Module Code: CS3TM20

Assignment report Title: Text Mining Coursework

Student Number (e.g. 25098635): 27011203

Date (when the work completed): 08/04/2020

Actual hrs spent for the assignment: 30

Assignment evaluation (3 key points):

* A point
* Another point
* A third point

**Content outline**

**1. Task #1: NLP analysis [65 marks]**

**2. Task #2: Prediction Challenge [35 marks]**

**References and citations [10 marks]Task 1 – NLP Analysis**

**Baseline Model:**

**Morphological Model:**

Lemmatization and Stemming – Coded

Lemmatization and Stemming – Results

Lemmatization and Stemming – Write Up

**Lexicon Model:**

TF-IDF Weighted Unigram Bag of Words Model 🡪 What is the computational meaning of individual words in context

Word2Vec as an alternative to TF-IDF

**Syntax Model:**

NLTK.POS tagging

Constituency parsing, chunking

**Semantic Model:**

Word Embedding ( Follow tutorial)

Dataset: sentiment analysis dataset (dataset.zip, on blackboard). It includes a training

set (train.csv), a development set (dev.csv), and a test set (test.csv). Each tweet has a

sentiment label (Positive, Negative, Neutral). In the test.csv file, the correct sentiment

label of each tweet has been replaced with a symbol “?”.

* Apply NLP analysis methods of each linguistic level:
* Morphology (Lemmatization, review different types of stemmer see which is better)
  + Formed of morphemes.
  + Two types of morphemes, lexical and grammatical.
  + Stemming is the simplest form of morphological processing.
  + Stems are the base of an English word, which can be surrounded by secondary morphemes called affixes.
  + Stemming involves reducing a word to its lemma.
  + http://www.nltk.org/howto/tokenize.html
* Lexicon (Lexical analysis is dividing the whole chunk of text into paragraphs, sentences and words)
* Syntax ( Grammatical structure of sentences, helps to identify sentence structure and relationship between entities)
* Semantics ( <http://www.nltk.org/howto/semantics.html> )
* To process the input text and extract features (positive, negative, neutral)?
* Then use Logistic Regression as the classifier ( <https://www.nltk.org/api/nltk.classify.html> )
* Discuss results and effectiveness of each method for **sentiment prediction** based on the development set
* Give references and citations to the model you used
* Use td-idf weighted uni-gram bag-of-words model as baseline model

{ Text normalization:

* Tokenizing (segmenting) words 🡪 Splitting into words
* Normalizing word formats 🡪 Case folding (all to lowercase? Case may be useful for sentiment, stemming)
* Segmenting sentences