**UML602**

****

**SUBMITTED TO :- SUBMITTED BY:**

**ARSHDEEP KAUR DAMANPREET SINGH (101611017)**

**BHAVNEET KAUR (101603075)**

**INTRODUCTION**

Summarization is an essential aspect of any learning activity, especially when the text involved is lengthy and talks about a complex subject. Summarizing a text allows the reader to discern its most important ideas, ignore information that is irrelevant and often redundant, and obtain a central, fundamental idea of the text in a meaningful way.

However, summarization can be a daunting task since it asks for the reader to skim through the entire text and extract the core and essential ideas, and continuously take large portions of text and reduce each to its main points for more concise understanding. Summarization strategies, however effective and fast, cannot ignore the entire skimming act.

Therefore, an application that does said task for the user could be a very welcome and unique idea in that most people are non-readers and mostly have the patience to read entire texts let alone extract ideas out of them. The text summarizer program we provide in the following pages does exactly this, and does so in simple brief steps, generating a lucid, effective summary covering all the essential ideas of the text that is fed to it. The program is short and efficient and holds potential to be liked by many along with providing more and more people with their busy lives and precious time the opportunity to learn new ideas and insights without having to go through the entirety of big, heavy books.

STEPS OF WORKING

1. Importing the basic libraries .

2. Selecting the file from internet for making summary.

3.Formatting the html file with the help of beautiful soup.

4. Selecting the paragraphs from the html file.

5. Removing the numbers stopwords and excess spaces .

6. Making sentences out of separated tokens.

7. Making frequency distribution chart of all words.

8. Normalising all the counts by dividing by the no having max count.

9. The sentences now having the maximum weightage are now selected and added to the summary.

CODE

import bs4 as bs

import urllib.request

import re

import nltk

import heapq

source = urllib.request.urlopen('https://en.wikipedia.org/wiki/Game\_of\_Thrones').read()

soup = bs.BeautifulSoup(source,'lxml')

text = ""

for paragraph in soup.find\_all('p'):

text += paragraph.text

text = re.sub(r'\[[0-9]\*\]',' ',text)

text = re.sub(r'\s+',' ',text)

clean\_text = text.lower()

clean\_text = re.sub(r'\W',' ',clean\_text)

clean\_text = re.sub(r'\d',' ',clean\_text)

clean\_text = re.sub(r'\s+',' ',clean\_text)

sentences = nltk.sent\_tokenize(text)

stop\_words = nltk.corpus.stopwords.words('english')

word2count = {}

for word in nltk.word\_tokenize(clean\_text):

if word not in stop\_words:

if word not in word2count.keys():

word2count[word] = 1

else:

word2count[word] += 1

max\_count = max(word2count.values())

for key in word2count.keys():

word2count[key] = word2count[key]/max\_count

sent2score = {}

for sentence in sentences:

for word in nltk.word\_tokenize(sentence.lower()):

if word in word2count.keys():

if len(sentence.split(' ')) < 25:

if sentence not in sent2score.keys():

sent2score[sentence] = word2count[word]

else:

sent2score[sentence] += word2count[word]

best\_sentences = heapq.nlargest(5, sent2score, key=sent2score.get)

print('---------------------------------------------------------')

for sentence in best\_sentences:

print(sentence)

APPLICATIONS

**1. Media monitoring**

The problem of information overload and “[content shock](https://contentmarketinginstitute.com/2014/01/content-shock-trend-content-marketing-myth/)” has been widely discussed. Automatic summarization presents an opportunity to condense the continuous torrent of information into smaller pieces of information.

**2. Newsletters**

Many weekly newsletters take the form of an introduction followed by a curated selection of relevant articles. Summarization would allow organizations to further enrich newsletters with a stream of summaries (versus a list of links), which can be a particularly convenient format in mobile.

**3. Search marketing and SEO**

When evaluating search queries for SEO, it is critical to have a well-rounded understanding of what your competitors are talking about in their content. This has become particularly important since Google updated its algorithm and shifted focus towards [topical authority](http://blog.frase.io/artificial-intelligence-in-seo-and-content-optimization/) (versus keywords). Multi-document summarization can be a powerful tool to quickly analyze dozens of search results, understand shared themes and skim the most important points.

**4. Internal document workflow**

Large companies are constantly producing internal knowledge, which frequently gets stored and under-used in databases as unstructured data. These companies should embrace tools that let them re-use already existing knowledge. Summarization can enable analysts to quickly understand everything the company has already done in a given subject, and quickly assemble reports that incorporate different points of view.

**5. Financial research**

Investment banking firms spend large amounts of money acquiring information to drive their decision-making, including automated stock trading. When you are a financial analyst looking at market reports and news everyday, you will inevitably hit a wall and won’t be able to read everything. Summarization systems tailored to financial documents like earning reports and financial news can help analysts quickly derive market signals from content.

**6. Legal contract analysis**

Related to point 4 (internal document workflow), more specific summarization systems could be developed to analyze legal documents. In this case, a summarizer might add value by condensing a contract to the riskier clauses, or help you compare agreements.

**7. Social media marketing**

Companies producing long-form content, like whitepapers, e-books and blogs, might be able to leverage summarization to break down this content and make it sharable on social media sites like Twitter or Facebook. This would allow companies to further re-use existing content.

**8. Question answering and bots**

Personal assistants are taking over the workplace and the smart home. However, most assistants are fairly limited to very specific tasks. Large-scale summarization could become a powerful question answering technique. By collecting the most relevant documents for a particular question, a summarizer could assemble a cohesive answer in the form of a multi-document summary.

**9. Video scripting**

Video is becoming one of the most important marketing mediums. Besides video-focused platforms like YouTube or Vimeo, people are now sharing videos on professional networks like LinkedIn. Depending on the type of video, more or less scripting might be required. Summarization can get to be an ally when looking to produce a script that incorporates research from many sources.

**10. Medical cases**

With the growth of tele-health, there is a growing need to better manage medical cases, which are now fully digital. As telemedicine networkspromise a more accessible and open healthcare system, technology has to make the process scalable. Summarization can be a crucial component in the tele-health supply chain when it comes to analyzing medical cases and routing these to the appropriate health professional.

**11. Books and literature**

Google has reportedly worked on projects that attempt to [understand novels](https://www.theguardian.com/books/2016/sep/28/google-swallows-11000-novels-to-improve-ais-conversation). Summarization can help consumers quickly understand what a book is about as part of their buying process.

**12. Email overload**

Companies like Slack were born to keep us away from constant emailing. Summarization could surface the most important content within email and let us skim emails faster.

**13. E-learning and class assignments**

Many teachers utilize case studies and news to frame their lectures. Summarization can help teachers more quickly update their content by producing summarized reports on their subject of interest.

**14. Science and R&D**

Academic papers typically include a human-made abstract that acts as a summary. However, when you are tasked with monitoring trends and innovation in a given sector, it can become overwhelming to read every abstract. Systems that can group papers and further compress abstracts can become useful for this task.

**15. Patent research**

Researching patents can be a tedious process. Whether you are doing market intelligence research or looking to file a new patent, a summarizer to extract the most salient claims across patents could be a time saver.

**16. Meetings and video-conferencing**

With the growth of tele-working, the ability to capture key ideas and content from conversations is increasingly needed. A system that could turn voice to text and generate summaries from your team meetings would be fantastic.

**17. Help desk and customer support**

Knowledge bases have been around for a while, and they are critical for SAAS platforms to provide customer support at scale. Still, users can sometimes feel overwhelmed when browsing help docs. Could multi-document summarization provide key points from across help articles and give the user a well rounded understanding of the issue?

**18. Helping disabled people**

As voice-to-text technology continues to improve, people with hearing disabilities could benefit from summarization to keep up with content in a more efficient way.

**19. Programming languages**

There have been multiple attempts to build [AI technology that could write code](https://www.futurity.org/artificial-intelligence-bayou-coding-1740702/) and build websites by itself. It is a possibility that custom “code summarizers” will emerge to help developers get the big picture out of a new project.

**20. Automated content creation**

“Will robo-writers replace my job?” That’s what writers are increasingly asking themselves. If artificial intelligence is able to replace any stage of the content creation process, automatic summarization is likely going to play an important role. Related to point 3 (applications in search marketing and SEO), writing a good blog usually goes by summarizing existing sources for a given query. Summarization technology might reach a point where it can compose an entirely original article out of summarizing related search results.

FUTURE SCOPE

The growth of digital information in a decade has led to the problem of information overload. Text analytics for such data presents many new challenges for research and development, and has also gained interest from industry. Automatic text summarization is a well-known solution to the problem of information overload. Text summaries are an essential guide to the users to form an opinion on the relevance of the document. In other words, summaries save time of internet users in their daily work. From literature survey it was observed that most of the existing summarization systems have been built either on statistical approaches or on linguistic approaches. Statistical techniques started with shallow features such as term frequency and gradually extended to positional features and domain-specific thematic features to improve the quality of summary. The statistical techniques were found to be simple and faster in implementation. They worked efficiently with larger documents also. The statistical techniques lacked in semantic analyses of the textual units and thus generated summary that lacked cohesiveness and coherence. The linguistic techniques explore the discourse structure of the document by using semantic analyses of the text. It needs the support of Lexical database to find the relatedness (connectivity) of the textual units. This technique generates cohesive summary as compared to statistical techniques using shallow features. It has high complexity level of implementation as compared to statistical techniques and works slower for large documents. It is not useful for domain-specific summarization as it does not use domain specific features. To achieve the benefits of statistical and linguistic methods a hybrid approach is used to generate a summarization system that uses semantic analysis of document along with important features of textual units in news domain and anaphora feature for resolution of correlation of sentences To study the effect of the hybrid approach, three methods have been implemented separately and tested for the same datasets: • Text summarization using lexical chaining The advantage of this Linguistic method is: It is appealing because it offer perspectives for more semantically and linguistically rich treatment of text for summarization. Lexical chains help to capture all the sentences related to the central theme of the document providing the coverage of the topic and thus cohesive summary is generated. The limitation of this method is: It does not consider domain-specific features and hence cannot be used for domain-specific summarization. It works efficiently for small documents. • Text summarization using fuzzy logic This method uses feature extraction and fuzzy logic for decision module. Fuzzy logic handles the uncertainty and impreciseness in feature extraction. The framework can be extended to handle any number of features without any major changes. Its speed is high and can work efficiently on large documents also. The limitation of this method is the lack of semantic analysis due to which summary generated may not be cohesive. It does not handle correlation of sentences. • Text summarization using LexicalFuzzySum: An hybrid approach that results in an efficient domain-specific text summarization using lexical chaining and fuzzy logic. It uses combination of statistical and Linguistic methods. It also includes anaphora resolution to handle correlating sentences