In-browser Text Importance Detector
BSP - Phase I deliverable

Monday 28th March, 2022 - 20:26

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 enough knowledge about Web Technologies, such as HTML and the JavaScript programming language.

3. Scientific Deliverable Description

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

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

How can we automatically detect text blocks on a web page?

To answer the first question, the concept of automatic text summarization needs to be explained in the scientific deliverable. Automatic text summarization is the technique of creating a shorter version of a text using a computer. This shorter version of the text, the summary, contains the most relevant information of the original text.

Automatic text summarizing methods are desperately needed to deal with the ever-increasing volume of text data available online, both to help identify relevant information and to consume relevant information more quickly. There are many reasons why one would need automatic text summarization. For example, summaries reduce reading time and should be less biased than summaries made by a human person, according to [1].

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 scientific deliverable of this project focuses on extractive text summarization, as it is the technique we use in the technical deliverable. There exist many approaches for extractive text summarization, like statistical approaches, lexical chain based approaches, graph-based approaches, cluster-based approaches and fuzzy logic based approaches.

Moreover, to answer the first question, we are focusing on graph-based text summarization, as it commonly used by many algorithms to classify text in order of importance. For example, Google's PageRank is a graph-based ranking algorithm used by Google to rank webpages on Google. This algorithm is the basis of other well-known graph-based text summarization algorithms, like TextRank and LexRank.

The graph-based method of text summarization is an unsupervised technique in which sentences or words are scored using a graph. The basic goal of the graph-based method is to extract the most essential sentences from a single source. Thus, the scientific deliverable focuses on explaining how graph-based algorithms work to identify the most important sentences in a block.

Then, to answer the second question, the notion of text blocks should be explained. Text blocks in a webpage are simply multiple paragraphs grouped together. Usually, one would want to select the most important text blocks in a webpage, the main content, and not unimportant text context, such as the text in the header or footer of a webpage.

The main issue with the second question is that automatic text detection is not reliable. Since all webpages are different, detecting blocks of text can be a hard task, as different websites have different structures. The scientific deliverable of this project studies this question more thoroughly as there does not seem to be a trivial solution to this question.

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.

A browser extension is a small software package that adds features and functions to a browser. Extensions are built using well-known web technologies such as HTML, CSS, and JavaScript. Extensions possess their own set of APIs varying with the browser, which, for example, enables them to change the browser's default behaviours or access a webpage's content.

The extension should be built using JavaScript and should work on Google Chrome browsers, without breaking the functionality of existing websites. The extension should allow the user to automatically highlight the most important sentences of an article or blog post in a specific webpage.

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

[1] J.-M. Torres-Moreno, *Automatic text summarization*. John Wiley & Sons, 2014.

[2] BiCS(2021), "BiCS Semester Projects Reference Document," University of Luxembourg, Tech. Rep., 2021.

[3] ——, "BiCS Bachelor Semester Project Report Template," University of Luxembourg, BiCS - Bachelor in Computer Science, Tech. Rep., 2021. [Online]. Available: https://github.com/nicolasguelfi/lu.uni.course.bics.global