

TEXT SUMMARIZATION OF NEWS ARTICLES

1.DESRIPTION:

In recent years, automatic text summarization has become a well-studied NLP problem. Machine learning models are built that can extract important information from a bigger piece of text and condense it into a smaller one. Deep neural networks have shown to be an excellent method for text summarization, and the concept has a wide range of applications. Text ad creation in internet search advertising is one conceivable use. Extractive summarization process involves giving scores to sentences using some method and then using the sentences that achieve highest scores as summaries. This project aims to develop a model which gives text summarization from news articles.

2.DATASET:

The dataset is taken from Kaggle: <https://www.kaggle.com/pariza/bbc-news-summary>

This dataset is used for extractive text summarization has business, political, sports, entertainment, and tech news articles of BBC from 2004 to 2005 in the News Articles folder. For each article, five summaries are provided in the Summaries folder. The first clause of the text of articles is the respective title.

3. METHODOLOGY

The major outside tools that we plan to use are Scikit-learn and Tensorflow, as well as matplotlib and seaborn for some visualization. We are going to use LDA from Gensim package and 3- stacked LSTM to summarize the news article. Stacked LSTM has multiple layers of LSTM stacked on top of each other. This leads to a better representation of the sequence. We hope to visualize the data, and test whether the summarization for each news article is accurate or not.

4. TIMELINE:

Week 9: Clean and pre-process the data (Tokenize, remove stop words, and lemmatize)

Week 10 & 11: Implement the two models (LSTM and LDA) and train them.

Week 12: Test and do final minor adjustments. Visualize the results.

Week 13: Write the report and record the final presentation.

5. RESPONSIBILITIES

There are four people in our group: Varsha Reddy Anugu, Lasya Manthripragada, Saketh Dathrika, and Nithish Goud Podeti. Each member will contribute evenly to the project. Varsha and Lasya will contribute mainly towards cleaning and getting the data ready to implement the models. Saketh will work on the model LDA, and Nithish will contribute towards LSTM model. We all will evaluate the models and visualize the results at last to see whether we got the desirable results.