**ITCS 4111/5111 Intro to NLP**

**Text Summarization - Readme**

Shashikant Jaiswal

Tejaswini Naredla

Anusha Balaji

## **Environment**

* Python 3.6
* PyCharm/Spyder IDE for Python (optional, GUI interface)

## **Required Packages**

* NLTK
* Sklearn (scikit-learn)
* Numpy
* rougescore

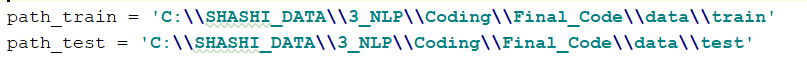
## **Data**

* CNN dataset (Training data : 3200 news articles and Test data : 800 news articles).

Google Drive link for Dataset: <https://drive.google.com/drive/u/1/folders/1F80QHO2qiZHNNK4njR2T5466nS00uujJ>

* Please download and extract the data ***from our submission*** link to the your project working directory

* Please change the values of the variables **‘path\_train’** and **‘path\_test’** to your local directory path of train and test datasets in both the python files i.e. ‘textSummarizer.py’ and ‘textRank.py’. Screenshot is given below.



.../data/train # train data with news articles (.SENT extension files)

.../data/test # test data with news articles (.SENT and .SUMM extension files)

* .SENT files are news articles and .SUMM files have corresponding summary to those news articles.
* The actual summaries given in the data set are used to evaluate the generated summaries.

Initial setup

* Install the python rougescore package from the following link: https://github.com/bdusell/rougescore
* Run **one of the below** command to install python rougescore package :

i) pip install git+git://github.com/bdusell/rougescore.git

ii) git clone https://github.com/bdusell/rougescore.git  
 cd rougescore  
 python setup.py install

## **Run!**

Tf-Idf approach to generate single document summary:

* Python textsummarizer.py

TextRank based approach:

* Python textRank.py

Instructions for running programs:

1. Import both .py files in Pycharm/Spyder IDE
2. Change the ‘path\_train’ and ‘path\_test’ variables in python files to your local paths
3. Run the .py files one by one