Capstone Project Summary

Analysis of SEO data at BrokerChooser and competitors:

Similarity and Google updates

Alima Dzhanybaeva

June 2023

Content

1. Introduction

The client of my project is an independent online discovery and comparison platform that thrives to help independent investors and traders to find a broker fitting to their needs. The company reviews top-tier regulated, well-established players in the financial industry, focusing on brokers.

The aim of this project is to investigate the similarity of web pages on the client's and its competitors' websites. This investigation holds a great significance as search engines strive to present users with the most relevant and distinct content. When webpages contain excessive similarities or identical material, search engines may perceive it as duplicate content. Consequently, the presence of numerous similar webpages can lead to penalties that adversely impact a website's search engine rankings. Understanding the level of similarity between webpages of BrokerChooser and competitors is crucial for optimizing content strategies, avoiding penalties, and maintaining strong visibility and rankings within search engine results.

2. Data and Methodology

2.1. Data

To perform a comparative analysis of the client's web pages five competitors were selected. The choice is justified by the fact that these websites were found to be competing with the client for the same search engine results page (SERP).

The content from all websites was scraped using Python. To ensure accurate similarity analysis, I excluded headers and footers that were present on every web page. By omitting these repetitive elements, I focused solely on the unique and informative content.

2.2. Data Preprocessing

Text preprocessing is a crucial step in various natural language processing tasks, including analyzing the similarity between web pages on a website. To make the content obtained in the previous part more standardized and enhance the accuracy and reliability of similarity analysis across web pages it was decided to carry out the following manipulations:

- Lowercasing the text: It ensures that words with different cases are treated as equivalent, eliminating potential discrepancies in similarity calculations
- Removing punctuation: It helps to focus on the essential words by reducing noise caused by punctuation differences
- Removing stopwords: Excluding words that are commonly used but are semantically uninformative, reduces noise and allows for a more meaningful comparison of content.

2.3 Similarity between webpages

2.3.1 Tf-Idf

To enable meaningful similarity calculations, the text needs to be transformed into numerical vectors. For this work, I opted to use the Tf-Idf vectorizer. While embeddings could have been an alternative choice, there are several reasons why in the context of this project Tf-Idf is a more suitable approach to vectorizing the content. First of all, the content was scraped as a single text, therefore a word-level analysis using Tf-Idf was considered to be more applicable than a semantic analysis using embeddings. Additionally, the massive data size in the project and token limitations in many embedding models posed challenges, making Tf-Idf a better choice for the particular work.

2.3.2. Cosine similarity

The next step is to calculate the similarity between the vectorized documents. To achieve this, the cosine similarity metric is commonly employed. Cosine similarity is particularly suitable for this project for several reasons:

- Cosine similarity is robust to the varying lengths of documents, ensuring that similarity is determined based on the direction or orientation of the vectors rather than their overall size.
- It provides a normalized measure, ranging from -1 to 1. This range allows for clear interpretation and comparison of similarity scores across different documents.
- It is computational efficient.
- Furthermore, cosine similarity is widely adopted in text analysis tasks, including document similarity, information retrieval, and clustering. Its popularity and extensive usage in the field of natural language processing provide a strong foundation for leveraging it in this project.

As some websites contained a large number of web pages calculating the cosine similarity matrix in a single operation could be computationally expensive and memory-intensive. To handle

this efficiently, I computed it in batches. By processing the data in manageable subsets, I mitigated the computational demands and memory constraints associated with large-scale calculations.

While the cosine similarity matrix provided the necessary statistics for analyzing the competitors, for future analysis of the client, I needed to store the pairwise similarity results in a dataset with three columns: 'link1', 'link2', and 'cosine_similarity'. I processed the data in the upper part of the matrix in manageable batches and wrote the resulting rows directly to a CSV file. At the end I obtained a substantial number of 635,015,703 link pairs for the client, representing the pairwise cosine similarity calculations.

3. Results

Among the analyzed websites, the client company was found to have the lowest mean cosine similarity score of 0.003568, indicating a relatively lower content overlap.

To validate the results and gain further insights, I aimed to compare the obtained mean similarity scores with changes in organic traffic for the websites. The objective for the examination of the correlation between mean cosine similarity and changes in organic traffic was to determine if websites with lower mean cosine similarity scores experienced a more significant increase in organic traffic, while websites with relatively high mean cosine similarity scores exhibited a decline in organic traffic, potentially indicating penalties for low-quality content. Assessment of this relationship aimed to shed light on the potential impact of content similarity on search engine rankings and user engagement.

For this analysis, I selected two specific dates: March 15th (Google's core updates), and March 29th (two weeks after the core updates). These dates were chosen to assess whether there were notable changes in organic traffic following the updates.

While it is true that having only six observations limits the reliability and statistical robustness of the obtained results, a negative relationship between the mean cosine similarity and the change in organic traffic was still observed. It implies that the websites with higher mean cosine similarity are more likely to be penalized for the content overlap, and may experience a decline in their visibility or search engine rankings, resulting in a decrease in organic traffic. Although the small sample size restricts the ability to draw definitive conclusions, this preliminary observation aligns with the notion that search engines penalize or devalue websites with low-quality or duplicate content.

Additionally, I observed a negative relationship between the number of web pages on a website and the mean cosine similarity. This could indicate that websites with a larger number of pages exhibit more diverse and varied content, resulting in lower content overlap or resemblance between individual pages. On the other hand, websites with a smaller number of pages may have a higher likelihood of containing similar or duplicated content across their limited number of pages, leading to higher mean cosine similarity scores.

In addition to the analysis conducted, I also aimed to explore the percentage of link pairs that exhibited a high level of similarity for every website. To define "very similar" webpages, I set a threshold of 0.8 for the cosine similarity score. While the client exhibits the lowest mean cosine similarity score among the websites analyzed, when considering the percentage of webpage pairs with a cosine similarity higher than 0.8, another company stands out with an exceptionally low value of just 0.008%.

4. Modeling

In addition, I decided to further explore the task of content similarity identification by training a model. To accomplish this, I began by selecting a random subsample of 1,000,000 link pairs from the original dataset for the client, which consisted of 600,000,000 links, and saving it in ROC format. This subsampling process was performed using Apache Spark.

Further, I created a new variable named 'target,' which takes the value of 1 if the link pair has a cosine similarity score above 0.8, and 0 otherwise. Upon inspecting the class distribution, I noticed a severe imbalance in the dataset. Out of the 1,000,000 link pairs, only 32,597 pairs were classified as 'very similar', indicating a significant disparity between the majority and minority classes. Imbalanced datasets can pose challenges for machine learning models, as the majority class overwhelms the minority class, leading to biased results. To address this issue, I performed downsampling on the majority class (target = 0) to achieve an equal number of observations in each class, ensuring a balanced dataset for model training.

For this project, I decided to train two models: Logistic Regression (Logit) and Random Forest. Before fitting the Logit and Random Forest models, I once again employed the Tf-Idf vectorizer to transform the text predictors (contents from two webpages) into numerical representations.

Consequently, both models showed a good performance. Nevertheless, Random Forest achieved the best result with an accuracy score of 0.98. It indicates that the model correctly classified 98% of the webpage pairs in the test set, demonstrating strong predictive capability.

5. Recommendations

Based on the potential relationship observed between mean cosine similarity of the webpages and the change in the organic traffic, it is recommended that the company continues to monitor and analyze the similarity its and its competitors' metrics. This ongoing observation can provide valuable insights into the performance and content quality of the websites.

To enhance the analysis, it would be beneficial to increase the number of websites included in the study. By expanding the dataset to incorporate more competitors' websites, a broader understanding of the relationship between similarity, traffic changes, and content quality can be achieved. This perspective will contribute to more robust and reliable conclusions.

Furthermore, based on the results obtained for the most similar webpages, it is suggested to identify the "weakest links" within the website's content. By pinpointing these specific pages or sections, the company can focus on improving its content quality, thereby reducing the risk of penalties from search engines.

6. Limitations

It is important to acknowledge certain limitations that could have impacted this project's scope and conclusions. Firstly, a limitation arises from the choice of analyzing the entire content from the webpages as a single text without segmenting it into meaningful sections. By dividing the content into segments, a more granular analysis could have been performed, enabling the utilization of advanced techniques like word embeddings to capture semantic relationships between words and phrases more effectively. This approach could have provided deeper insights into the similarity patterns within specific sections of the webpages.

Additionally, the project's reliability and comprehensiveness could have been further enhanced by scraping and including more competitors' websites in the analysis. Expanding the dataset to include a larger pool of competitors would have allowed for a more robust assessment of the relationship between mean cosine similarity, changes in web traffic after Google Updates, and the total number of links on the websites.

7. Conclusion

In conclusion, this project aimed to assess the similarity between the webpages of the client and its five competitors using cosine similarity as a measure. The findings shed light on several significant observations. Firstly, among the six analyzed websites, the client's website displayed the lowest mean cosine similarity score (0.003568), indicating a distinctiveness in its content compared to the competitors.

Furthermore, the project uncovered potential relationships between mean cosine similarity and two factors. Firstly, there appeared to be a negative correlation between mean cosine similarity and the change in website organic traffic. This implies that the websites with higher mean cosine similarity between webpages are more likely to be penalized for the content overlap, and may experience a decline in their visibility or search engine rankings. Consequently, companies with more similar content may face challenges in maintaining their organic traffic levels, indicating the importance of unique and valuable content for search engine optimization.

Secondly, a negative relationship was observed between mean cosine similarity and the total number of links. This implies that websites with a larger number of webpages may exhibit lower similarity levels in their content.

Nevertheless, it is important to note that these findings are based on a limited sample, as only six websites were analyzed. Thus, for more reliable conclusions, a broader range of websites should be considered.

Additionally, the project extended its focus by training a model to classify webpage pairs as "very similar" or not, using a threshold of 0.8 for cosine similarity. The Random Forest model demonstrated exceptional performance, achieving an accuracy score of 0.98 on the test set. This signifies the model's ability to effectively distinguish between highly similar and dissimilar webpage pairs.