# CS5481: Data Engineering - Projects

## Instructions

1. Due at 12:59pm, November 15, Tuesday, 2022.
2. This is the group project. Each group has 2-4 members. Please set up your group by 12:59pm, October 15, 2022 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 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 TA Mr. Han Wu (email: hanwu32-c@my.cityu.edu.hk).

# Topic 1 - Session-based Recommender System

The session-based recommendation has been an emerging topic in recent years. Different from other traditional recommender systems, such as content-based recommender systems or collaborative filtering recommender systems, which usually model users' long-term and static preferences, session-based recommender systems aim to capture users' short-term and dynamic preferences to provide users with real-time and accurate recommendation service. In recent years, a number of methods have been proposed to address this task, such as GRU4Rec[1], NARM[2], and SR-GNN[3].

Yoochoose[4] is a widely used session-based recommendation datasets. The Yoochoose dataset is obtained from the RecSys Challenge 2015, which contains a stream of user clicks on an e-commerce website within 6 months. Please try different models on Yoochoose and evaluate over the following metrics, including Precision and Mean Reciprocal Rank (MRR).

Additionally, you are also encouraged to explore the application of other methods, such as self-supervised learning or meta learning, on session-based recommendation. For example, you could use self-supervised learning to conduct data augmentation, aiming at enhancing the model performance on the session-based recommendation. Please refer to [5] for more details about session-based recommendations.

Note that:

1. Owing to the large scale of the dataset and the limited computing resources, you sample a subset from the official Yoochoose dataset. You can download the dataset from Canvas-Files-Project.

2. During testing, you can filter the sessions that only have a record, and predict the second or later item_ids based on the history.

**Reference**


1. Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2015). Session-based recommendations with recurrent neural networks. arXiv preprint arXiv:1511.06939.

2. Li, J., Ren, P., Chen, Z., Ren, Z., Lian, T., & Ma, J. (2017, November). Neural attentive session-based recommendation. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 1419-1428).

3. Wu, S., Tang, Y., Zhu, Y., Wang, L., Xie, X., & Tan, T. (2019, July). Session-based recommendation with graph neural networks. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 346-353).

4. https://www.kaggle.com/datasets/chadgostopp/recsys-challenge-2015

5. Wang, S., Cao, L., Wang, Y., Sheng, Q. Z., Orgun, M. A., & Lian, D. (2021). A survey on session-based recommender systems. ACM Computing Surveys (CSUR), 54(7), 1-38.


# 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.

## 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 - 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. Just open your mind and take a try. **Please contact TAs to verify your own selected topic before starting the group work**.