SI650 Project Proposal

Name	E-mail	UMID
Duanmu Mingliang	duanmuml@umich.edu	95520934
Gu Yucheng	ycgu@umich.edu	97263073

Project Objective

Task and Deliverables:

Our project aims to build a second-hand car search engine for people **who do not know cars well**. Our system will take in a query containing general descriptions of the expectations of the user, and return a list of second-hand cars available on the market that best suits the requirements of the user, in the order of relevance.

Data:

To build the second-hand car dataset, we plan to use web scraping on popular car dealer websites like cars.com and Cargurus. To make a comprehensive evaluation on each car, our data should include the major fields of **price, mileage, make, drivetrain, external color, engine and gas mileage**. As manual labelling is necessary in deciding the relevance of results, we decide to limit the body style to **sedan and SUV** to ensure a moderate size of the data we will be working with.

Importance of Topic

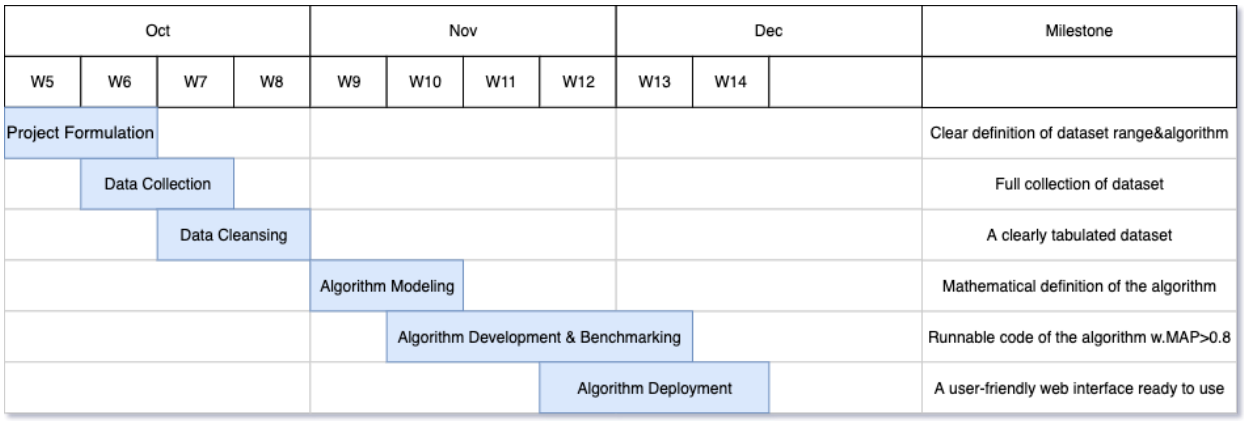
The current car dealer websites are mostly filter-based, which means people must decide on a series of parameters like mileage, drivetrain, make and engine before being presented with choices. However, the fact is most people have no concept of these values, especially for those buying a car for the first time. A descriptive sentence that implicitly conveys the needs of a car buyer is more common in the real world, for instance, “I want a small car for me and my roommate that has a good speaker and supports Apple CarPlay”, “I want a big car that can take my family of four and a dog for long trips”. Our system is acting like a sales manager focusing on such fuzzy search to better help nonprofessional car buyers choose the best deal according to their needs.

Feasibility of the Project

Due to the rich programming experience with Python and knowledge about cars for both of the team members, we have confidence in obtaining data from the mainstream car dealer websites. Also, we will search for open datasets of used cars to perform an EDA to better grasp the features of different cars. In terms of ground truth decision, we will perform a series of generated queries (enough to cover a wide range of customer needs) to our model and manually label the relevance of each result (about 50 for each query). As we are limiting the dataset only for sedans and SUVs, the annotation amount should not be substantial. After a few rounds of queries, we can cover the majority of candidates.

Tentative Timeline

The following graph shows our estimated timeline of development.



Oct				Nov				Dec			Milestone
W5	W6	W7	W8	W9	W10	W11	W12	W13	W14		
Project Formulation	▨										Clear definition of dataset range&algorithm
	Data Collection	▨									Full collection of dataset
		Data Cleansing	▨								A clearly tabulated dataset
				Algorithm Modeling	▨						Mathematical definition of the algorithm
					Algorithm Development & Benchmarking	▨	▨	▨			Runnable code of the algorithm w.MAP>0.8
							Algorithm Deployment	▨	▨		A user-friendly web interface ready to use

Risks of the Project

The risks of this project are classified into three categories, as the risks of data collection, the risks of data processing and the risks of algorithm effectiveness.

Risks of Data Collection

As described in the previous section, we plan to scrape raw data from a set of car dealer websites. We will develop a series of automated scrapers to accomplish this task, and all the subsequent works will depend on the quality and quantity of the data collected.
However, there is a possibility that those websites are not scraper-friendly. Potential problems include the following, as their corresponding solutions listed:

- **The website has anti-scraper strategies.** To overcome this, we will need to write smarter scrapers, implementing techniques such as self-adopted user agents, periodic sparse scraping and distributed scraping.
- **The website uses JavaScript to load dynamic contents.** The solution to this problem is to use browser puppet frameworks instead of static HTML frameworks. For instance, use `Selenium` to replace `BeautifulSoup`, and develop corresponding scripts. Since these frameworks usually have poorer efficiency, parallel computing frameworks may be introduced to boost data collection.

Technical Skills required:

- Scrapers such as `Selenium` and `BeautifulSoup`.
- Understanding of basic `HTML` and `JavaScript`.
- (Optional) Distributed frameworks such as `MRJob`.

Risks of Data Cleansing

After raw data collection, we will begin data cleansing work, including tabulating the data, removing corrupted rows, filling up N/A values and preprocessing the data. The potential risks in this step include the following, as their corresponding solutions listed:

- **The data from different sources have different keys.** This will result in loss of keys if we join the data together. Possible solutions including definition of a set of default values.
- **The scrapable data may have plenty of missing values.** In this case, we plan to fill those values via an additional source, such as cars-data.com.

Technical Skills required:

- Data processing tools such as `Pandas` and `SQL`.

Risks of Algorithm Effectiveness

The algorithm may not achieve our expected performance. We list some possible scenarios, with as their corresponding solutions:

- **The algorithm has poor performance due to the quality of data.** In this case, we should examine the dataset again, expand it or do another cleansing according to the problem.
- **The algorithm has poor performance due to parameters that are not fully trained.** In this case, we can introduce different methods for parameter tuning, such as `Ray`, a distributed framework for algorithm weights training.

Technical Skills required:

- Basic IR frameworks including `Pyserini`.
- (Optional) IR system benchmarking frameworks such as `beir`.
- (Optional) Algorithm tuning frameworks such as `Ray`.

Currently, we believe that no access to special kinds of resources is required.

Evaluation Metrics

To evaluate the effectiveness of the system, we plan to perform a series of searches. The following scheme will be applied:

1. A set of key points will be predefined, we will use a script to do a bootstrap sample from the key points, and combine the sampled keys to generate a query.
2. The process will be repeated, so as a result a batch of different queries will be generated. NLP methods may be introduced to smooth the queries and make them more like natural language. The queries will first be verified manually to ensure that they are reasonable (e.g. There won't be queries such as "A black sedan that can hold 6 people").
3. Once the queries are obtained, they will be inputted into the system to get the corresponding results.
4. The results will be manually labeled, and MAPs will be calculated based on the labels as a metric. Other metrics such as NDCG may be considered if time allows, but since it requires a ranking label, it might not be implemented.

To claim success, we would set a milestone of MAP>0.8. The milestone may be adjusted to fit the dataset and other changeable situations, but shouldn't be varied too much from this baseline.

Team Members

Our team includes two members.

- Duanmu Mingliang (端木明亮), first year MSI student of Data Track.
- Gu Yucheng (顾煜程), first year MSI student of Data Track.

In this project, we are working on a new field and there are - as we investigated - no available datasets that are ready-to-use. Therefore the collection of data will add some workload to the project. Additionally, due to the many features that a car would have, it is important to develop a feature extraction algorithm to preprocess the user query, and to rank the information according to different features. Therefore, this project requires two people to cooperate.
In principle, **Mingliang will focus on the development of algorithms & benchmarking**, while **Yucheng will prepare the dataset as well as the deployment of results.** The work will still have some overlaps, and whoever is free at the moment should offer help to better accomplish each other's work.