**AI Project Report**

**DAVE3625**

Dennis Johnsen

Alexander Paulsen

**Project Report: Development of a Language Model AI Chrome Extension for Summarizing Web Browsing Sessions**

Introduction

Problem Context

In the digital age information comes from every angle. This often leads to big information overload for people. It can be hard to remember and summarize your time on the internet after each session. This being articles, new topics, research or other things. In this case it could be beneficial to save or summarize the activity in a short form summary. This could help users to recall crucial information they might have forgot, but also help users to track their thoughts back to the original place. This can be used in research, school or work.

Problem Occurrence

The problem generally affects students, researchers or professionals that rely on many internet sources for gathering information. As a student group we know how stressful it is when surfing the web about a topic, but forgetting it minutes later. The reason for this is our brains is very limited in what it can store. Multitasking and memory isn’t a very good combination. When trying to remember many things at once our brains usually picks some of the information and lays off the rest. To counteract this, we need an effective and automated solution to make the information easy to recall and summarize from each session.

Outcomes

This solution will provide:

* Better information retention
* Higher productivity
* Save time

This solution can have:

* Privacy concerns

Business opportunity

By using this solution, we can monetize it as a productivity tool for the chrome browser. Where we can make it subscription-based and make student and professional subscriptions. We can also make it free to use but use donations as monetization. Another example can be making it public and getting help from collaborators to make the product even better. By the last idea we can scale it to help many more people.

Resources

Data Availability

The data for this project is usually unstructured. It is for the most part just text that is not labelled on the websites. The extension will collect the text data from the web pages. It will have to go through the data and get context for it since the data is unlabelled. Challenges with the collecting is to ensure privacy of the users and how it handles multimedia and dynamic content. When collecting the data, the summary depends on how much data is collected. It is important that the data collected is not too much, so the summary gets too long. Its therefore important to implement a type of evaluation that limits how much is summarized from each site for each session. Another problem we will encounter is ads, navigation elements and other roadblocks on a website. It is very important that we can provide clean data to the model. We should not mistakenly provide ads to the model.

More Challenges

It is important that the data is anonymized when needed and user consent is ensured. It is also important that dynamic data is considered. If the website frequently changes content, it might be smart to provide dates for the data collection. The cost for the solution is mostly development costs whereas the data is already free and out there.

Method

Machine Learning Problem Type

For making this model we would probably go for a mix of Natural Language Processing tasks with the main focus on summarizing and topic modeling. Important things we need to account for is determining headings, main content, important keywords. It is also important that we determine what is deemed as unnecessary as ads, popups and other elements in HTML/CSS. To make the processing model we need to have a way of chopping up all words. When the data is cleaned, and we only have the raw text we can start tokenizing. By breaking the text into words, letters, or phrases we can start the process of modeling. We will be using a transformer-based model to do this. Probably a model based on the way GPT would work. It predicts the next word after the other. This model though requires a big database with all the relevant words. For testing we could get users to give feedback for the summary quality. We can also use some labelled datasets as well to make the learning process of the language model better.

Api

Another problem we are facing is operating the model. When feeding data or using the model the computer works very intensively. Running the ai on the personal computer would be too hard of a task. What we could do instead is use an api based service for the ai. The chrome extension will automatically call the model on a remote supercomputer that does all the computation. Then it gets sent back to the browser extension. This will cost significantly more, but the performance will be much better for the user’s computer.

One good way to make the AI fast and productive is to generate summaries of the most visited websites. Then when the extension sees the session history url it just calls for the summary. This can work for high traffic websites where the content is not dynamic.

Example

Step-by-Step Example

When using this extension, we want it to be as automated as possible. This means from starting the extension to getting the summary. Less steps usually equates to easier user availability. In this part we will go through a step-by-step example and explain how easy our extension should be.

User: User starts up the browser. The extension is already on as default. The user navigates to a website.

Extension: The extension gets the url for the website and sends an api request to the supercomputer with the language model. The url for the website might already have the url summary stored or it will have to go through the model and create a new summary. The summary will then be stored by the supercomputer for later use.

User: User gets the summary when the user ends the session. The user can decide between reading it in the browser or download the file.

Conclusion

This AI powered extension offers user a new way of productivity and memory-retention. To counteract the information overload in today’s digital age we have to think smarter. Therefore, using AI tools like this extension to summarize web browsing session will help both the younger generation, but also older people that might struggle to remember what they actually spent their time on the web for. There are still many challenges with developing such a big extension. The AI in itself being the most complicated, but privacy and data cleaning also poses problems for the extension. By getting User-feedback and training the AI continuously we will be able to develop a good product.