CISC/CMPE Advanced Data Analytics W22
Exercise 3

March 22, 2022

1 Introduction

The goal of this assignment is to get familiar with textual data analysis.
Submission: you need to submit your analysis as an executable R Markdown file or Python Jupyter Notebook file.

2 Data Reprocessing (30 points)

The given dataset is a table containing questions about R on StackOverflow site. Your first task is to perform standard text prepossessing steps introduced in lectures for future tasks. You can perform analysis on title or body of the questions.

3 Learning word vectors from text corpus (30 points)

Using existing libraries, such as gensim `https://radimrehurek.com/gensim/` to learn word embeddings from the prepossessed text from previous step. At the end of this step, you should save the learned word embeddings in a file.

4 Topic Modeling (30 points)

Perform topic analysis on the prepossessed textual data. Briefly specify how you pick the number of topics. Ref: `https://towardsdatascience.com/topic-modelling-in-python-with-nltk-and-gensim-4ef03213cd21`

5 Summary (10 points)

Present your findings from the topic modeling process.