Extractive Summarization

Geetam
Department of Electrical Engineering, Indian Institute of Technology Kanpur, Uttar Pradesh, 208016
Email: {geetam}@iitk.ac.in

***Abstract*—There is an enormous amount of textual material, and it is only growing every single day.There is a great need to reduce much of this text data to shorter, focused summaries that capture the salient details, both so we can navigate it more effectively as well as check whether the larger documents contain the information that we are looking for. Text summarization is the problem of creating a short, accurate, and fluent summary of a longer text document.Automatic text summarization methods are greatly needed to address the ever-growing amount of text data available online to both better help discover relevant information and to consume relevant information faster. [?]**

I. Introduction

This paper has been inspired by "Summarizing Situational Tweets in Crisis Scenarios - K Rudra et al " and I have also applied these methods on a tweet data-set . We do summazarization in two steps ,first we do extractive summarization to shorten the text , remove duplicate sentences or sentences which are very similar . Here, content is extracted from the original data, but the extracted content is not modified in any way. Examples of extracted content include key-phrases that can be used to "tag" or index a text document, or key sentences (including headings) that collectively comprise an abstract, and representative images or video segments, as stated above. For text, extraction is analogous to the process of skimming, where the summary (if available), headings and subheadings, figures, the first and last paragraphs of a section, and optionally the first and last sentences in a paragraph are read before one chooses to read the entire document in detail . This steps also takes care of the redundant sentences which are contain words that are normally used in sentences which dont convey much meaning .This step gives rise to text which is shortened but is unpolished . Due to the Covid constraints , my project could only be completed upto this stage . The 2nd step is condensing this text into a coherent and informative summary using path matching by graph algorithms . Abstractive methods build an internal semantic representation of the original content, and then use this representation to create a summary that is closer to what a human might express. In theory , two sentences like "Tribhuvan international airport will remain closed" and "airport will remain closed from 6pm to 9pm " will have highly similar and overlapping content and thus will be condensed into a single sentence like "Tribhuvan international airport will remain closed from 6pm to 9pm .In other words this step aims to further reduce redundancy and increase informativeness .

II. Procedure for Extractive Summarization

We first look how this process works with summarization of disaster tweets , then discuss its effectiveness in summarizing general text . We primarily use content words to measure the informativeness of a sentence . We extract a set of tweets with 1,000 word limit constraint in our initial extractive phase of summarization. In a nut-shell the system takes tweets and selects the tweets having the best value or more "content words" and removes duplicates with a given constraint like a set limit of final summary length . First of all , the tweets are split up into their constituent words and their frequencies into a dictionary . We exclude "stop words" like too , if , my , a etc because they dont add much value to a tweet or a text . A vector of tweets is made and at this stage , duplicate tweets are discarded by using comparison techniques . At this stage , we start to give value to words by calculating their TF-IDF scores . Scores of words having length less than or equal to 2 are set to 0 because they don't add much value to a summary and thus should not effect a tweet's score .

A. Introduction to Mathematics behind the problem

The problem at hand reduces to the following - We have to maximize the informativeness of the extracted summary while keeping the length of the summary within a specified limit . Mathematically the problem now becomes - We have to maximize the overall TF-IDF score of the summary while limiting the total number of words in the summary below a specified limit (1000 in this case). We create a linear programming problem and solve it using GurobiPy solver . The Gurobi Optimizer is a commercial optimization solver for linear programming, quadratic programming etc . We create a model with various constraints and solve the linear problem . We define variables corresponding to tweets and content words which will be governed by constraints . The same variables for tweets will act like variables for sentences when we summarize text instead of tweets . We denote the ith tweet with tweet[i] and the jth content-word with word[j] . All the variables are binary having values 0 or 1 .

If tweet[i] is 1 , it means the ith tweet will be included in the final summary , if its 0 then ith tweet will be omitted from the final summary .

If word[j] is 1 , it means the jth content-word will make atleast one appearance in the final summary .

B. Formulating and solving the Mathematical problem

We define an objective function P which has to be maximized by this process .

P = tweet[0] + tweet [1] + ... +tweet[i] + ... tweet[n-1] + tf-idf[0]*word[0] +tf-idf[1]*word[1] + ... +tf-idf[j]*word[j] + ... +tf-idf[m-1]*word[m-1]

Our objective is to maximize the numerical value of this function subject to the following constraints :-

1) Constraint 1: If a tweet(n) is selected in the final summary , then all the content-words in it must also be selected . We mathematically represent this by :-

word[i] + ... + word[j] >= (number of content words in the tweet)*tweet[n-1]

here , tweet[n-1] contained content-words i through j .

2) Constraint 2: If a content-word(m) is selected then atleast one tweet containing this content-word has to be selected . We mathematically represent this by :-

tweet[i] + ... + tweet[j] >= word[m-1]

here , tweets i through j contained the content word word[m-1] .

All of these constraints are fed in the GurobiPy solver along with the objective function which gives the output . Tweet variables with value "1" are included in the summery and the ones with the value "0" are excluded from the summary .

3) COnstraint 3: This he sum of lengths of all the selected tweets must be less than the summary length limit specified "L" .

We mathematically represent this by :-

tweet[0]*Length[0] + tweet[1]*Length[1] + ... + tweet[i]*Length[i] + ... tweet[n-1]*Length[n-1]

III. Some Results

A. Algorithm test on disaster tweets

.

About 8000 of the following type of tweets obtained randomly from a disaster were passed through the summarizer :-

.

RT @kundadixit: Most new high rises in Kathmandu ok, old buildings down. Temples reduced to rubble. 0.9

Kathmandu airport shut, flights from India cancelled: New Delhi, April 25 (IANS) The Tribhuvan International A... http://t.co/n7wEYT03Uv 0.9

RT @ndtv: After massive 7.9 earthquake, commercial flights to Kathmandu put on hold http://t.co/ZcCOxQ8SBE http://t.co/uA7HYDAFnL 0.9

RT @drkerem: Kathmandu airport now open amp; operating 0.9

RT @cctvnews: China's Tibet severely affected by NepalEarthquake; houses collapsed, communications cut off http://t.co/NoT6imSVWu 0.9

RT @Superneha83: My cousin just sent me these pictures. The entire bhaktpur durbar square is no more! Kathmandu earthquake http://t.co/ec… 0.9

RT @kundadixit: Most new high rises in Kathmandu ok, old buildings down. Temples reduced to rubble. 0.9

RT @ndtv: After massive 7.9 earthquake, commercial flights to Kathmandu put on hold http://t.co/ZcCOxQ8SBE http://t.co/uA7HYDAFnL 0.9

Promtaction by modisarkar to send relief materials to Nepal. 2 flights to take off frm UttarPradesh soon. narendramodi NepalEarthquake 0.9

Flights seem to be getting out though https://t.co/b6tAiDqMAB 0.9

RT @HeadlinesToday: NEWS FLASH: Indian flights to Kathmandu put on hold earthquake http://t.co/RbNiUoQvpE 0.9

RT @kundadixit: Kathmandu Valley shrouded in dust from 7.4 mag quake. Airport closed. Lots of damage. http://t.co/QF3kE65Tzf 0.9

RT @kundadixit: Kathmandu airport now open. 0.9

RT @anilkapur:$TribhuvanAirportatKathmanduisopenednowforC$ $130Aircraftfrom Indiacarryingreliefmaterialstolandhttp$: $//t.co/e9\ldots$ 0.9

RT @IndianExpress: Flights to Kathmandu put on hold following powerful earthquake | Read more here: http://t.co/IofmNwVtzT http://t.co/1hqN… 0.9

RT @kundadixit: Kathmandu airport now open. 0.9

RT @AnupKaphle: That's good news. RT @kundadixit Kathmandu airport now open. NepalQuake 0.9

RT @kundadixit: Most new high rises in Kathmandu ok, old buildings down. Temples reduced to rubble. 0.9

RT @cnnbrk: Buildings are down and roads are out after major Nepal earthquake, CNN sister network CNN-IBN reports. http://t.co/E8Fh03tnSi 0.9

RT @kundadixit: Kathmandu airport now open. 0.9

After Massive 7.9 Earthquake, Flights to Kathmandu Put on Hold http://t.co/eCLwPIydSc 0.9

NEWS FLASH: Indian flights to Kathmandu put on hold earthquake http://t.co/RbNiUoQvpE 0.9

RT @kundadixit: Most new high rises in Kathmandu ok, old buildings down. Temples reduced to rubble. 0.9

RT @cnnbrk: Buildings are down and roads are out after major Nepal earthquake, CNN sister network CNN-IBN reports. http://t.co/E8Fh03tnSi 0.9

.

The following is a part of the 1000 word summary obtained :-

.

RT @cctvnews: China's Tibet severely affected by NepalEarthquake; houses collapsed, communications cut off http://t.co/NoT6imSVWu

RT @HeadlinesToday: NEWS FLASH: Indian flights to Kathmandu put on hold earthquake http://t.co/RbNiUoQvpE

RT @airlivenet: ALERT Kathmandu airport shut post earthquake; flights diverted to India (Pic: Reuters) http://t.co/IXCUUgmMKg http://t.co/j…

RT @emilyrauhala: NepalQuake: Tibet also hit hard. China Daily reports 70

RT @emilyrauhala: Nepal earthquake also toppled buildings in Tibet, Chinese state media reports. No word on casualties there. http://t.co/S…

RT @RTcom: NepalEarthquake: Dozens trapped inside collapsed Dharahara tower - local media http://t.co/E6wZrronVg http://t.co/fBozbaKKm5

RT @kundadixit: Kathmandu Valley devastated by huge quake. Lots of buildings down. Pall of dust over city.

NYT story: Earthquake in Nepal Kills Hundreds and Levels Buildings http://t.co/YQHJqHEzu8

RT @NBCPhiladelphia: 7.9 magnitude earthquake rocks Nepal, leveling buildings, causing an avalanche on Mt. Everest: http://t.co/FTPIwbOFTs … Bodies seen after Nepal quake topples landmark tower http://t.co/lwWWNT0eKl

As is already clear , the duplicate tweets have been discarded . Furthermore , most of the tweets in the summary are meaningful ones and supply unique information .

B. Algorithm test on a short story

We use the woodcutter's story as it is well-known :-

Long ago, there lived a woodcutter in a small village. He was sincere in his work and very honest. Every day, he set out into the nearby forest to cut trees. He brought the woods back into the village and sold them out to a merchant and earn his money. He earned just about enough to make a living, but he was satisfied with his simple living.

One day, while cutting a tree near a river, his axe slipped out of his hand and fell into the river. The river was so deep, he could not even think to retrieve it on his own. He only had one axe which was gone into the river. He became a very worried thinking how he will be able to earn his living now! He was very sad and prayed to the God. He prayed sincerely so the God appeared in front of him and asked, "What is the problem, my son?" The woodcutter explained the problem and requested the God to get his axe back.

The God put his hand deep into the river and took out a silver axe and asked, "Is this your axe?" The Woodcutter looked at the axe and said "No". So the God put his hand back deep into the water again and showed a golden axe and asked, "Is this your axe?" The woodcutter looked at the axe and said "No". The God said, "Take a look again Son, this is a very valuable golden axe, are you sure this is not yours?" The woodcutter said, "No, It's not mine. I can't cut the trees with a golden axe. It's not useful for me".

The God smiled and finally put his hand into the water again and took out his iron axe and asked, "Is this your axe?" To this, the woodcutter said, "Yes! This is mine! Thank you!" The Goddess was very impressed with his honesty so she gave him his iron axe and also other two axes as a reward for his honesty.

here is the 140 word summary of this story

Long ago, there lived a woodcutter in a small village He was sincere in his work and very honest Every day, he set out into the nearby forest to cut trees He brought the woods back into the village and sold them out to a merchant and earn his money He earned just about enough to make a living, but he was satisfied with his simple living The river was so deep, he could not even think to retrieve it on his own He became a very worried thinking how he will be able to earn his living now! He was very sad and prayed to the God I can't cut the trees with a golden axe It's not useful for me"

.

As we can see , the summary is not satisfying as the original story itself is very short and its very difficult to summarize an already small text . I have done a similar test on a bigger story(it is in the demo folder) and results are better on that . This confirms that the extractive method is effective on large texts since it doesnot changes the skeletal structure of the text .

IV. CONCLUSION AND ANALYSIS

I was working on the methods to quantify the effectiveness of this method in summarizing huge volumes of tweets ,feedback surveys and stories . A skim over the resulting summary from different media - tweets , stories etc does show promise and show that the method is indeed effective in summarizing huge volumes of data to workable length text . Tweets summary has minimal redundancy and is mostly informative for its 1000 word length limit . The method at hand is very effective

V. FUTURE WORK

Unfortunately , due to the ongoing Covid-19 pandemic I could only work on the extractive phase of the summarization process which essentially , objectively shortens the text from its initial humongous size down to about a 1000 word length . Next step originally planned was the Abstractive phase which merges sentences of similar meaning to prepare a summary that further maximizes the informativeness and reduces the redundancy of the text . This steps takes the summary closer to what a human might prepare . I plan to work on this step further after the normal semester starts .

VI. REFERENCES

[1] Koustav Rudra , Pawan Goyal, Niloy Ganguly, Muhammad Imran, and Prasenjit Mitra :- Summarizing Situational Tweets in Crisis Scenarios: An Extractive-Abstractive Approach