In-browser Text Importance Detector
BSP - Phase I deliverable

Monday 28th March, 2022 - 12:04

Paulo Ricardo Botelho Oliva
University of Luxembourg
Email: paulo.botelho.001@student.uni.lu

This deliverable has been produced under the supervision of:
Luis Leiva
University of Luxembourg
Email: luis.leiva@uni.lu

1. Project main objectives

The Bachelor Semester Project is about finding important sentences in a text. The main objectives of this project are to create a browser extension that highlights the most important information of an article's webpage. Moreover, this project addresses the topic of automatic text summarization, as well as extractive text summarization algorithms.

2. Prerequisites

The only required competency before starting to work on this Bachelor Semester Project is possessing knowledge of the JavaScript programming language.

3. Scientific Deliverable Description

The following questions should be answered in the scientific deliverable of this project:

How can we automatically detect text blocks (groups of paragraphs) on a web page?

How can we identify the most important sentences within a text block?

In the scientific deliverable of this project, the answer to these two question is given in more detail.

Before answering these questions, the concept of automatic text summarization needs to be explained. The technique of computationally shortening a text to generate a summary that contains the most significant or relevant information within the original content is known as automatic text summarization.

In general, there are two main kinds of text summarization techniques, extractive text summarization and abstractive text summarization. Extractive text summarization consists in selecting the most important parts of a text and extracting them to create a shorter summary of the same text. Abstractive text summarization consists in gathering the main ideas of a text and using those abstract ideas to generate a summary of the text, where the summary isn't just a copy of the old text, but actual new sentences.

The main issue with the first question is that automatic text detection is not reliable. Since all webpages are different, detecting blocks of text can be a hard task. For that reason, this BSP focuses on reliably detecting the main text of a WordPress webpage.

Basically, WordPress structures its webpages using a predefined hierarchy based on themes and common class names for the HTML elements. Thus, to reliably be able to detect text blocks on a WordPress webpage, it suffices to find the most common classes used on WordPress to define the main content of the webpage.

As for the second question, to identify the most important sentences within a text block, extractive text summarization techniques can be used. To answer this question more thoroughly, the topic of extractive text summarization needs to be addressed more in depth and different extractive text summarization algorithms are studied.

4. Technical Deliverable Description

The technical deliverable of this project focuses on creating a browser extension that detects the most important sentences on a specific web page. The extension is to be built using JavaScript and should work in both Google Chrome and Firefox browsers, without breaking the functionality of existing websites.

The extension should allow the user to automatically highlight the most important sentences of an article in a specific webpage, if that webpage is built using WordPress. If the extension cannot find text or the page was not built using WordPress, the extension allows the user to select a block of text from the webpage to highlight text from. A

JavaScript text summarization library called Summary.js is used to obtain the most important sentences from a block of text using an extractive text summarization algorithm.

The most essential functions of the extension are: a function that returns the main text content in a WordPress webpage, a function that generates a text summary using Summary.js, and a function that highlights the text in a webpage based on the sentences contained in the summary generated previously. For the first function, the querySelectorAll() is used to find all the text paragraphs matching certain classes based on different WordPress page content class definitions.

The point of the technical work is to implement text summarization techniques in a browser environment, while creating a simple extension that is easy to use and that works on most of the WordPress websites.

References

[BiCS(2021)] BiCS Bachelor Semester Project Report Template. https://github.com/nicolasguelfi/lu.uni.course.bics.global University of Luxembourg, BiCS - Bachelor in Computer Science (2021).

[BiCS(2021)] Bachelor in Computer Science: BiCS Semester Projects Reference Document. Technical report, University of Luxembourg (2021)

[Armstrong and Green(2017)] J Scott Armstrong and Kesten C Green. Guidelines for science: Evidence and checklists. *Scholarly Commons*, pages 1–24, 2017. https://repository.upenn.edu/marketing_papers/181/