_nlp22_quiz_1.pdf and upload to the designated assignment in LMS. Do not email.

5. Write question number clearly for each answer. Draw a line after the answer.

6. No hard or soft material are permitted for consultation during the exam.

7. There will be partial markings for the questions, so even if you are not able to solve the entire problem be sincere with the steps.

8. **Be precise**.

1. Let us assume that you are given the task of creating a dictionary of words in your mother tongue. Each entry in the dictionary should be a word in your mother tongue and its meaning in English. (3)
   How will you approach this task? What will be the key challenges?

2. What are the disadvantages of ambiguity in a human language? Are there any advantages of ambiguity? (3)

3. What is the importance of smoothing in training language models? What are the key limitations of the smoothing techniques we discussed in the class? (3)

4. Please solve the problem "Z's Law" on the next page. The problem is authored by Prof. Dragomir Radev. The problem has been reproduced with the author's permission. (6)

Z's Law

Dr. Z, a field linguist, was studying the word frequencies in a newly discovered language. She counted the number of occurrences in a text of the 15 most frequent words in that language (shown in alphabetical order below) and wrote them on a piece of paper. However, she poured some tea on her paper and, as a result, two of the numbers were damaged. For one of them, a single digit is no longer legible. The other number is completely unreadable. The only thing which she remembers for certain is that *kumun* is the most frequently used word in the language. Can you please help the linguist recover the original numbers?

Word	**Frequency**
domin	6749
dotem	8998
dun	3001
ga	4503
grimun	2697
grumid	2075
kugrum	1801
kumun	27005
letun	3374
mat	2249
mig	2454
led	**?**854
mugun	**????**
mulunt	1930
munt	13497