MANDATE 1 - NEWS SUMMARIZATION

Mani Nandadeep Medicharla(IMT2019051)
R Prasannavenkatesh(IMT2019063)

ABSTRACT:
The goal of this project is to build a text summarizer and use it to summarize the news articles. A text summarizer is a NLP based tool that wraps up text to a short length. It condenses a long article to main points. The need for text summarizers are increasing in everyday life since people want news quick and to the point rather than long and shabby walls of text. Social media sites like reddit also use a text summarizer called autotldr which is a bot that uses SMMRY to automatically summarize long reddit submissions. People are looking for shortcut methods to learn ideas in less time. Text summarizers are also helping people to decide whether a book, a research paper, or an article is worth reading or not.

1 Mandate - 1

1.1 Dataset

https://www.kaggle.com/sunnysai12345/news-summary

The dataset consists of 4515 examples and contains Author name, Headlines, URL of Article, Short text, Complete Article. The news has been summarized from Inshorts and only scraped the news articles from Hindu, Indian times and Guardian in time period ranges from February to August 2017.

This is an example of a dataset we are planning to take. We will be using this as well as other similar datasets to train and test the model. We are also planning to experiment with indic languages once the initial model for english news summarization is done.

1.2 Text Summarization Categories

1.2.1 Extractive Summarization

The extractive approach involves picking up the most important phrases and lines from the documents. It then combines all the important lines to create the summary. So, in this case, every line and word of the summary actually belongs to the original document which is summarized.

1.2.2 Abstractive Summarization

The abstractive approach involves summarization based on deep learning. So, it uses new phrases and terms, different from the actual document, keeping the points the same, just like how we actually summarize.

1.3 Preprocessing approaches

Some of the preprocessing approaches we will be using are:

- Contraction expanding
- Case checks, i.e, using a standard case system
- Remove Punctuations, emojis, URLs and other non standard text
- Remove Stopwords
- Rephrase Text
- Stemming
- Lemmatization
- Remove White spaces

1.4 Text Summarization Approaches

These are some of the approaches that we try to implement

- **Seq2Seq:** Sequence to Sequence (often abbreviated to seq2seq) models is a special class of Recurrent Neural Network architectures that we typically use (but not restricted) to solve complex Language problems like Machine Translation, Question Answering, creating Chatbots, Text Summarization, etc.
- **LSTM:** Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. The two technical problems overcome by LSTMs are vanishing gradients and exploding gradients, both related to how the network is trained.

- **Glove:** GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.
- **T5 Transformer model:** T5 is an encoder-decoder model pre-trained on a multi-task mixture of unsupervised and supervised tasks and for which each task is converted into a text-to-text format. It uses relative scalar embeddings. All NLP tasks are converted to a text-to-text problem. Tasks such as translation, classification, summarization and question answering, all of them are treated as a text-to-text conversion problem, rather than seen as separate unique problem statements.

Evaluating the above models will mostly be done using human evaluation and ROGUE evaluations. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It is the method that determines the quality of the summary by comparing it to other summaries made by humans as a reference.

2 References

- https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/
- https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/
- https://towardsdatascience.com/understand-text-summarization-and-create-your-own-summarizer-in-python-b26a9f09fc70