MAI372– Natural Language Processing

III MSc AIM

CIA Component-II

15-03-2024 (2-4 pm)

1. **Implement the Pre-processing steps in the corpus (given below).**

Transformer is an exceptional innovation in the field of Deep Learning, contributed by Ashish Vaswani et al. (2017), Google. The transformer is the most influential Neural Network model that has shown outstanding performance on various NLP tasks including Machine Reading Comprehension, Machine translation and sentence classification. Attention mechanism and parallelization are the prominent features in the transformers. Consequently, it can facilitate long-range dependencies without any gradient vanishing or gradient explosion problems and it overcomes the drawbacks of the existing methods such as RNN and LSTM. The transformer is executed with an encoder-decoder mechanism and the original article of transformers # "Attention All You Need".

a. Word and Sentence Tokenization (2 marks)
b. Stopwords removal (2 marks)
c. Punctuation removal (2 marks)
d. Find the Frequency Distribution after (Steps B and C) and visualize the same. (2 marks)
e. Stemming (use any 2 stemmers) and lemmatization (3 marks)
f. Tag the Words using PoS (2 marks)
g. Implement the Named Entity Recognition (2 marks)

2. Implement the Word2Vec on **"Mickey Mouse, a cheerful and optimistic mouse clad in red shorts and yellow shoes, is the iconic mascot of The Walt Disney Company. Debuting in 1928, this spunky character has charmed audiences for generations with his adventures and can-do attitude."** (4 marks)

a. Visualize the vectors in a plane. (3 marks)
b. write (type) the working and flow of the code. (3 marks)

Program Evaluation Rubrics

Evaluation Rubrics

1. Implementation
   a. Section – 15 Marks
   b. Section – 10 Marks
2. Concept Clarity (VIVA)– 5 Marks

General Instructions

1. The file you have to save with your name, last 3 digits of register number and program number "Aaron_201_CIA II".
2. The implemented code you have to download as pdf format and upload in the Google Class room in the given scheduled time. (Should not copy paste and convert).