# NLP, Mid-Semester Exam 2018

1. Answer **yes / No** with correct reason. +1 for both correct answer and correct justification and negative marking 0.5 for any one being wrong. Write max 2 line justification and to the points. [ 5 marks ]

    I. Accuracy alone is a good metric for model evaluation ?
    II. Naive Bayes classifier performs relatively better than other classifiers for text classification problem if training corpus is very small ?
    III. Perplexity is a function for only language model ?
    IV. Lemmatization keeps the vocabulary size same?
    V. Orthographic rules can distinguish goose vs geese ?

2. Given corpus: “They like icecream.They like pizza too.” [2*1 = 2 marks]
    **Calculate the following bigram probabilities with add-one smoothing.**
    A. P(icecream|love) [bigram model]
    B. P(icecream|like) [bigram model]

3. Calculate perplexity of given sentence, using both “bigram model and stupid back-off”: [4 marks]
    “<s> Sachin is a great guy </s>”

    **Corpus is :**
    <s> Everyone likes Sachin </s>
    <s> They regard him as a great cricketer and a humble guy</s>
    <s> He is a national asset </s>

4. (a) What relationship among the extracted features must be maintained to apply Naive Bayes Classifier for any classification problem. (1 line answer) [ 1+ 1 ]
    (b) Can you express your answer mathematically?

5. [ 2 marks ]

For each (language, regular expression) pair does the regular expression correctly cover all strings in the language. In case the regular expression is false for the language, report an example of string that is a false accept / false reject and give the right solution.

Alphabet set {0,1} (for all languages)

(a) Language: Strings which are all 0s or all 1s
Regular Expression: (0|1)*
(b) Language: Strings with even number of zeros
Regular Expression: (1*01*0)*

6.

Match the following

[2 marks]

| | |
|---|---|
| 1. Inflections | a. appoint, appointee |
| 2. Derivations | b. I will, I'll |
| 3. Compounding | c. eat, ate |
| 4. Cliticization | d. Doghouse |

7. [ 5 marks ]

A diagnostic test has a probability 0.95 of giving a positive result when applied to a person suffering from a certain disease, and a probability of 0.10 of giving a (false) positive when applied to a non sufferer. It is estimated that 0.5% of the population are sufferer. Suppose that the test is now administered to a person about whom we have no relevant information relating to the disease. Calculate the following probabilities.

(a) The test result will be positive.
(b) Given a positive result the person will be sufferer
(c) Given a negative result the person will be a sufferer
(d) Probability that the person will be misclassified.

8. Your friend Puja is a journalist and she covers political news. One day she finds one of her political news repository got mixed with sports news. Considering the fact that manually separating the news is impossible she seeks for your help as you told her earlier that you took NLP course & using NLP similar kind problem can be solved easily!! But you have only learned Naive Bayes Classifier as of now.

So you build a Naive Bayes Classifier model using the following corpus.

| **News** | **Category** |
|---|---|
| A great game | sports |
| The election was over | politics |
| Very clean match | sports |
| A clean but forgettable game | sports |
| It was a close election | politics |

[ 3 marks]

Now, to check if your solution really works, she asks you to tell her in which category the following news should belong to :

“**a very close game**”

What should be your reply to her? (You must show the correct individual probability calculation for classification) (Use total vocabulary size as 14)

9. Being impressed with your solution Puja gives you another challenge. [3 marks ]
She has lots of facebook friends but she hardly interacts with all of them. So she decides to create a new facebook profile with only her close friends with whom she interacts and deletes the previous account. She ask if you could do that using NLP knowledge.
You created a binary classifier which can predict whether someone is going be her friend in new fb account or not.
Before delivering the model you wanted to check the model performance and you test on following validation sets.

**Test set 1**
**Actual Output should be :**
Continue friendship in New Account [ Akash, Nisha, Karan, Amrita, Saikat, Sandip ]
No more friendship in new Account [ Prakash, Ankit, Rubel, Lahari, Namita ]
**Your Model Prediction:**
Continue friendship in New Account [ Prakash, Nisha, Amrita, Saikat, Namita]
No more friendship in new Account [ Akash, Ankit, Rubel, Lahari, Karan, Sandip]

How do you measure the accuracy of the model in a more specific manner to know detailed accuracy about different classes. Remember you have to report the accuracy for the entire