Style Change Detection

Style change detection project is to identify the number of authors who have written a document by checking their style change in a given context and to identify text positions within a given multi-author document at which the author switches. Te project is subdivided into 2 tasks

- Task 1 is to find whether the given document is written by multiple authors or not.
- Task 2 is to identify the locations at which there is a change in styles.

Getting Started

Instructions for the project to execute in a step by step format on our local machine for development and testing purposes.

Clone fimgit repository using Source Tree 3.3.8 using university credentials from:

https://git.fim.uni-passau.de/padas/20ss-tmp/team14

Then install the following dependencies listed below.

Install dependencies

pip3 install -r requirements.txt

Optional dependencies

pip3 install jupyter

pip3 install xbboost

Installing

Install dependencies

pip3 install -r requirements.txt

Optional dependencies

pip3 install pydot (keras model visualization)
apt install graphviz (keras model visualization)
pip3 install jupyter
pip3 install h5py (saving keras models to disk)
pip3 install textstat

Prerequisites

Lightgbm 2.3.1
Matplotlib 3.1.3
Nltk 3.4.5
Jupyter notebook 6.0.3
Numpy 1.18.1
Pandas 1.0.4
Python 3.8.2
Scikit-learn 0.23.1
Seaborn 0.10.1
Textstat 0.6.2

Optional
xbboost 1.1.1

Source Tree

Running the tests

First we have to open the main jupyter file to follow each tasks,

    Code/style_change_detection.ipynb

Begin the project by importing all the relevant libraries which is available in the Section 1.1 of the jupyter notebook. Then download the required nltk resources by running the next cell. To read dataset and run the cells under section 1.2. Another way to read the file is by running only the last cell in this section which will read a CSV preloaded with the data. Lastly call the Section 1.3 to call all functions relevant to extracting features from the text.

Implementation to Task 1

To find out whether the given document is written by multiple authors or not. Run all the cells under the section 2.1 in jupyter notebook to achieve the task.

- The variable accuracy_task1_train will have the training accuracy and the variable accuracy_task1_test will have test accuracy.

Implementation to Task 2

To find the locations at which there is a change in styles.

- Run all the cells under the Section 2.2 in jupyter notebook to achieve the task. The variable df_task2 will have all the results for the task2.

Authors

- Anthony Odinukwe
- Shine Jose
- Subin Sam Varghese