Running Shoe Search Engine and Ranking System

Joshua Chang UMID #17805264

Introduction

Physical activities have been proven to highly correlate with both physical and psychological well-being. Among various physical activities, running is one of the most popular forms of sports internationally due to its low cost and easy accessibility. From an economic standpoint, all a runner need is a pair of running shoes to be able to enjoy the sport. However, for novice runners who want to get into the sport, sometimes it might be daunting for them to find the right running shoes. Also for experienced runners, they might find it time-consuming and difficult to find the right racing shoe to add to their collection. “Is the shoe good for me if have wide feet?”, “Does the shoe have enough cushion if I am logging a lot of miles throughout the week?”, these are all some general questions that whether you are new to running or have been training for years might have. Hence, I hope to build a running shoe search engine that could benefit all runners to find the right model for their specific needs and also how actually the shoe performs given all the comments and reviews online. In addition to that, I hope to create a shoe ranking system that recommends the top running shoes, which are still attainable on the market, for different distances(5k, half marathon, or marathon).

Data Source

Most of the data I would expect to get is by searching sole collector database and web-scraping through Twitter, StockX, and the websites of major running shoe brands, such as Nike, Hoka, Adidas, etc to get the comments and reviews on each shoe. While collecting the data, I need specific attributes, such as retail price, on-market price range, release date, brand, size range, shoe comments, reviews, etc. Also, I would need to annotate the ground truth for each query result to compare them to my IR model.

Motivation

Although there are many shoe review websites. However, there are no general running shoe websites that give well-rounded reviews from multiple sources. Moreover, although the shoe reviews can be viewed from their own brand website. Sometimes, the comments are not reliable since the brand/company will filter out bad/unwanted comments. Hence, I hope to obtain data from Twitter and any other reliable sources to give more insights into each running shoe. By creating this search engine, runners do not have to worry about who to ask for the best advice for running shoes, since it is just one query away that they can get the right shoe they want, and get all the information

they need. Moreover, by creating the running shoe ranking system, not only runners can get to know the right information about what are the most popular shoes. Shoe companies can also realize how competitive their shoe is on the market, which I hope to become a good metric for shoe brands to reference to.

Feasibility and Timeline

This single-person project is doable if all the data scraping, data collecting, and indexing are planned ahead of time. With 10 weeks left to complete the project, I plan to set the first four weeks for web scraping, data collection, and data annotation, since I believe how to get the right data and how to annotate the data would be the most time-consuming tasks. Next, the next four weeks would be building the search engine and defining and implementing the evaluation metrics to evaluate our search engine. For the last two weeks, I would take this period as a buffer time to solve any problems that I run into along the way, and hopefully complete the project ahead of the deadline.

Potential Challenges

While this project might sound super exciting and ideal, I believe the first challenge for me to overcome would be learning how web scraping works and getting the right data I require since I have no specific knowledge or experience in this, so if possible, I really hope the teaching staff could help me out to avoid any unnecessary time on things that would actually not work out for my model. In addition, with a huge amount of data expected in this project, I believe annotating the data on my own would take so much time for me. Hence, hopefully, if someone is doing a similar project to mine, I would really love to team up to share the burden.

Evaluation Performance

For this shoe search engine project, I would use NDCG and BM25 or other evaluation metrics that I find better to evaluate the ranking function in my retrieval model. For the final shoe ranking system, since running shoe reviews and rankings online are mostly subjunctive and not collected from a huge amount of data sources, I might need to still learn the correct way to measure and compare my running shoe ranking results. Besides, to avoid biased and missing features and results, I would take out some testing data to evaluate the biasness of the model and hope to create some sort of debiasing algorithm to preprocess the data.

Equity

In terms of the equity of the project design, the collection of data and the objective function I hope to use will be in line with ethics.

Team Members

For now, this is a project only a one-man band by myself, I believe there will be less time discussing the methods to be used in my project with other groupmates but more time asking for advice and suggestions from the teaching staff.