**TEXT MINING – EXAM**

**Total Marks : 30**

**2 Marks Question : 4 \*2 = 8 Marks**

1. What is the difference between Stemming and Lemmatization?
2. What is significance of TF-IDF?
3. Explain dependency parsing in NLP?
4. Which is better to use while extracting features character n-grams or word n-grams? Why?

**3 Marks Question : 5 \*3 = 15 Marks**

1. Write a function to calculate the word length of each sentence(3marks)
2. Write a code to remove rarely occurring words from the text (3 marks)
3. Explain the steps involved in web scrapping.
4. Many mobile applications provide a way to screen incoming SMS as spam. With 95% accuracy the company is able to classify sms as spam or ham. And there is always space for improvement. Your task is to use any of the machine learning models to improve the accuracy of the model to above 96%
   1. Use encoding=’cp1252’ while reading the data (spam.csv) using pandas, since this data set contains other languages
   2. Make sure that you split the data in to train (70%) and test (30%). Use random\_state=100, while using train\_test\_split function to set the seed
   3. Print the following metrics
      1. Accuracy, sensitivity, specificity, F1 score
5. Let’s take a corpus

All my penguins in a row’,  
‘When my penguin sits down, it looks like a Furby toy!’

Create a bag of words code for this corpus which includes feature vector representation.

**7 Marks : 1 \* 7 = 7 marks**

1. Using Yelp reviews data set (<https://bit.ly/2U7mLmC>), create a supervised model to predict the sentiment of user reviews.
   1. Create a word cloud using the user reviews
   2. Use random\_state=100 while splitting data in to training (80%) and testing(20%).
   3. Comment if the target variable contains balanced or imbalanced classes (In target variable: 0 = negative, 1 = positive sentiment)
   4. Build various supervised classification models using the training data set
   5. Predict and compute accuracy of the model on test data set
   6. Identify which model is best for sentiment prediction

Identify sentiment for the yelp reviews using vader package. Comment the performance of supervised vs unsupervised sentiment prediction.