**AEGIS SCHOOL OF DATA SCIENCE**

CAPSTONE PROJECT: FINAL SUBMISSION

|  |  |
| --- | --- |
| Name | Prachi Ghadge |
| Batch | DF-1809 |
| Roll Number | - |
| Other Group Members (if any) | Prachi Bari, Falgunee Warutkar |
| Project Title | News Article Text Summarisation |
| Project Description | 1.The aim is to build an automatic news summarization system  2.Provide the user with a short summary of the news articles |
| Final deployment of the project | Using Python and Flask developed a Website which displayed summary generated by model of news article from different Indian news website. |
| List your major learnings from this project | 1. Learnt to generate extractive summary by using text rank and word frequency count algorithms  2. Learnt various concepts of data cleaning for textual data  3. Working of Lstm using sequence to sequence models. They are extremely slow. It took around 13 hours for 12 epochs of training.  4.Got familiar working with Flask  5.Learnt to do SQL database connectivity  6.Learnt to use boostrap,html,css and Jumbotron for designing UI. |
| Major tools used | Python 3.6: Anaconda  Libraries used- Flask, nltk.corpus, MySQLdb,sqlalchemy, create\_engine,glob,pandas,numpy,tensorflow |
| Final number of data points used in the project | 2GB Data |
| Mention cross validation techniques used, if any | None |
| Data sourced from | <https://timesofindia.indiatimes.com/>  <https://www.hindustantimes.com/>  <https://scroll.in/>  <https://www.indiatoday.in/>  <https://indianexpress.com/>  <https://www.deccanchronicle.com/>  <https://www.thehindu.com/>  <https://www.financialexpress.com/>  <https://zeenews.india.com/>  <https://www.inshorts.com/en/read> |
| Methods used to source the data | Selenium and bs4 Library |
| Challenges faced to source data |  |
| Describe the efforts spent and challenges overcome during data preparation, and the steps used | UI Development , Clean Summary |
| Outline the major steps used to implement the project | The deployment is using flask. |
| List the major nlp techniques used / implemented in your project | Extractive techniques  Lstm using sequence to sequence models. |
| Mention the major challenges faced and overcome during model creation | Time consumed was alot |
| List the techniques used to explore the data | Per day wise data analysis and model building. |
| List the outcome of data exploration | Inspect the length of texts |
| List any unsupervised learning technique used during data exploration | None |
| How many features were finally identified to create the model | None |
| Describe your Feature Engineering steps, if any |  |
| List and describe any feature reduction techniques used |  |
|  | **DETAILS ABOUT NLP TECHNIQUES USED** |
| Technique 1 | Text Rank Algorithm |
| Technique 1 quality metrics | Rouge Score : 0.7 |
| Technique 2 | Word Frequency Algorithm |
| Technique 1 quality metrics | Rouge Score : 0.6 |
| … | … |
| … | … |
| … | … |
|  |  |
| Recommendations for future work, if any | None |
|  |  |
| Final presentation prepared and uploaded to mUniversity | Date:01/08/2019 |
| Code uploaded to mUniversity | Date:01/08/2019 |
| Data uploaded to mUniversity | Date:01/08/2019  (If data size is very large, coordinate with Aegis Class Schedule, Aegis IT) |
| Presentation Video uploaded to mUniversity | Date:01/08/2019 |
|  |  |

oooOOOooo