Advance Database Project Proposal

Shashank Gupta

October 2022

1 Problem Statement

Google Search Engine Algorithms are ranking more pages than ever while making the ranking system not so transparent. Several factors account for page ranking, but keywords and the title of the pages remains the top contributing factors. The strategic addition of relevant keywords can help improve the ranking of web pages by a far margin. However, finding such keywords is a daunting task. Therefore, in this project, we aim to create an end-to-end utility by leveraging cloud services to find the relevant keywords for any general topic. Such keywords will improve the web page ranking and increase the overall business value.

2 Methodology

The data source for this project would be the text content available from the top Google Search result links for the relevant searched text given by the user. A lambda function will be written in Python to web-scrape the text data on the top links, apply some basic text-processing steps to clean the text data, and finally apply the TF-IDF algorithm to find the relevant keywords in the document. These keywords will be converted to JavaScript JSON files and written to the DynamoDB tables. Before inserting the data via JSON file, we must design the data model carefully. The high-level architecture for this project is attached below:

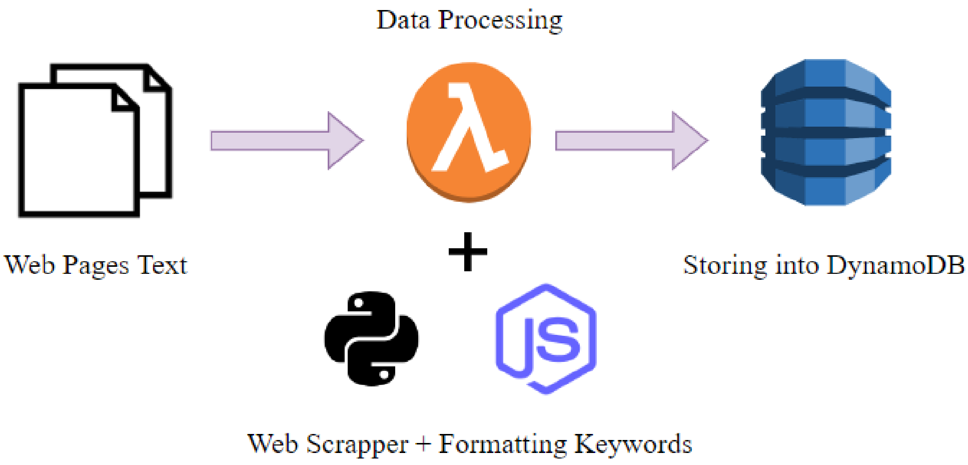

Figure 1: Architecture

3 Significance & Rationale

- The generated keyword dataset will help explain the possible optimizations and improvements in the page ranking algorithm.
- This will help in understanding the user search pattern and building high-quality links for businesses.
- Help finding relevant keywords with good traffic potential.

4 Requirements

- AWS Lambda Web service
- AWS Dynamo DB
- Jupyter Notebook (Python 3.7)
- JavaScript IDE