December 14, 2022

Assignment 3

Course: Data Mining (IT734A)
Estimated time needed to complete the entire lab: 4-10 hours.

Instructor:
[redacted]

Introduction

In this assignment, we will learn neural networks, text mining and NLP. You are allowed to use packages providing basic mathematical operations, plotting, and algorithms, such as *NumPy*, *Pandas*, *Matplotlib*, *Plotly, sci-kit-learn, Spacy, and Tensorflow*. Besides the code for your solution, you should also write a small report (or Markdown) documenting your solution. Your code should be compressed into a zip archive, and your report should be handed in as a pdf file. Your files (zip, report) should be named accordingly as “DM22-C3-Code-ID.zip” and “DM22-C3-Report-ID.pdf”, where the ID is your Canvas user name.

Assessment

The assignment is graded with U and G. In order to pass the assignment, all individual tasks of the assignment ***must be passed***. Observe that this is an individual assignment. This means that everything that you submit for grading must be created by you. Plagiarism is not allowed in any form, but you may, of course, discuss concepts with peers.

Important links:

- Google Colab link : https://colab.research.google.com/
- Kaggke Link : https://www.kaggle.com/
- Scikit-learn: https://scikit-learn.org/
- Plotly : https://plotly.com/python/
- Spacy: https://spacy.io/
- Tensorflow: https://www.tensorflow.org/learn

Good Luck!

Dataset: link

Task 1: Text Mining & NLP

Sentiment analysis is a simple form of natural language processing in which we try to determine if a piece of text is positive, negative, or neutral. There are many techniques for performing sentiment analysis. In this assignment, we will use an approach based on TF-IDF, BOW, and word embeddings.

You will use the ***sentiment140(training)*** dataset. It contains 1,600,000 tweets extracted using the Twitter API. The tweets have been annotated (0 = negative, 4 = positive), and they can be used to detect sentiment. *Sentiment140* contains the following 6 fields:

1. target: the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
2. ids: The id of the tweet (2087)
3. date: the date of the tweet (Sat May 16 23:58:44 UTC 2009)
4. flag: The query (lyx). If there is no query, then this value is NO_QUERY.
5. user: the user that tweeted (robotickilldozr)
6. text: the text of the tweet (Lyx is cool)

Questions:

1. Explore and prepare the data (Tokenization, Stemming, Stopwords, visualization, etc.)
2. Build a BOW and train a KNN, Decision Tree, and SVM model
3. Evaluate the above models (confusion matrix, accuracy, classification report, etc.)
4. Use one of the word embeddings (word2vec, Glove, fasText) and build a CNN model and compare the result with question 2.
5. Build an API or a user interface to use the trained CNN model in production (this question is optional)

Tips:

Saving a Keras model:

```
model = ...  # Get model (Sequential, Functional Model, or Model subclass)
model.save('path/to/location')
```

Loading the model back:

```
from tensorflow import keras
model = keras.models.load_model('path/to/location')
```