CS5481: Data Engineering - Projects

Instructions

1. Due at 11:59 pm, November 26, Tuesday, 2024.
2. This is the group project. Each group has up to 8 members. Please set up your group by 11:59pm, October 5, 2024 on Canvas-People-Groups.
3. You are required to submit the project report and source code via Canvas and give a 15-min presentation for your project in class. The project report should at least consist of following parts, including introduction, methodology, experiments and discussions. The report may contain up to 6 pages of main content, plus unlimited pages of references and appendix. The source code can be submitted by Jupyter Notebook or Python files.
4. Please attach your presentation slides at the end of the report.
5. Please state the individual contributions and ratios (%) in the report.
6. This project is very open, and you are highly encouraged to come up with fantastic ideas and designs!
7. If you have any questions, please post your questions on the Canvas-Discussion forum or contact our TAs.

Topic 1 - Recommender System

LLM-based recommendation systems have emerged as a significant topic in recent years. Unlike traditional recommender systems, such as content-based or collaborative filtering approaches that primarily model users' collaborative preferences, LLM-based systems leverage large language models (LLMs) to enhance recommendations [1]. LLMs can be utilized either to make direct recommendations or to assist in the recommendation process. Recently, several methods have been proposed to address this task, including TALLRec [2], LLaRA [3], etc.

The MovieLens dataset[1] is a widely used recommendation dataset, obtained from user interactions with movies. It provides a rich source for evaluating different models. We encourage you to experiment with various models on MovieLens and other datasets to assess their performance.

Additionally, you are also encouraged to explore other methods, such as prompt engineering and instruction tuning, to further enhance LLM-based recommendations. For instance, employing chain-of-thought prompting can help elicit the reasoning capabilities of LLMs, while instruction tuning can align the models more closely with recommendation tasks. For more details on LLM-based recommendations, please refer to [4,5].

Reference

1. Lin, J., Dai, X., Xi, Y., Liu, W., Chen, B., Zhang, H., ... & Zhang, W. (2023). How can recommender systems benefit from large language models: A survey. arXiv preprint arXiv:2306.05817.

2. Bao, K., Zhang, J., Zhang, Y., Wang, W., Feng, F., & He, X. (2023, September). Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In Proceedings of the 17th ACM Conference on Recommender Systems (pp. 1007-1014).

3. Liao, J., Li, S., Yang, Z., Wu, J., Yuan, Y., Wang, X., & He, X. (2024, July). Llara: Large language-recommendation assistant. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1785-1795).

4. Wu, L., Zheng, Z., Qiu, Z., Wang, H., Gu, H., Shen, T., ... & Chen, E. (2024). A survey on large language models for recommendation. World Wide Web, 27(5), 60.

5. Li, L., Zhang, Y., Liu, D., & Chen, L. (2023). Large language models for generative recommendation: A survey and visionary discussions. arXiv preprint arXiv:2309.01157.

Topic 2 - Event Timeline Generation

With the rapid development of social media platforms, there are massive news/posts appeared when an breaking event occurs. However, those social media data might always be fragmented. With the evolution of the event, more social media data related to this event would be posted. Therefore, a critical problem is how to gather those social media data together and produce an event timeline to help people better learn about the event.

To this end, we are encouraged to crawl the social media data and generate a timeline for a specific event. You can finish this task in two fashions: 1) firstly, you crawl much social media data from the Internet, and then detect the events from your data and finally produce a timeline for the event. 2) given event keywords, you firstly crawl the social media data related to the event from the Internet, and then produce the event timeline.

Note that:

1. For each node of the timeline, following information should be included: time, event description, news source.

2. You can directly temporally organize your crawled social media data related to the target event, but remind to filter the overlapped data from different sources.

Reference

1. Li C, Sun A, Datta A. Twevent: segment-based event detection from tweets[C]//Proceedings of the 21st ACM international conference on Information and knowledge management. 2012: 155-164.

2. Hasan M, Orgun M A, Schwitter R. A survey on real-time event detection from the Twitter data stream[J]. Journal of Information Science, 2018, 44(4): 443-463.

3. Guo B, Ouyang Y, Zhang C, et al. Crowdstory: Fine-grained event storyline generation by fusion of multi-modal crowdsourced data[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2017, 1(3): 1-19.

[1] https://grouplens.org/datasets/movielens/

Topic 3 - Search Engine

Try to build your own search engine. To obtain a strong search engine, you can consider from three aspects:

1. Massive data. Firstly, you should have sufficient data to support your query. So, please crawl as much as possible data from the Internet. The data can be anything, including news, blogs, posts, etc. At least 1 million pieces of data are needed.
2. Data management. Think about how to store and manage the data on your device.
3. Data query. Given a query, how to retrieve the most related data from the database as fast and accurate as possible.

Based on the above requirements, please try to implement your own search engine. Note that you are not encouraged to directly query the data by SQL statements. Try to apply information retrieval techniques.

Reference

1. https://towardsdatascience.com/how-to-build-a-search-engine-9f8ffa405eac
2. https://www.opengrowth.com/article/how-to-build-a-search-engine

Topic 4 - Data Statistics and Visualization

Generally, some keywords can reflect the tendency of a specific domain since they are frequently mentioned in the related news of the domain. For example, "stock", "dollar" and "business" might be the keywords of financial news. Therefore, please try to crawl the social media data in a specific domain, then find the keywords of the domain. The keywords can be the most frequently used meaningful words or phrases in the domain, but not the general or stop words. Furthermore, try to conduct some data statistics and visualize the results.

Examples of Keywords Visualization

1. https://www.wordclouds.com/
2. https://www.mentimeter.com/features/word-cloud

Topic 5 - Personalized Chatbot using Large Language Models

In recent years, Large Language Models (LLMs) such as GPT-4, LLaMA, and others have shown remarkable capabilities in understanding and generating human-like text. A key application of these models is in creating personalized chatbots that can engage with users in a more contextual and meaningful way.
For this project, you are tasked with developing a personalized chatbot using an LLM. The chatbot should be able to:

1. **Understand User Profiles:** The chatbot should take into account the user's past interactions, preferences, and stored data to provide personalized responses.
2. **Real-Time Adaptation:** Implement mechanisms that allow the chatbot to adapt in real-time based on ongoing conversations, refining its responses to align with the user's style and needs.
3. **Contextual Understanding:** Ensure the chatbot can maintain the context across sessions, remembering past conversations and using that information to inform future interactions.
4. **Data Privacy:** Address challenges related to storing and managing user data securely. Implement data privacy practices, such as anonymization and encryption, ensuring compliance with data protection regulations like GDPR.
5. **Evaluation Metrics:** Develop metrics to evaluate the performance of the chatbot, including user satisfaction, response accuracy, and adaptability.

Requirements:

1. **Data Collection:** You can collect user interaction data by simulating conversations or using existing datasets.
2. **Utilizing Open-Source LLMs:** Use open-source LLMs like LLaMA or Qwen to build the chatbot, you can directly utilize the API.

3. **Visualization:** Visualize the chatbot's decision-making process and how it adapts to different user profiles.

Reference

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. NeurIPS.
2. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
3. Qwen-7B: An Open-Source Large Language Model by Alibaba Cloud. (2023). Alibaba Cloud Documentation.

Topic 6 - Open Your Mind

This is a very open-minded project that you can try anything you are interested and related to the course content. For this topic, the only requirement is that you should crawl massive data from the Internet, and then do something with the data. For example, you can predict the stock tendency based on the social news; you can build the social networks based on the interactions among users on social media platform. **We highly recommend using ChatGPT to do some fun things**. Just open your mind and take a try. **Please contact TAs to verify your own selected topic before starting the group work**.