AUTOMATIC TEXT SUMMARIZATION

KDM Project Submission 1 - Spring 2021

Team number : 3

Team Members :

Charles Scola, Bayard Rucker, Vyoma Desai, Claire Ndofor

Motivation:

Our motivation is driven by the interesting topics we came across during our knowledge and discovery management course and how convenient some of the tools we learned made work easy and time saving. With the text summarization idea , we will be able to accomplish a similar goal by making research easier for students and industries. Inorder to obtain our goal, we ask ourselves the following questions to get an objective

- ○ Can we create a summary with the major points of an original document?
- ○ Abstractive (write your own summary) and Extractive (select pieces of text from original) are two popular approaches

Dataset: CNN and DailyMail News Pieces by Google DeepMind.

CHAPTER 1 (LIFE) :

1) **Define your scope or domain where the use case is relevant or prevalent?**

   Summarizing text of reasonable length based on online data. This model should be able to take non-technical text and summarize it for the end users. This will be relevant to organizations who work with large sets of data and need to obtain a summary for quicker understanding. For this project, we will be working with CNN and DailyMail News pieces by google DeepMind. This dataset contains the documents and accompanying questions from the news articles of CNN. There are approximately 90k documents and 380k questions.

2) **What is your main story?**

   SInce we are working with a dataset from CNN, our main story is focusing on a small broadcast tv channel who just started in the business and is trying meet up with it’s peers. This company often finds itself with large loads of data to summarize and report to the community. Our Automatic text summarization idea comes in to assist them with the summarization tasks. This will save the company time and money. We also focus on graduate students who find themselves often summarizing articles or writing research papers. We look at a case a student who is unable to collect the correct sources since most of his sources were filled with cumbersome data and was hard to read through the entirety of each source. We take an example of students who have to work fulltime in order to pay tuition fees and still maintain good grades.

3) Who are the characters or people in the main story?

Our main character focuses on news reporters and data analysts who gather large amounts of data in order to filter what is most important. Furthermore, we focus on Full Time students who try to balance work and school life in order to meet up with their daily needs, pay student loans and still maintain good grades. Lastly, we focus on people who are looking to gain information about a specific field of study

4) What problem happened to them?

The amount of time taken to go through large amounts of data only to summarize what is important. Due to the overwhelming amount of text (lots of paragraphs/lengthy sentences etc.), they were unable to understand the source material or were unable to find good sources (unable to quickly skim through multiple resources).

5) Where did the problem take place?

a. “Where” means two things: 1) The environment and settings that the people or the community is living in, and 2) the place/location where the problem take place.

This problem occurs throughout both academic and industry working environments. This probably typically takes place while doing surfing / research on the web.

6) Why?

a. Why means the possible causes and/or origin of the problem?

This problem has come from the massive increase in data in this internet age. This huge set of internet data decreases the content readability. The origin of the problem has come from downloading/surfing automatic text summarization web tools and softwares. These web tools and softwares cut-copy-paste sentences (they basically pick and place texts) and thus this became the most common cause of unhappy students / net surfers who are unable to extract relevant information at one go

7) **How?**

Text summarization is the process of extracting phrases and sentences from a document to make up a new concise summary. Techniques involve shortening long pieces of text into one single meaningful sentence and creating a summary covering only the main / relevant point addressed in the document. Due to increase in content and web technologies, many tools and software cheat the students/customers/net users by charging them for the services to generate concise summaries but end up giving a crappy cut-copy-paste solution. Our algorithm will define these drawbacks and will create a new summary without changing the meaning of the sentence. And will also eliminate copy/paste structure inside the document.

Our team is not using a Kaggle challenge for this project.