Amirkabir University of Technology
(Tehran Polytechnic)

**Final Project**
**We need privacy!**

## Guidelines and Policies

- It is expected that the students submit a project report (Project_[student_id].pdf) as well as required source codes (.m or .py) into an archive file (Project_[student_id].zip). Please combine all your Persian reports just into a single .pdf file by problems order. Code without report has an exact zero point.
  Please do not use implementation tools when it is asked to solve the problem by hand, otherwise you will be penalized and lose some points.
- You are free to solve by-hand problems on a paper and include their pictures in your report. Here, cleanness and readability are of high importance. Images should also have appropriate quality.
- Your work will be evaluated mostly by the quality of your report. Do not forget to explain your answers clearly, and provide enough discussions when needed.
- In this project, 4 points (out of a possible 100) belong to compactness, expressiveness, and neatness of your report and codes.
- By default, we assume you implement your codes in Python. If you are using MATLAB or R, you have to use the equivalent functions when it is asked to use specific Python functions.
- Your codes must be separated for each question.
- Make sure you have access to Courses, because that is where all assignments and project definition as well as course announcements are posted. Submissions are only made through Courses.

- Please submit your work **before the end of July 13rd**.
- Submissions must be made by the regular deadline; **No delay** are allowed due to the grade publication schedule.
- You are not allowed to share codes/answers or use works from the past semesters. Violators will receive a zero for that particular problem.
- If there is any question, please do not hesitate to contact us through the Telegram group chat.



## LSH-based Collaborative Recommendation Method with Privacy-Preservation

Welcome to an exciting journey into the world of Recommendation Systems!
In this project, you'll delve into the realm of recommendation algorithms, specifically focusing on a novel approach that incorporates Locality Sensitive Hashing (LSH) for recommendation while also ensuring privacy preservation. We'll be using the MovieLens 20M Dataset as our playground, a rich source of movie ratings and tag applications from a vast community of users. By the end of this project, you'll not only gain insights into the intricacies of recommendation systems but also understand the importance of safeguarding user privacy in the era of Big Data.

1. **Dataset Exploration and Structure:** Your first task is to embark on a journey through the MovieLens 20M Dataset. You'll dive deep into its structure, understanding how the data is stored and organized. Through this exploration, you'll familiarize yourself with the intricate details of the dataset, laying the foundation for the subsequent steps. By visualizing the dataset's structure, you'll gain valuable insights into its complexity, setting the stage for efficient preprocessing and analysis.

2. **Preprocessing and Sparse Representation:** Once you've grasped the structure of the dataset, it's time to prepare it for analysis. Preprocessing is key to ensuring that the data is in a suitable format for our recommendation system. You'll undertake tasks such as loading, transformation, and normalization, ensuring that the data is optimized for further processing. Given the vast scale of the dataset, you'll also explore sparse representation techniques to alleviate hardware resource limitations, enabling efficient computation on large datasets without compromising accuracy.

3. **Dataset Splitting and Train-Test Division**: Navigating through the vast expanse of the MovieLens 20M Dataset, you'll encounter a large user-item table, a treasure trove of user preferences and movie ratings. But how do we effectively split this behemoth into train and test sets? It's a meticulous process, one that requires careful consideration of the user-item interactions. You'll embark on this journey by outlining a systematic procedure for splitting the dataset, detailing the steps and rationale behind each decision. Furthermore, you'll roll up your sleeves and implement your own code, breathing life into the splitting process. Through hands-on experimentation, you'll gain a deeper understanding of the nuances involved, ensuring a robust and reliable train-test division for evaluating our LSH-based recommendation method.

In your project, you'll embark on a journey to delve deep into the intricacies of the "LSH-based Collaborative Recommendation Method with Privacy-Preservation" paper, presented at the 2020 IEEE 13rd International Conference on Cloud Computing. Here's what awaits you:

4. **Understanding Privacy-Preservation:** As you dive into the paper, you'll encounter the term "Privacy-Preservation" and its paramount importance in the realm of recommendation systems. Your task is to unravel the significance of this term and how it's addressed within the paper. Explore the methodologies proposed to safeguard user privacy while enhancing the recommendation process, shedding light on the innovative approaches outlined by the authors.
5. **Time Complexity Analysis:** Every algorithm comes with its computational demands, and the proposed method in the paper is no exception. Your challenge is to justify or explain the time complexity of the method. By dissecting the algorithmic intricacies, you'll uncover the underlying computational requirements, providing insights into the efficiency of the proposed approach.

6. **Implementation in Functional/OOP Architecture:** Now comes the hands-on part – implementing the proposed method. You'll embark on a coding journey, structuring your implementation in a Functional/OOP architecture to ensure reusability and maintainability. Along the way, you may encounter challenges, such as optimizing performance or handling large-scale datasets. Fear not! Your task is not only to implement the method but also to devise innovative solutions to overcome any hurdles you encounter. Note anything in your report!

7. **Hyper-parameter Optimization and Conventional LSH Settings:** Hyper-parameters play a pivotal role in fine-tuning the performance of recommendation algorithms. You'll explore the challenges of setting hyper-parameters throughout the method implementation. Additionally, you'll investigate the conventional settings of LSH and their impact on recommendation performance. By conducting experiments, generating plots, and comparing results with those presented in the paper, you'll gain a comprehensive understanding of the method's efficacy and its potential for improvement.

With each step of the project, you'll unravel new insights, tackle challenges head-on, and emerge with a deeper understanding of collaborative recommendation systems and privacy preservation techniques.


Amirkabir University of Technology
(Tehran Polytechnic)

**More challenges, more scores!**

If you're aiming to maximize your project score, let's delve into the remaining sections for an opportunity to shine. Achieving an awesome grade requires engaging in extraordinary activities, and here's how you can score up to 2 out of 20 on your final grade!

8. **Instagram page selection:** Begin by selecting a subject of interest from the options provided below. Then, curate a list of 30 online shops, influencers, or bloggers on Instagram relevant to your chosen subject. Report the IDs and links of these selected webpages in a structured table format.
    a. **Sport:** Club pages, players, coaches, leaders
    b. **Food:** Chefs, recipes, restaurants, food bloggers, testers
    c. **Health and Fitness:** Personal trainers, nutritionists, fitness enthusiasts, workout routines, wellness coaches
    d. **Travel:** Travel bloggers, destination pages, travel agencies, adventure seekers, travel photographers
    e. **Fashion:** Fashion designers, clothing brands, fashion influencers, stylists, fashion photographers
    f. **Technology:** Tech bloggers, gadget reviewers, software developers, tech companies, innovation pages
    g. **Education:** Educational institutions, online courses, educators, education bloggers, study resources
    h. **Entertainment:** Actors, directors, production companies, entertainment news pages, film critics
    i. **Beauty:** Makeup artists, skincare experts, beauty influencers, cosmetic brands, hair stylists
    j. **Business and Entrepreneurship:** Entrepreneurs, startup pages, business coaches, venture capitalists, business magazines

**Note:** Each category should include a diverse but similar range of relevant Instagram pages to create a comprehensive dataset for analysis.

9. **Scraping and Feature Extraction:** Utilize open-source libraries to scrape the selected Instagram pages and extract various features such as follower count, following count, number of posts, average likes and comments per post, average number of slides per post, average length of captions, and top-10 frequent tags. Organize this data in JSON format and save it in a file named

"StudentID_InstaDataset.info". If you encounter any filtering or connection issues while scraping, feel free to use any free hosting (like: https://www.pythonanywhere.com/) service for your script; A sample file is attached.

10. **User Data Scraping and Compilation:** Scrape the users who have interacted with the mentioned pages by liking or commenting on posts. Gather data including follower count, following count, number of posts, likes and comments per page, and overall engagement metrics. Organize this data in JSON format and save it in a file named "StudentID_InstaUserDataset.info"; A sample file is attached.

11. **Numerical Rating Generation:** Develop rules for assigning numerical ratings to users based on their interactions with the Instagram pages. Consider factors such as following a page, liking a post, commenting on a post, and any other relevant engagement metrics. Discuss the rationale behind your rating rules and address potential challenges, such as biased engagement behaviors.
    **Sample Rating Rules:**
    a. If a user follows a page: 0.3 score
    b. Each like on a post: 0.1 score
    c. Each comment on a post: 0.2 score
    d. Additional engagement metrics (e.g., shares, saves): Assigned scores based on their perceived importance to engagement

12. **Recommendation System Implementation and Evaluation:** Apply the implemented recommendation approach to the compiled dataset and evaluate its performance. Report the results, including any plots or visualizations. Cherry-pick 30 samples from the evaluation results and discuss their ratings in detail, providing insights into the effectiveness of the recommendation system on your custom dataset.

By diligently completing these sections and demonstrating your understanding of recommendation systems and web data analysis techniques, you'll pave the way for an outstanding project outcome and a stellar final grade.