Augmented Web Browsing

College of Engineering and Informatics

Bachelor of Science (Computer Science & Information Technology)

Project Definition Document

Author:
Conor Gilmartin (19434246)

Academic Supervisor:
Dr Matthias Nickles

Contents

1. Project Overview

1.1. Introduction

I am going to create a web extension that runs script and summarizes text on a webpage for my Final Year Project. The extension will allow the user to select text, and it will embed the summary in a panel on the webpage. The purpose of this is to create an easy text summarization for articles on websites. The text will be processed by a machine learning algorithm that will then return the summarized the text.

1.2. Deliverables

- Project Definition Document - 11th November 2022
- Final Project Report - 1st April 2023
- Project Demonstration and Viva Voce - 3rd – 6th April 2023

1.3. Project Topic Description

> “Web *Augmentation* (WA). Rather than creating a new application, WA builds on top of the rendering of an existing website. In some sense, WA is to the Web what Augmented Reality is to the physical world: layering relevant content/layout/navigation over the existing Web to customize the user experience.” [1]

The contents of a webpage are just a bunch of hypermedia items that are being presented and displayed, the contents are hyperlinked and arranged on the web browser window. The way the world is going tech is constantly evolving and developing therefore machine learning and web browsing also is developing rapidly. There are new techniques and web extensions that help to aid the user with what they want in effectively browsing through web contents.[2]

When the user wants to adapt a web app to their needs that haven't been met by developers, they can use web augmentation to solve this when what they want hasn't been considered. This is because web augmentation allows for the user to modify the page to display what they want by interacting with the user interface that has been loaded on the client side. When people use the internet, they use it for all sorts of reasons such as reading news articles or scavenging the internet for information using search engines.[3]

While doing these things we can be searching multiple things at once because there is a lot of support for completing these tasks. There are however these tools are generally stand-alone search engines or applications. This leads to a lack of continuity from the user when looking through pages and pages of documents. This is because they must switch between search engines or applications which leads to a loss of context and relation between them. The efforts of web augmentation tools help to support a reduction in the need for switching between applications or web pages when doing menial browsing tasks [3]

1.4. Project Topic Relevance

The relevance of this topic has increased throughout the expansion of the internet due to the amount of time people spend using the web. It's become a huge part of people's livelihood and jobs. The benefits of augmented web browsing can be extremely apparent especially for people that spend all day at a desk and on a computer. This can be risky however if you are to run script that might be unsafe and have some profound consequences such as an IP ban, this is a worrying issue. The world is now turning to working from home, so people are spending more time at their desk which means web augmentation has the potential to become more popular.

1.5. Document Structure

The concept of this document is to show the thinking and process behind creating this project from the inception of it. I’m going to throughout the document explain my goals that I am hoping to achieve. The document will also set out the requirements with the outline of tools, software and deliverables I’m hoping to implement throughout the project. The document will be a fantastic way for me to consistently see if my work is on the right track to be completed by the deadlines, this will set milestones for me to hit.

2. Project Objectives

2.1. Core Objectives

Analyze the text on the webpage

Analyze greasemonkey the script web extension and use it.

Build a system which takes text as input, and fetches the summarized text within that selected set.

The text is then analyzed in a number of ways.

The data points I would like to extract from the text include:

- The summarized version
- The sentiment of the text

Analyze text for a user account

I would return the text to the user in some form of integrated side panel into the webpage. Once the text has been fetched, we then analyze it and return the summary based on the percentage of text asked for on the slider by user, I would like to experiment with meta-parameters, depending on the parameters supported by the text summarization backend tool/API I choose to use.

Going to develop a simple user interface for this web extension
The interface will made simple and nice so that the user will know exactly how to use it and the learning curve is cut completely
The results of the users input will then be displayed in a way that avoids overwhelming the user with too much information
I want to first find a way to put the information in a popup box then figure out a way to implement it into the html of the page
A big importance will be placed on the simplicity of the data showed I'm going to try implement a sentiment analysis for this to be displayed with the summary
If the text selected might be quite long we will have to account for large data set returned from the query to be displayed in a aesthetically pleasing manner and understandable way.

2.2. Some Other Objectives

These are some objectives that if I have enough time I will try to implement at the end of the project

I would like to implement a percentage of summary wanted slider meter

Such as they want 10% of the original size of the text selected. Then the script would return the selected amount of summarization they want.
I feel this implementation would add an extra depth to the project because people don't want things predetermined for them, they want to be able to choose. This can solve the problem of either oversimplification or over abundance of info depending on what the user wants to get out of the software. This could be a very crucial implementation for certain users. There are implementations of this in other summarization tools such as text compactor [5]. I need to figure out a way of using this with all the tools available to me.

Develop some form of Integration of displaying the outputted text within the page

This will take quite a bit of tinkering and time on my part to figure out a bug free way of getting the text to be show side by side with the original text. This would help make the user interface look a lot cleaner and more professional. I would have to try and interact with the web pages html and insert it into the page.

Automatically identify summarizable text on the webpage

This would be a way of the extension automatically taking the text from the web page, without the need for the user to manually select text it would tell the user this text can be summarized.

2.3. Technical approach

This can be done in lots of ways

● Greasemonkey[7]

○ this is a user script manager that's an extension on browsers. It allows the user to work with scripts that make changes to the web page before or after it is loaded. It uses html and JavaScript.

● NLTK (natural language toolkit)[8]

○ This software library is great for getting a summary and the sentiment The benefits of picking NTLK is that its open source and completely free. The disadvantages NLTK is quite complicated and it has a harsh learning curve with some internal limitations.

● TensorFlow[9]

○ This is a machine learning open source software library. That has JavaScript library. Which is unique because all the other natural language processing apis use python. The disadvantage of using it would be that it's got a difficult learning curve.

● MeaningCloud Summarization [10]

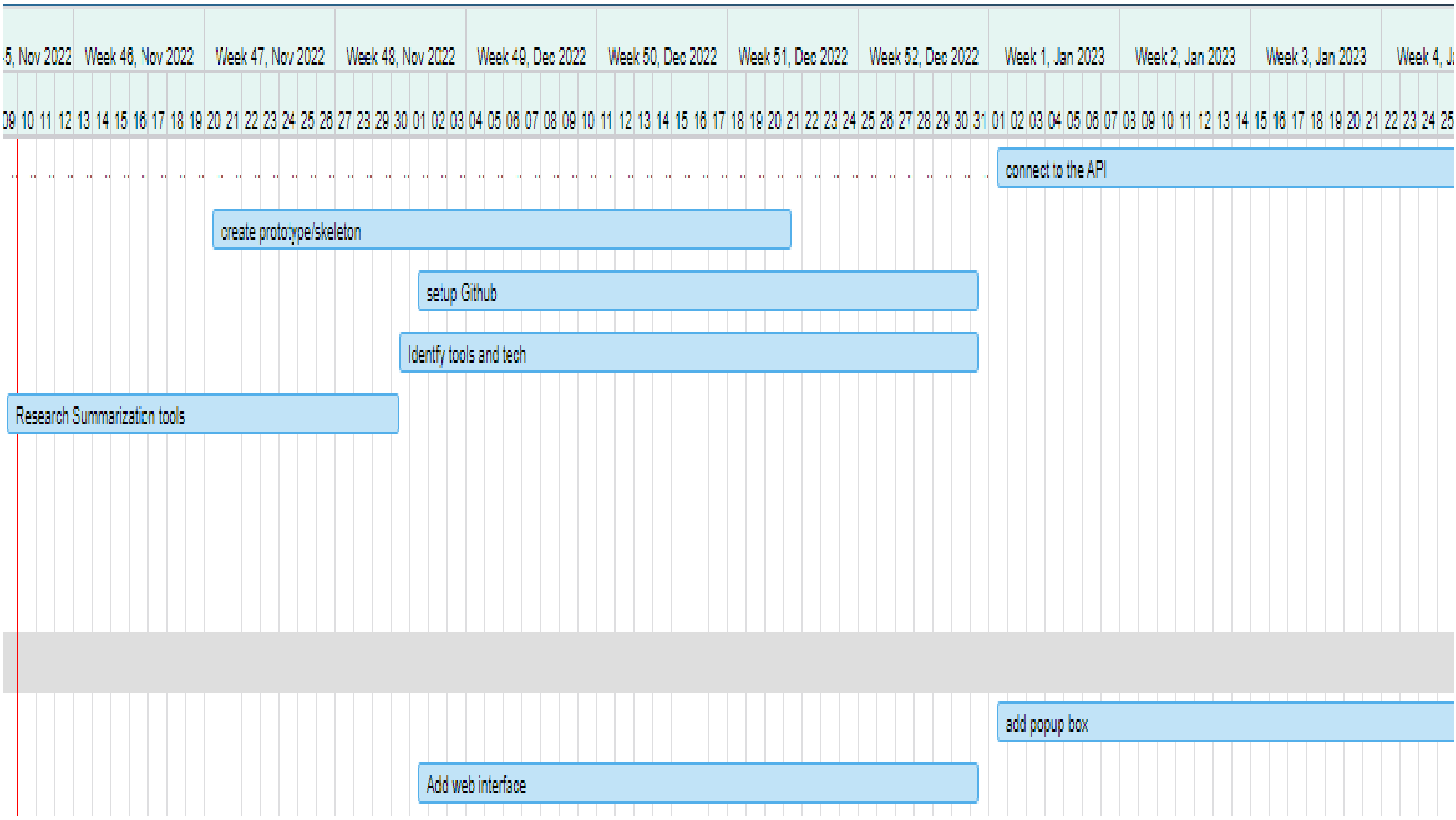
○ This is a summarization API that I can use to get the summary of the text. The learning curve is extremely easy to use and get up and running The disadvantage is that you only get 40,000 requests free a month limitation.

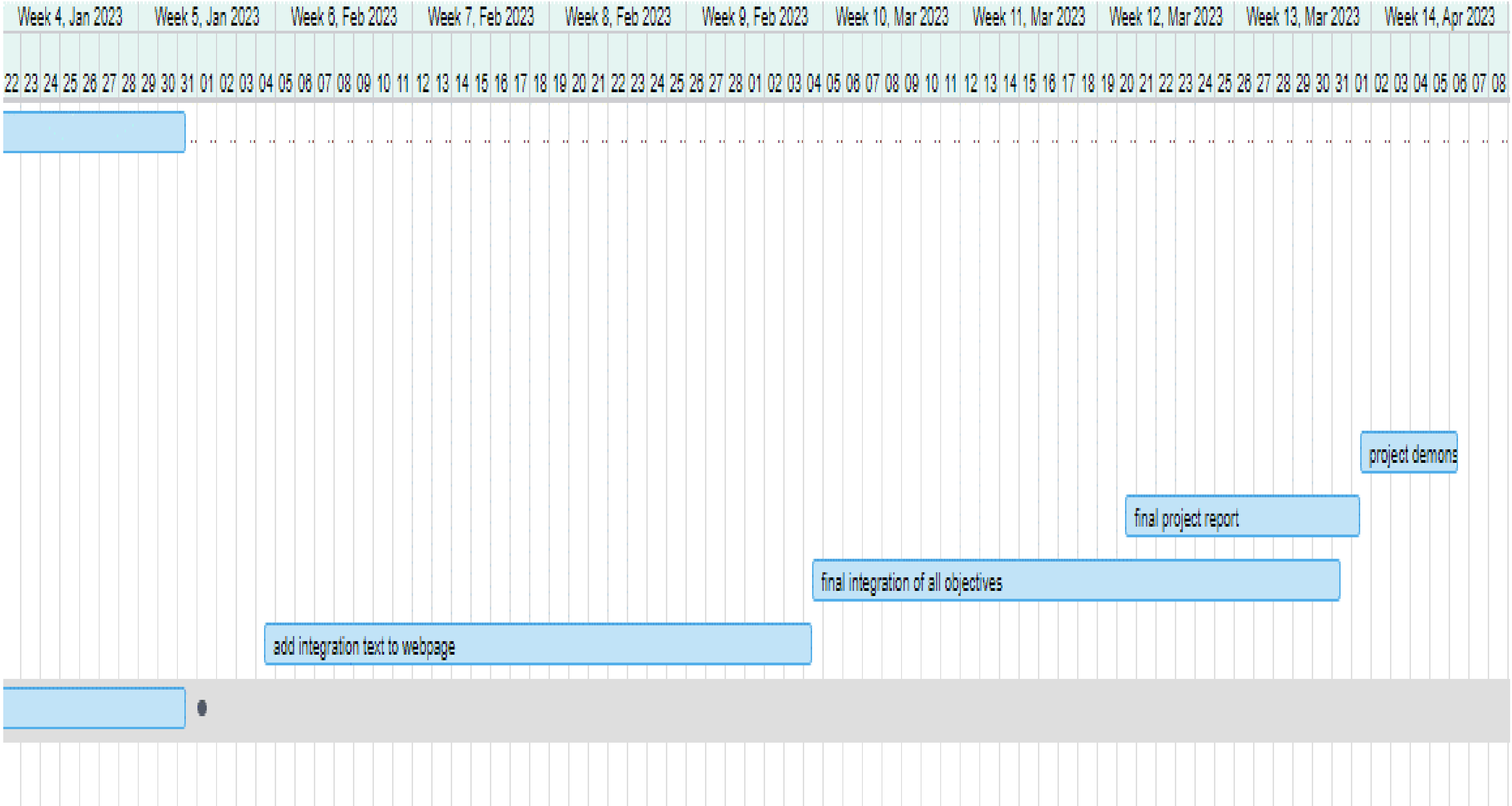
3. Research

When I began planning my project, I went searching for information on this topic by looking at the research that's been done in this subject. One of the most appropriate research papers for my project was completed by Oscar Díaz from University of the Basque Country, he published the research paper entitled "Understanding Web Augmentation" [1]. He explains web augmentation and what it is, He also explains the uses for it and what he would like to see develop within the community within the future. He calls for a development within WA for a "a set of good practice" [1]. Another interesting paper I found in researching this topic of web summarization "Web-page summarization using clickthrough data" [6], and "Visualization to Support Augmented Web Browsing" [2]. I have a lot of interest in this topic as someone who has been doing a lot of research throughout my degree. The idea of augmented web browsing always interested me as topic because I never fully understood why It isn't more popular as everyone uses adblocker these days as a form of extension which is a form of web augmentation. This tells me that there is a lot of potential within this field to be explored. Within my research of the summarization tools most are independent websites which is an inconvenience. This shows me that there is an effective use for my project. A lot of summarization apis are quite limiting, therefore I will have to extend my research throughout the coming weeks to figure out which one to use. Though MeaningCloud Summarization was easy to use for the prototype.

4. Planning and timeline

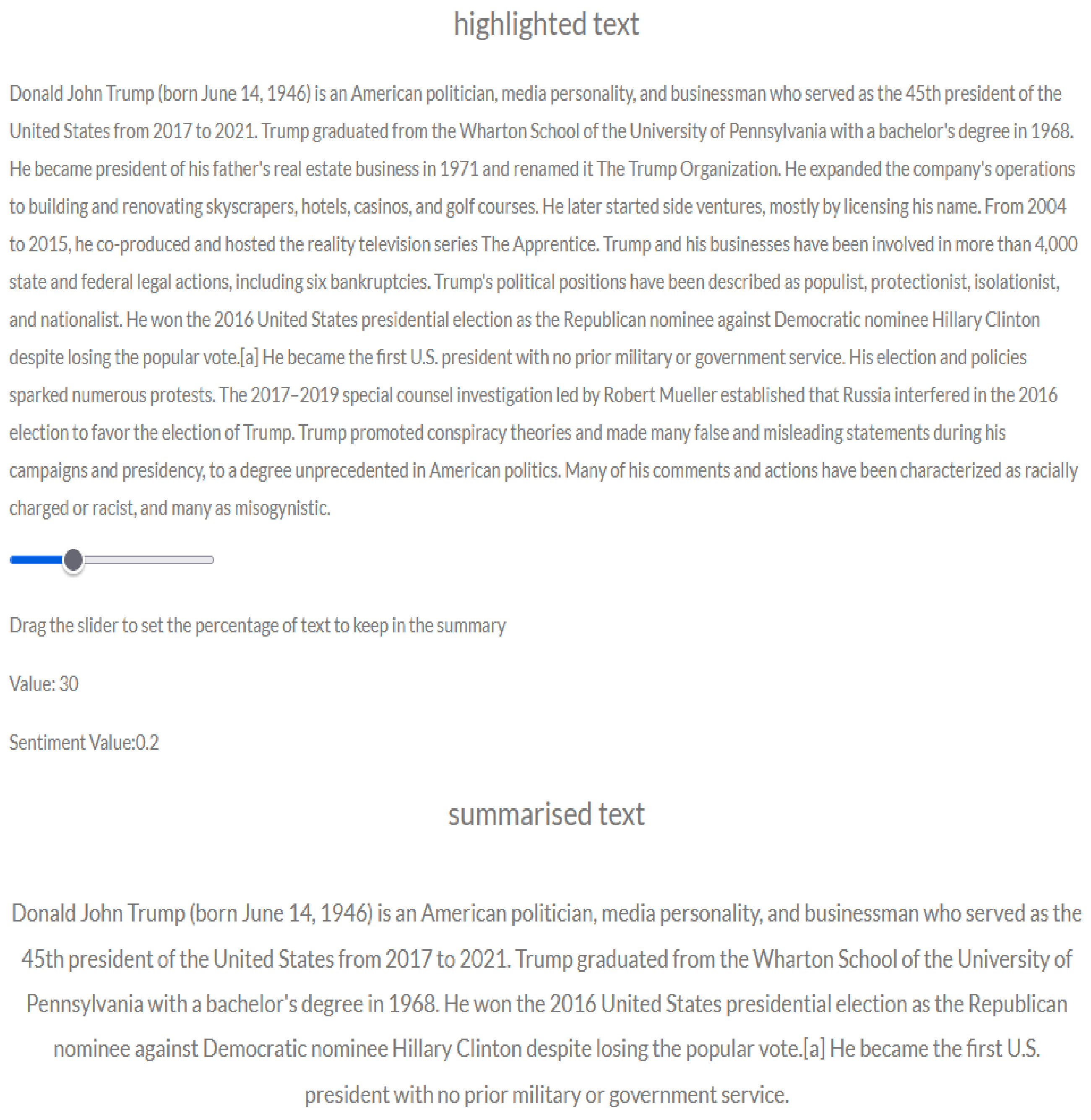
The key for starting a project is planning and making sure everyone hits their goals. Planning is necessary for identifying what parts must be completed first while minimizing risks and trying to hit all the goals and deadlines. This leads to a greater overall success for the project. I am using Microsoft planner to track my progress and set goals. This is necessary to keep track of things that are in progress or finished. I'm using Gant charts to track the weeks leading up to the final report and demo. This is extremely critical that I hit my targets of each section in the given timeframe to keep on track to finish the project on time.


5, Nov 2022
Week 46, Nov 2022
Week 47, Nov 2022
Week 48, Nov 2022
Week 49, Dec 2022
Week 50, Dec 2022
Week 51, Dec 2022
Week 52, Dec 2022
Week 1, Jan 2023
Week 2, Jan 2023
Week 3, Jan 2023
Week 4, J
09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
connect to the API
create prototype/skeleton
setup Github
Identfy tools and tech
Research Summarization tools
add popup box
Add web interface


Week 4, Jan 2023
Week 5, Jan 2023
Week 6, Feb 2023
Week 7, Feb 2023
Week 8, Feb 2023
Week 9, Feb 2023
Week 10, Mar 2023
Week 11, Mar 2023
Week 12, Mar 2023
Week 13, Mar 2023
Week 14, Apr 2023
22 23 24 25 26 27 28 29 30 31 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 01 02 03 04 05 06 07 08
project demons
final project report
final integration of all objectives
add integration text to webpage

5. User Interface Mockup

This is a quite simple mockup of what it would look like when the text is highlighted on the webpage.



highlighted text

Donald John Trump (born June 14, 1946) is an American politician, media personality, and businessman who served as the 45th president of the United States from 2017 to 2021. Trump graduated from the Wharton School of the University of Pennsylvania with a bachelor's degree in 1968. He became president of his father's real estate business in 1971 and renamed it The Trump Organization. He expanded the company's operations to building and renovating skyscrapers, hotels, casinos, and golf courses. He later started side ventures, mostly by licensing his name. From 2004 to 2015, he co-produced and hosted the reality television series The Apprentice. Trump and his businesses have been involved in more than 4,000 state and federal legal actions, including six bankruptcies. Trump's political positions have been described as populist, protectionist, isolationist, and nationalist. He won the 2016 United States presidential election as the Republican nominee against Democratic nominee Hillary Clinton despite losing the popular vote.[a] He became the first U.S. president with no prior military or government service. His election and policies sparked numerous protests. The 2017–2019 special counsel investigation led by Robert Mueller established that Russia interfered in the 2016 election to favor the election of Trump. Trump promoted conspiracy theories and made many false and misleading statements during his campaigns and presidency, to a degree unprecedented in American politics. Many of his comments and actions have been characterized as racially charged or racist, and many as misogynistic.

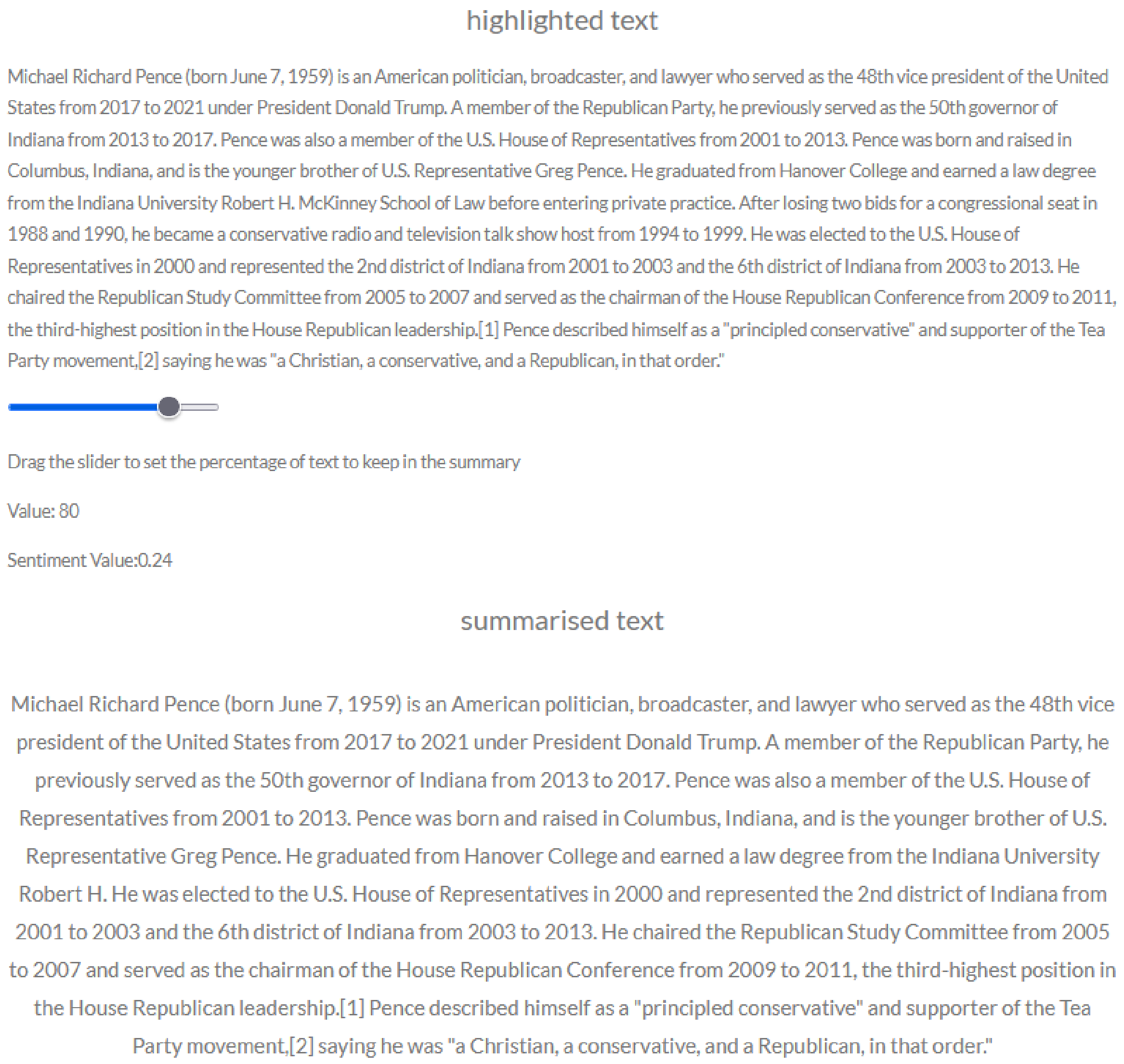
Drag the slider to set the percentage of text to keep in the summary

Value: 30

Sentiment Value:0.2

summarised text

Donald John Trump (born June 14, 1946) is an American politician, media personality, and businessman who served as the 45th president of the United States from 2017 to 2021. Trump graduated from the Wharton School of the University of Pennsylvania with a bachelor's degree in 1968. He won the 2016 United States presidential election as the Republican nominee against Democratic nominee Hillary Clinton despite losing the popular vote.[a] He became the first U.S. president with no prior military or government service.



highlighted text

Michael Richard Pence (born June 7, 1959) is an American politician, broadcaster, and lawyer who served as the 48th vice president of the United States from 2017 to 2021 under President Donald Trump. A member of the Republican Party, he previously served as the 50th governor of Indiana from 2013 to 2017. Pence was also a member of the U.S. House of Representatives from 2001 to 2013. Pence was born and raised in Columbus, Indiana, and is the younger brother of U.S. Representative Greg Pence. He graduated from Hanover College and earned a law degree from the Indiana University Robert H. McKinney School of Law before entering private practice. After losing two bids for a congressional seat in 1988 and 1990, he became a conservative radio and television talk show host from 1994 to 1999. He was elected to the U.S. House of Representatives in 2000 and represented the 2nd district of Indiana from 2001 to 2003 and the 6th district of Indiana from 2003 to 2013. He chaired the Republican Study Committee from 2005 to 2007 and served as the chairman of the House Republican Conference from 2009 to 2011, the third-highest position in the House Republican leadership.[1] Pence described himself as a "principled conservative" and supporter of the Tea Party movement,[2] saying he was "a Christian, a conservative, and a Republican, in that order."

Drag the slider to set the percentage of text to keep in the summary

Value: 80

Sentiment Value:0.24

summarised text

Michael Richard Pence (born June 7, 1959) is an American politician, broadcaster, and lawyer who served as the 48th vice president of the United States from 2017 to 2021 under President Donald Trump. A member of the Republican Party, he previously served as the 50th governor of Indiana from 2013 to 2017. Pence was also a member of the U.S. House of Representatives from 2001 to 2013. Pence was born and raised in Columbus, Indiana, and is the younger brother of U.S. Representative Greg Pence. He graduated from Hanover College and earned a law degree from the Indiana University Robert H. He was elected to the U.S. House of Representatives in 2000 and represented the 2nd district of Indiana from 2001 to 2003 and the 6th district of Indiana from 2003 to 2013. He chaired the Republican Study Committee from 2005 to 2007 and served as the chairman of the House Republican Conference from 2009 to 2011, the third-highest position in the House Republican leadership.[1] Pence described himself as a "principled conservative" and supporter of the Tea Party movement,[2] saying he was "a Christian, a conservative, and a Republican, in that order."

6. Constraints

The constraints that I will have to worry about when in progress with the project

- Hardware Issues – This could stop the project progress by missing deadlines.
- Tools– Having the correct tools to implement the necessary features that I want to add.
- Task – This is a big issue of trying to stay on course with the task a hand , rather then looking on a macro level sometimes just doing the task right now. The biggest issue will be trying to do everything at one.
- Risks – The biggest risk is trying to focus on the project while undertaking exams and assignments. This can lead to running out of time to get the project done
- Quality – This is a big point of emphasis to ensure the extension is of a high standard and quality such as the user interface

References

[1] Díaz, O. (2012). Understanding Web Augmentation. In: Grossniklaus, M., Wimmer, M. (eds) Current Trends in Web Engineering. ICWE 2012. Lecture Notes in Computer Science, vol 7703. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-35623-0_8

[2] D. Q. Nguyen and H. Schumann, "Visualization to Support Augmented Web Browsing," 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013, pp. 535-541, doi: 10.1109/WI-IAT.2013.75.

[3] Diego Firmenich, Sergio Firmenich, Gustavo Rossi, Manuel Wimmer, Irene Garrigós, César González-Mora,Engineering Web Augmentation software: A development method for enabling end-user maintenance,Information and Software Technology,Volume 141,2022,106735,ISSN 0950-5849,https://doi.org/10.1016/j.infsof.2021.106735. (https://www.sciencedirect.com/science/article/pii/S0950584921001853)

[4] Niels Olof Bouvin. 1999. Unifying strategies for Web augmentation. In Proceedings of the tenth ACM Conference on Hypertext and hypermedia : returning to our diverse roots: returning to our diverse roots (HYPERTEXT '99). Association for Computing Machinery, New York, NY, USA, 91–100. https://doi.org/10.1145/294469.294493

[5] https://www.textcompactor.com/

[6] Jian-Tao Sun, Dou Shen, Hua-Jun Zeng, Qiang Yang, Yuchang Lu, and Zheng Chen. 2005. Web-page summarization using clickthrough data. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05). Association for Computing Machinery, New York, NY, USA, 194–201. https://doi.org/10.1145/1076034.1076070

[7]greasemonkey, https://www.greasespot.net/

[8]NTLK, https://www.nltk.org/

[9] TensorFlow, https://www.tensorflow.org/

[10] meaningcloud, https://www.meaningcloud.com/