# Department of Data Science
# IIT Palakkad
## CS5624 : Natural Language Processing

**11:00-11:40** **Quiz 3 (24 Mar 2022)** **Marks : 10**

**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_3.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**.

9. **There are total 4 questions in this question paper**

1. What is the need of applying Bayes' theorem in the naive Bayes text classification approach? (2)

2. F-measure is the harmonic mean of precision and recall. What will be the effects of using the arithmetic mean instead of the harmonic mean? (2)

3. Consider the following text classification tasks. Each classification task has two classes (*C1* vs *C2*) and the goal is to classify a document as either *C1* or *C2*: (3)

    1. *politics* vs *religion*
    2. *politics* vs *sports*
    3. *hardware* vs *software*
    4. *mathematics* vs *biology*
    5. *algebra* vs *geometry*

    Let us assume that we train a naive Bayes classifier for each of the classification tasks with exactly the same experimental settings, such as the size of training and test datasets, no class imbalance, priors/smoothing mechanism, etc.
    Considering these facts and the semantics of class labels, which classification tasks will have relatively high performance and which tasks will have relatively low performance in terms of F1. Why?

4. One approach to handle negations in sentiment classification of sentence $S$ is as follows: (3)

    1. Remove all negation words such as *no, not, never, couldn't, didn't*, **etc.** in sentence $S$. Let this new sentence be $S'$
    2. Use any existing sentiment classification algorithm to identify the sentiment of $S'$
    3. If the original sentence $S$ contains at least one negation word then reverse the sentiment identified in the previous step.

    Example

    - $S$ : *The movie is not worth watching*
    - $S'$ : *The movie is worth watching*
    - Sentiment of $S'$ : **Positive**
    - Sentiment of $S$ : **Negative** (because $S$ contains *not*)

    What are the limitations of this approach?
    Write one sentence with positive sentiment and negative sentiment each that will *fool* this approach.