Group Member: Hanzhi Zhang 4395561906, Bin Zhang 5660329599

Technical challenge:

1. Dataset size: our dataset is large (till now it exceeds 4G) and we will use Spark to solve this problem.

2. Information extraction: We can extract the name entity such as food, room, or drink from user reviews. However, we want to also extract adjective of name entity such as delicious, juicy or bad. We will train our own NLP model based on the homework to analyze the text and get useful information for future recommendations.

3. Entity resolution: The food name is various. The computer cannot know that Macaron is a dessert. Therefore, we might use the existing KG such as Wikidata to build an Ontology. Furthermore, we might also train a LSTM model to make a food type prediction.

3. Frontend display: both of our team members are not familiar with frontend technology. So, we will choose FastAPI (a developer-friendly framework) to do rapid development.

Project Task Distribution:

Currently, we have three datasets need, TripAdvisor, Yelp and OSM, to be processed. Bin would do the Information Extraction on TripAdvisor. Hanzhi would do the same task on the Yelp. OSM is structured data. Bin would try to solve entity resolution and blocking. Hanzhi would build RDF model to knowledge presentation. Then Bin would focus on the graph visualization, and search system. Hanzhi would focus on the QA system of KG.

Progress summary

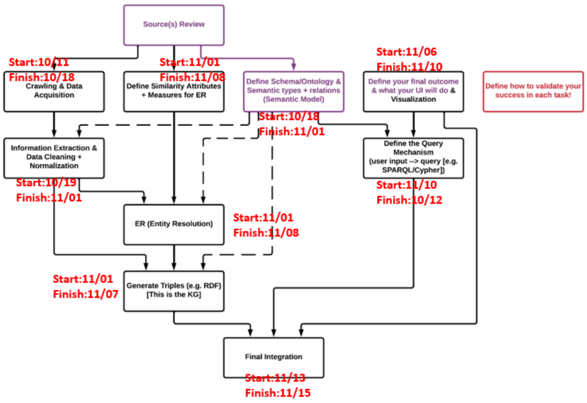
Until now, we have successful collected the raw data and currently are doing data cleaning and operating entity resolution and NLP model building. Our data is stored on the google drive: https://drive.google.com/drive/folders/1L1md1yI1t7WlijNRG-O9a75Sa5IbsapQ?usp=sharing. Our scripts are updated on GitHub: https://github.com/Bin-Go2/LA-Travels. Now, we are solving the problem related to information extraction which is already discussed Technical challenge.

Evaluation of Current Task

We extracted 1% percent data as our test data, to test the data's completeness, cleanness and following entity resolution's correctness.

Schedule

Index	Timeline	Description	Status
A	18-Oct	Crawling Data from website, Yelp, TripAdvisor and OSM	Finished
B	25-Oct	Finish the information Extraction of user reviews and Part of Ontology from Wikidata	Pending
C	1-Nov	Finish the total Schema/Ontology and the rest of information extraction	No started
D	8-Nov	Finish the entity resolution for difference datasets & generate knowledge graph	No started
E	15-Nov	Q/A system design, UI design & Validation of our system	No started
F	16-Nov	Presentation	No started
G	22-Nov	Report	No started


Source(s) Review
Start:10/11
Finish:10/18
Crawling & Data Acquisition
Start:11/01
Finish:11/08
Define Similarity Attributes + Measures for ER
Define Schema/Ontology & Semantic types + relations (Semantic Model)
Start:11/06
Finish:11/10
Define your final outcome & what your UI will do & Visualization
Define how to validate your success in each task!
Start:10/18
Finish:11/01
Information Extraction & Data Cleaning + Normalization
Define the Query Mechanism (user input --> query [e.g. SPARQL/Cypher])
Start:10/19
Finish:11/01
Start:11/10
Finish:10/12
ER (Entity Resolution)
Start:11/01
Finish:11/08
Start:11/01
Finish:11/07
Generate Triples (e.g. RDF) [This is the KG]
Final Integration
Start:11/13
Finish:11/15