Big Data
Final Project

Authors:
Alexandru Ciutacu - s1032974

June 29, 2023

IN·DEI·NOMINE·FELICITER

Radboud Universiteit

Contents

1 Introduction and Motivation

By combining word count and page rank algorithms, the project's aim is to analyze online content in a novel way. The main objective is to identify, count, and rank terms associated with the fitness world. This analysis has potential applications in several areas such as crafting more effective SEO strategies for marketers and understanding the trending topics and concerns in the fitness industry for fitness enthusiasts. In this report, I aim to highlight the top ten most frequently used fitness-related terms.

Initially, the approach was straightforward and swift: extract all text from a web-page and concentrate on counting terms linked to the fitness industry. However, this approach had to be modified as the extensive amount of data present proved too much. As a result, a smaller sample of 50 Web ARChive (WARC) files from a single segment was utilized as the data-set.

The inspiration for this project came from my personal passion for fitness and health. As an avid fitness enjoyer, I was intrigued to see which websites incorporated the most fitness-related terminology. To do this, I created a list of terms I thought were often used in this field. The aforementioned terms served as the framework for the analysis.

The keywords included:

- Fitness
- Exercise
- Workout
- Nutrition
- Health
- Wellness
- Weight Loss
- Strength Training
- Cardio
- Bodybuilding

Looking back, I considered adding more keywords to the list, but I was concerned about the time and computational resources required to process such a large amount of data. As the project progressed, these initial worries about time and computational complexity proved true.

In summary, this project represents a unique exploration of the fitness landscape online. The project elements of information retrieval, machine learning, and natural language processing, thus providing a valuable opportunity to better understand the use of AI-related terms in digital content.

2 Project Methodology and Process

In this project, the methodology was represented by iterative development, continuous refinement, and a focus on optimization. I first tried to understand the data at a low level before scaling to the entire data-set, thus facilitating effective debugging and optimization

2.1 Initial Analysis on a Single WARC File

To better understand the data structure and content, I started by processing a single WARC file present in **hdfs dfs -ls /single-warc-segment**. I got help from one of the TAs, namely Leah, who provided me with two such files as I encountered challenges downloading these files from the cluster. Following this, I transferred these files to my big-data container to take advantage of the Zeppelin notebook environment for subsequent steps. To accomplish this task I utilized the **docker cp /opt/hadoop/rubigdata/CC-MAIN-20210412053559-20210412083559-00387.warc.gz 729e80693375:/** command

After I initially explored the data analysis of the two WARC files, I gained a solid understanding of the data structure and its limitations. Because of this, I was able to create a parsing and analysis strategy for the HTML content. Recognizing a pattern where fitness-related terms predominantly appeared within the webpage bodies, I shifted my focus to parse specifically this portion rather than the entire HTML content. With this knowledge obtained, I proceeded to evaluate the efficacy of Spark and Jsoup for data parsing. This stage of the process was dedicated to grasping the HTML structure, extracting the body content, and implementing basic keyword matching to count the occurrence of fitness-related terms.

2.2 Parsing Optimization

After parsing the entire HTML content, the project encountered memory issues due to the size and complexity of the data. To address this, I refined the parsing strategy in order to remedy the aforementioned situation. This was possible due to the focus shift to the body content of web pages, as it is typically here that the main content is found. Based on this modification, the memory footprint was greatly reduced.

2.3 Keyword Search and Counting

The next crucial step involved defining a regular expression pattern to match the fitness-related keywords identified earlier in the process. Using this pattern, I was able to search the body contents of the web pages and count the occurrences of each keyword. This provided an initial indicator of a webpage's relevance to the fitness industry, under the assumption that a higher frequency of fitness-related terms signals a stronger alignment with the fitness context.

2.4 Ranking Webpages by Relevance

After the previously mentioned step was complete, the next goal was to rank the web pages according to their relevance to the fitness domain. Initially, I wanted to achieve this via the **sortyBy()** function, but the performance was found to be suboptimal. The sorting operation was later optimized by using SparkSQL, which significantly reduced the runtime and improved performance. It is important to note that the primary performance metric for this project was the frequency of fitness-related terms on a webpage. This straightforward approach provides a good initial estimate of relevance, but one of its limitations is that it does not take into account the context in which the terms are used. As a consequence, a high keyword count does not always equate to a high relevance to fitness, which is a consideration for further improvement of the methodology.

2.5 Memory Management

As the complexity of the operations increased, the project faced memory-related challenges. Despite my best efforts, I couldn't rectify the issue when working on 2 WARC files in the Zepellin notebook. Fortunately, the professor, Arjen, provided me with a crucial observation, namely the big-data Docker container might be running out of memory as it wasn't allocated enough memory by default. His input was invaluable and proved to be the missing piece I needed to address the issue. Prior to this observation, I had spent a substantial amount of time attempting to debug the problem. Consequently, to remedy these issues, I fine-tuned the Spark configuration. Those remedies consist of:

- Optimizing the memory allocation for the driver and executor
- Enabling offHeap memory
- Setting the Kryo serialization classes

Other improvements to the code:

- removed .persist() as it was deemed troublesome from urlWordCount.
- Tried to pass all the warcFiles as a comma-separated string to newAPIHadoopFile instead of using a var in a for loop
- Used the take method directly on sql_query to get the top five records

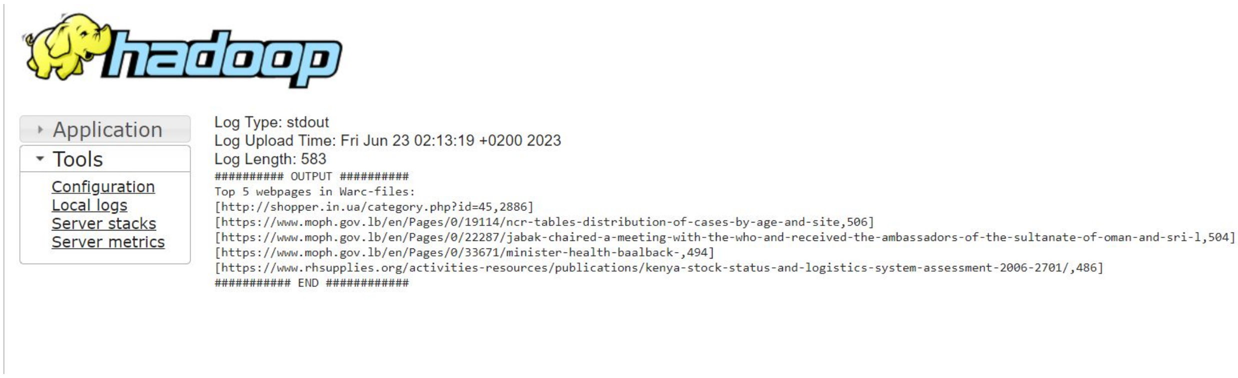
This resulted in improved memory usage and performance. In addition, this helped prevent the premature termination of executors.

2.6 Scaling to Multiple WARC Files

After I successfully implemented the code on a single WARC file, the next step constituted of scaling the project to process multiple WARC files on the cluster. However, this task wasn't as straightforward as it first seemed. Since my project relied on external libraries, namely Jsoup and the HadoopConcatGz package for reading the WARC files, I needed to create a new JAR using **sbt assembly**. This step was critical in ensuring that all necessary dependencies were included in the JAR file. Once I assembled the new JAR, I had to make sure to submit the correct one to the Spark cluster. I did this using the command **spark-submit –deploy-mode cluster –queue gold target/scala-2.12/RUBigDataApp-assembly-1.0.jar**. Fortunately, and in part due to my testing on a single WARC file, I did not encounter any issues when I scaled my project up.

3 Analysis & results

The project successfully processed 50 WARC files from the specified segment. The goal was to identify and rank web pages based on the frequency of fitness-related terms. The list of terms was derived from commonly used fitness-related terminology. The top 5 web pages with the highest count of fitness-related terms were found on the redbad website. The output may be observed in the following figure



```
Log Type: stdout
Log Upload Time: Fri Jun 23 02:13:19 +0200 2023
Log Length: 583
########## OUTPUT ##########
Top 5 webpages in Warc-files:
[http://shopper.in.ua/category.php?id=45,2886]
[https://www.moph.gov.lb/en/Pages/0/19114/ncr-tables-distribution-of-cases-by-age-and-site,506]
[https://www.moph.gov.lb/en/Pages/0/22287/jabak-chaired-a-meeting-with-the-who-and-received-the-ambassadors-of-the-sultanate-of-oman-and-sri-l,504]
[https://www.moph.gov.lb/en/Pages/0/33671/minister-health-baalback-,494]
[https://www.rhsupplies.org/activities-resources/publications/kenya-stock-status-and-logistics-system-assessment-2006-2701/,486]
########### END ############
```

Figure 1: Ouput from the 50 WARC files

Notably, the page with the highest count was shopper.in.ua, with an overwhelming 2886 mentions, while the subsequent four pages had counts in the 400-500 range.

4 Conclusion

This analysis has offered some intriguing insights. However, the top results might not reflect what one might expect for pages with the highest frequency of fitness-related terms. This discrepancy could be due to the use of fitness-related terms in contexts outside of the fitness and health industry.

For instance, a few of the pages in the top results are from the Ministry of Public Health (MOPH) in Lebanon, which might be using these terms in a health-related context rather than specifically a fitness one.

In terms of improvements for future work, it might be beneficial to refine the regular expression pattern used for matching fitness-related terms. For example, the current pattern matches terms irrespective of their context, which can lead to high counts on pages that might not necessarily be focused on fitness. The inclusion of additional keywords might also be useful, as well as a method for evaluating the context in which a keyword is used.

Another point to note is that the tool was limited to processing 50 WARC files due to constraints in time as I took longer than expected to finish the assignment as previously discussed with the memory issues related to Zepellin. Ideally, with more time, this limitation could be overcome which could potentially lead to more accurate and representative results as more WARC files would be analyzed.

Lastly, memory management proved to be an essential part of this project. The initial problems encountered during the project underscored the importance of allocating sufficient memory for big data processing. The optimization of the memory allocation for the driver and executor, enabling offHeap memory, and setting the Kryo serialization classes facilitated more efficient processing of the data.

Overall, this project provided valuable insights into big data processing and demonstrated the applicability of various tools and techniques in analyzing web content. For future work, I aim to enhance the precision of results by incorporating NLP techniques to understand the context of term usage. This could help identify truly fitness-focused web pages rather than those with high counts of fitness-related terms.