Natural Language Processing (DS 5007)

Project Proposal

Submitted By: **Muhammad Raamish Alam (20k-1326), Ashar Ansari (20k-1409), Syed Muzammil (20k-1394)**

Submitted To: **Dr. Raza Abbas**

Course Name: **NLP(DS 5007) - Fall22**

Dated : **Sep 18 2022**

Title:

Text/Data Summarization end to end deployed on Flask Server.

Objective:

The goal of this project will be to demonstrate one of the many applications of NLP which is text summarization. Our final project will be a web application deployed on a flask server which takes any coherent sequence of data and with the approaches discussed below, compress and summarize it so it is leaning towards the actual point that is being discussed within it.

Background:

If a book captures our interest, do we start reading it at once? Not likely. We are more inclined towards reading its "Blurb" or an "Review" to determine how the book likely is presented and if we should dedicate the necessary time to go through it or not.

Another example is when we open any online news sites, do we simply begin reading each news story? Presumably not. We commonly look at the short news summary and afterward read more details provided we are intrigued.
Short, informative outlines of the news are currently everywhere like magazines, news aggregator applications, research sites , and so on.

Indeed, It is feasible to make the synopses naturally as the news rolls in from different sources all over the planet.

The strategy for removing these outlines from the first tremendous text without losing fundamental data is called Text summarization. It is fundamental for the outline to be familiar, nonstop and portray the critical.

As a matter of fact, the google news, the inshorts application and different other news aggregator applications exploit text synopsis calculations.

Literature Review:

Text summarization is an emerging research field. H.P. Luhn in 1958(Luhn, 1958) presented this examination philosophy. He proposed a technique to remove the significant sentences from the text utilizing elements like expression and word recurrence (Allahyari et al, 2017).].

Interaction of Programmed Text Outline incorporates removing or assembling huge information from unique substances and displays that information as synopsis (Nitu et al, 2017). It decreases the powerful chance to get the core of the information. Need for synopsis should be visible for various reasons and in various spaces, for instance synopsis of news stories , messages, statistical surveying, data connected with government specialists, clinical history of patients and sicknesses and so on. Summarization has variations as it should be possible on a solitary record and furthermore on various archives on comparative topics.

Summarization tools are likewise accessible online in view of the sort of information to be handled for various fields like news article summarizers like Columbia News blaster (Saranyamol and Sindhu, 2014) and clinical field related rundown devices like Total Essential (Gaikwad and Mahender, 2016).Classification of text rundown depends on different norms (Rani and Tandon, 2018). In view of these norms three rules are being set : input sort of report, reason criteria , archive yield measures (Aries et al, 2019). Early rundown was finished on the single report which produces the rundown of a solitary record (Khan and Salim, 2014). Be that as it may, as the information increases, a multi archive outline arose.

Approach:

Within the bounds of NLP, Text summarization can be constituted within two major categories. **Extractive** and **Abstractive methods**.

When it comes to Extractive Text Summarization It is the conventional method developed first. The principal objective is to recognize the critical sentences of the text and add them to the summary being accumulated. We really want to take note of the summary that contains accurate sentences from the original data.

While Abstractive Text Summarization. It is a relatively advanced technique, numerous progressions continue to come out frequently. The approach is to identify the significant segments, decipher the specific situation and reproduce in another manner. This guarantees that the core data is passed on through the most limited text conceivable. Note that here, the sentences in outline are created, not simply separated from original text.

We will be relying on both approaches, and experiment to see which one produces the desired results that harmonizes with being displayed and worked on the web application that will be deployed on flask.

Dataset:

https://metatext.io/datasets-list/summarization-task

The link displayed above contains links to some of these data sets that can be used in summarization applications.

WikiSummary: A summarization dataset extracted from Wikipedia.

IndoSum: Dataset for text summarization in Indonesian that is compiled from online news articles and publicly available.

Multi-Xscience: A multi-document summarization dataset created from scientific articles. MultiXScience introduces a challenging multi-document summarization task: writing the related-work section of a paper based on its abstract and the articles it references.

NewSHead: Dataset contains 369,940 English stories with 932,571 unique URLs, among which we have 359,940 stories for training, 5,000 for validation, and 5,000 for testing, respectively. Each news story contains at least three (and up to five) articles.

NCLS-Corpora: Contains two datasets for cross-lingual summarization: ZH2ENSUM and EN2ZHSUM. There exists 370,759 English-to-Chinese cross-lingual summarization (CLS) pairs from ENSUM and 1,699,713 Chinese-to-English CLS pairs.