Generating Summary for Terms and Conditions Document Based on the Query Provided.

SAI TEJA PALADI, University of South Carolina, USA

We introduce a system for summary generation based on the terms of service documents we usually agree to by reading the entire or useful content. Generally, we skip reading large documents and end up agreeing to some things that we don't like to do. What if we have a system that can summarize the piece of information that you are worried about in a large document and get the important sentences? Our system takes the URL of these terms and conditions documents and the query we are concerned about as the input and produces the summarized sentences.

ACM Reference Format:
Sai Teja Paladi. 2022. Generating Summary for Terms and Conditions Document Based on the Query Provided. . *J. ACM* 37, 4, Article 111 (August 2022), 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Usually, when we come across the terms and conditions documents, most of the users just simply click on the agree button and move on rather than reading the entire document. And some people might search for some points, like is their data is confidential or not, and to do so, they have to search the entire document finally, they may find the information that is relevant, but they have to read the information and again fetch the important points. Despite doing all these tasks, there is a chance of missing the overall objective of the information that we see. What if we have a system that can take the URL of the online document that we are referring to and the focused area as input and give us the sentences that cover the area we need to focus on. If we do so, we don't need to worry about reading the entire document or missing any important points.

2 PROBLEM

In our system, we have two inputs one is the document that we are currently dealing with In our project, we are summarizing the terms and conditions documents which mainly talk about the privacy of the data. And we are experimenting on mainly three documents which are mentioned below.

2.1 Input

- Samsung T&C.
- Apple T&C.
- OnePlus T&C.

Area of focus: Data, user data, confidentiality, private information

And based on the provided input document and the area of interest we will be fetching out the sentences that provide the useful information.

Author's address: Sai Teja Paladi, University of South Carolina, USA.

0004-5411/2022/8-ART111 $15.00
https://doi.org/XXXXXXX.XXXXXXX

2.2 Output

Output is the summarized text based on the document and the initial query provided by the user.

Sample output: When you upload, transmit, create, post, display, or otherwise provide any information, materials, documents, media files, or other content on or through our Services, you grant us an irrevocable, unlimited, worldwide, royalty-free, and non-exclusive license to copy, reproduce, adapt, modify, edit, distribute, translate, publish, publicly perform, and publicly display such User Content to the full extent allowed by Applicable Law.

3 RELATED WORK

In the text summarization there are two types of summarizations. They are extractive and abstractive text summarizations. These is lot of work has focused on Extractive summarization. Extractive summarization takes the original text and extracts information that is identical to it. In other words, rather than providing a unique summary based on the full content ((Neto et al., 2002), (Erkan and Radev, 2004), (Filippova and Altun, 2013), (Colmenares et al., 2015), (Riedhammer et al., 2010), (Ribeiro et al., 2013)). There has been some work on abstractive summarization in the context of DUC-2003 and DUC-2004 contests (Zajic et al.). We refer the reader to (Das and Martins, 2007) and (Nenkova and McKeown, 2012) for an excellent survey of the field.

4 APPROACH

In our problem, we cannot directly perform the text summarization to get the sentences based on the area of focus we have to first get the relevant data or passages from the long documents that we use as input based on the area of focus that we provide and ask for.

After getting the content from the extractive text summarization, we provide this content as input and perform abstractive text summarization on this input and finally, we get our desired sentences.

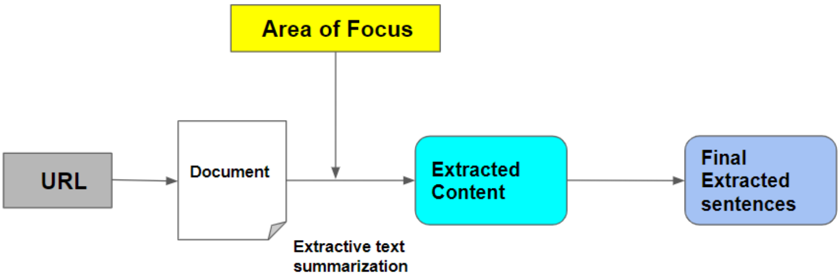


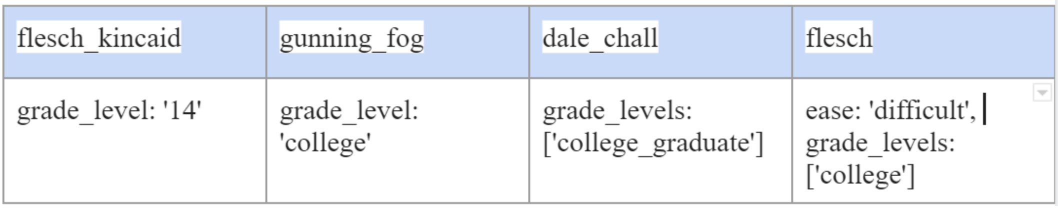
Fig. 1. Text Summarization

4.1 Readability metrics

Before getting into the actual data pre-processing and summarization. We are using some readability metrics for the raw data in the document. These readability metrics help us to understand the ease of how a user can read the actual document. It depends on the document's content and the document's size.

Here are some results of four types of readability metrics that we used to evaluate the document.

We can clearly observe that for each readability metric here the level of difficulty is above the grad level which is difficult to read.



flesch_kincaid	gunning_fog	dale_chall	flesch
grade_level: '14'	grade_level: 'college'	grade_levels: ['college_graduate']	ease: 'difficult', grade_levels: ['college']

Fig. 2. Readability Metric Results for the Input Document

4.2 Text summarization is mainly divided into two categories

4.2.1 Extractive Summarization. The extractive text summarization methods rely on extracting sentences and phrases from the documents and creating summary. Identifying the correct sentences for summarization is an important task in this method.

4.2.2 Abstractive Summarization. The abstractive text summarization methods use advanced NLP techniques to generate an entirely new summary. Some parts of the summary will not even appear in the original text.

4.3 Extractive Summarization

In our project, we will first generate the important sentences based on the extractive text summarization that we perform on our documents, and then we will rank the sentences based on the query that we are looking for in summary. We will use the TextRank algorithm on the data we extract from the URLs and create a brief summary based on the provided content. This will be the first step in our summarization task. The TextRank algorithm is an extractive and unsupervised text summarization technique which is used to find the similarity between the sentences and store the similarity scores in the square matrix for the sentence rank calculation. And finally, the top-ranked sentences are used as the summary.

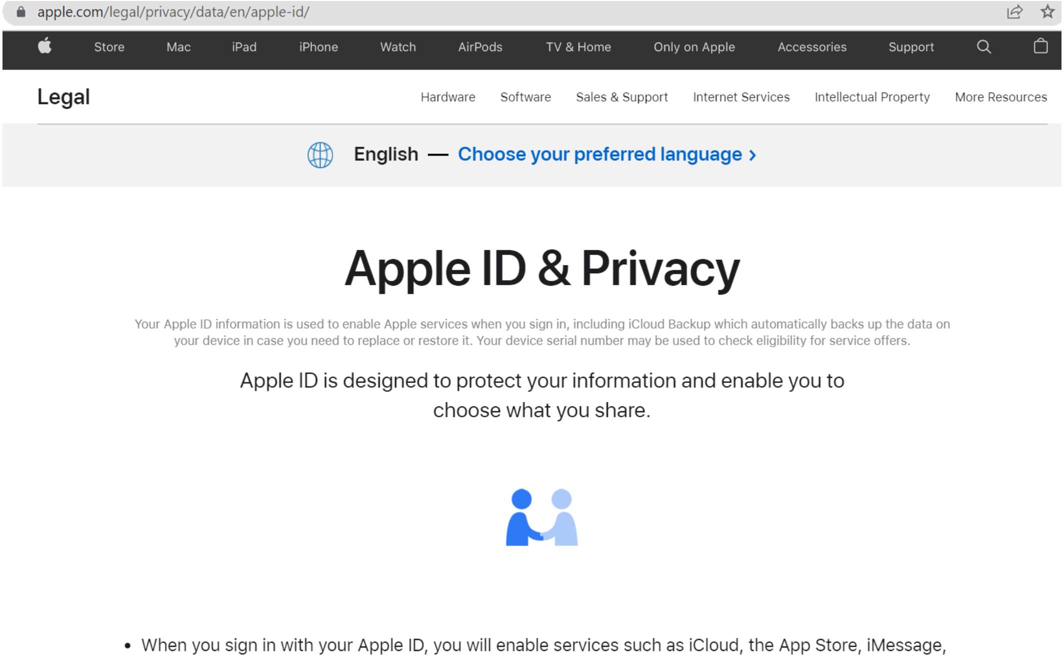
4.3.1 Steps in Extractive Summarization.

- Parsing the HTML content from the URL that we have provided as an input. We create a .txt document based on the parsed data
- We will split the text into into individual sentences. We will use the senttokenize( ) function of the nltk library to do this.
- We download GloVe Word Embeddings to create features for our sentences
- We perform text cleaning by removing the punctuations, numbers and special characters. We will also get rid of the stopwords(is, am, the, of, in, etc) by dowloading nltk-stopwords and defining a function to remove the stop words.
- And finally we use the sentences to create vectors for sentences in our data using the GloVe word vectors.
- Finally we will prepare a similarity matrix. We will initialize this matrix with the cosine similarity scores of the sentences. And we convert the similarity matrix into a graph. The nodes of this graph will represent the sentences, and the edges will represent the similarity scores between the sentences. We will apply the PageRank algorithm on this graph to arrive at the sentence rankings. And we will extract the top sentences based on their ranking for a summary generation.

5 EVALUATION

Our main aim is to generate summarized sentences based on the query that we provide.

For Example, if a user wants to know about the data he is sharing with the company. Using our model, we can provide the query as Data Security and fetch the top extracted summarized sentences.



Apple ID & Privacy

Your Apple ID information is used to enable Apple services when you sign in, including iCloud Backup which automatically backs up the data on your device in case you need to replace or restore it. Your device serial number may be used to check eligibility for service offers.

Apple ID is designed to protect your information and enable you to choose what you share.

- When you sign in with your Apple ID, you will enable services such as iCloud, the App Store, iMessage,

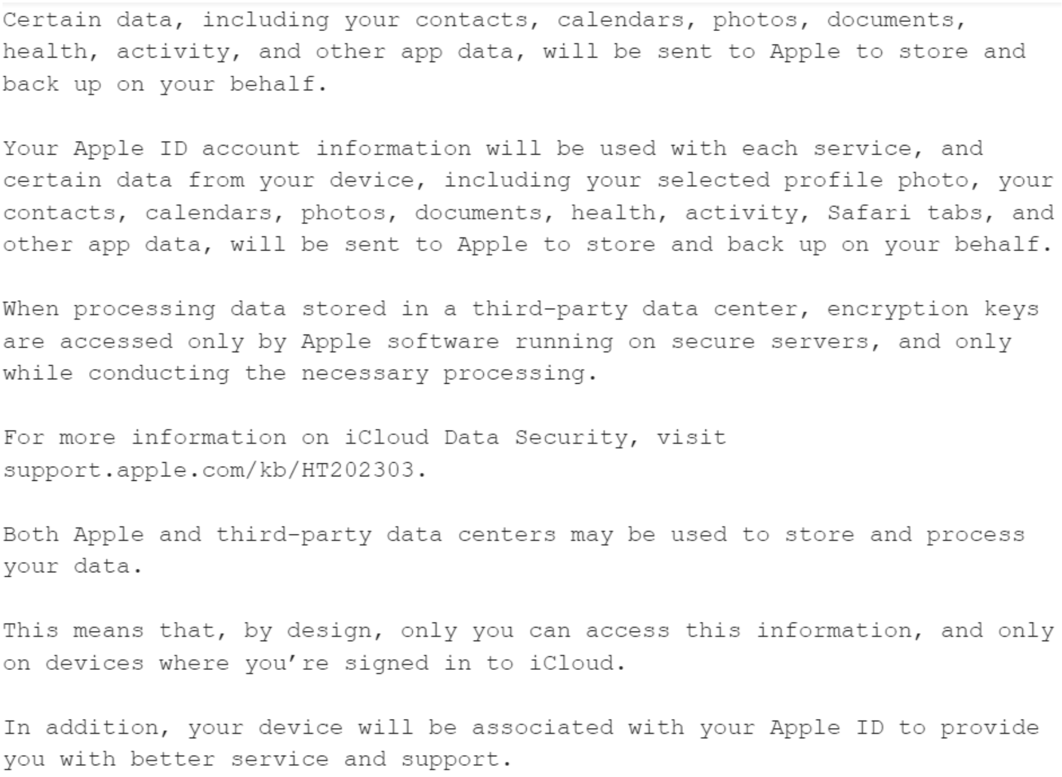
Fig. 3. Input document for the summary

As shown in the above figure, We use terms and conditions documents of various service providers like apple, Samsung, etc. And try to summarize the content and get all the summarized sentences that are relevant to the query that we are looking at, which is "Data Security" or "Data Privacy."

As shown in the figure below, all the top extracted sentences involved in data security are listed concisely. A user can go through these sentences to learn about his data privacy and security rather than the entire document.

6 CONCLUSION

Summarization provides the user with a small amount of important data that he's concerned about when reading or agreeing to a large amount of data or secured documents. Which may involve his confidential data. Using this query-based summary, we can get the required data and in a short format.



```
Certain data, including your contacts, calendars, photos, documents,
health, activity, and other app data, will be sent to Apple to store and
back up on your behalf.

Your Apple ID account information will be used with each service, and
certain data from your device, including your selected profile photo, your
contacts, calendars, photos, documents, health, activity, Safari tabs, and
other app data, will be sent to Apple to store and back up on your behalf.

When processing data stored in a third-party data center, encryption keys
are accessed only by Apple software running on secure servers, and only
while conducting the necessary processing.

For more information on iCloud Data Security, visit
support.apple.com/kb/HT202303.

Both Apple and third-party data centers may be used to store and process
your data.

This means that, by design, only you can access this information, and only
on devices where you’re signed in to iCloud.

In addition, your device will be associated with your Apple ID to provide
you with better service and support.
```

Fig. 4. Output document of the summary based on the Data Security Query on Apple terms and conditions

7 REFERENCES

(1) Moratanch N, Gopalan C (2017) A survey on extractive text summarization. pp 1–6

(2) Rahim Khan, Yurong Qian, Sajid Naeem. Extractive based Text Summarization Using K-Means and TF-IDF, International Journal of Information Engineering and Electronic Business, May 2019.

(3) PL.Prabha, M.Parvathy. Extractive and Abstractive Text Summarization Techniques, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-9 Issue-1, May 2020.

(4) https://www.sciencedirect.com/science/article/abs/pii/S0885230820300991

(5) https://blog.agolo.com/query-based-summarization-in-action-ea729df3109c

(6) https://ieeexplore.ieee.org/document/7557434