Getting Started with Your Final Project

Introduction:

Welcome to the Final Project! This is your opportunity to step into the shoes of a Big Data Engineer\Scientist at LinkedIn, where you have the freedom to choose your path. You can join:

1. A team dedicated to developing AI-driven solutions for **internal company use**.
2. A team specializing in refining AI-driven features for **everyday end-users**.
3. A team focused on innovating AI-driven features tailored for **power users**, such as recruiters or sales personnel.

Your primary objective is to craft an **AI solution**, either for the broader LinkedIn user base or exclusively for your internal LinkedIn team, based on the specific focus area you choose from the three options provided. We encourage you to focus on creativity for your project concepts and their visual aspects. Original ideas will be given more importance than the performance of the models.

Presentation:

Upon the conclusion of the course, a poster presentation session will be arranged, providing you with the opportunity to showcase your final project to the course staff, fellow classmates, and select visitors.

Project Preparation:

Before diving in, we recommend creating a LinkedIn account and immersing yourself in the platform. Explore its social dynamics and functionalities to gain insights that will inform your project. Understanding how LinkedIn works is key to developing a successful AI product.

Project Examples:

To kickstart your creativity, here are a few project examples:

1. **Internal company use example**: Premium Detector
   Predict which users are more likely to opt for a Premium account, as this insight can assist our company in tailoring features and advertisements to encourage them to subscribe to the premium service.

2. **Everyday end-users example:** Profile Enhancer
   Develop an AI-powered feature within LinkedIn to assist users in enhancing their profiles, ultimately improving their visibility and professional appeal.

3. **Power users example**: My Great Catch
   For recruiters, envision "My Great Catch," an AI-driven system designed to streamline the candidate discovery process. This tool assists recruiters in finding ideal candidates based on their unique hiring needs or the attributes of their existing high-performing hires.

Data Collection Guidelines:

Data (LinkedIn Databases):

You have access to two essential datasets: **linkedin/profiles** and **linkedin/companies**. Explore and understand their contents to make informed decisions in your project.

Enrichment (Supplementary Data):

Besides the databases provided, gather additional data using automatic code-based tools (scraping or similar, not manual scarping) crucial for your project's success. Document your approach, considering relevance, accuracy, and ethical considerations.

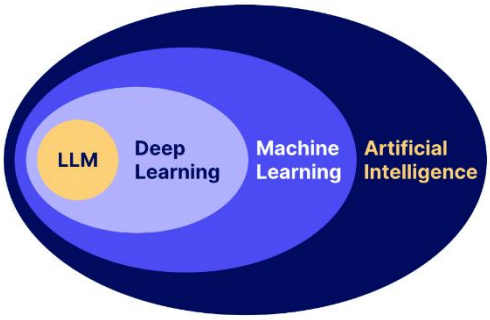
At least $1000 \cdot \alpha(group)$ items, where $\alpha(group) \geq 0.1$ is defined per group (at the project proposal submission), as a combination of task complexity, taking into account the dimensionality of the data item (rows vs columns).

Define what you consider as an item and your **estimate** for enrichment size (rows and columns), and $\alpha(group)$.

Libraries and Tools:

- Platform: Databricks clusters are used for efficient code execution with large datasets.
- Visualization: You are welcome to use the visualization features that databricks give you (queries, dashboards, etc.), and any plotting library.
- Libraries: You may use any external library you choose that can be installed on your cluster.
- Scraping: Scraping should be an integral part of your project –
    - You may use BrightData platform, native Python libraries (selenium, beautiful soup or similar), Ready to use Dataset etc.

Leveraging Large Language Models:



- LLMs usage is optional and should be used in moderation.
    - *Striking the right balance may elevate your project without overwhelming it.*
- Consider integrating advanced language models such as ChatGPT, Gemini, or similar technologies, if you think they can improve specific aspects of your AI product's intelligence and capabilities.
- Feel free to employ LLMs as you see fit within your project. However, we encourage you to think creatively about how these tools can be used not just for their own sake but to genuinely elevate the quality and innovation of your work.
- LLM should not be the whole "AI" in your solution. Try to incorporate both classic and advanced methodologies to elevate the project.

Project Milestones:

Grading

15% project proposal.
15% project presentation.
70% project report.

See general grading scheme at the bottom of this document.

Project Proposal

Deadline: 04.03.24 at 23:59

Submit a 2-pager PDF with the following sections:

Introduction and Chosen Path:

Briefly introduce the project, emphasizing the chosen path (internal solutions, end-user features, or power user-focused tools), and your research question/s.

Benefits of the New Solution:

Clearly outline the benefits that the new AI product aims to bring, whether it's improved efficiency, enhanced user experience, strategic business advantages or other.

Relevant Data from General Databases:

Specify how you intend to utilize data from the **linkedin/people** and **linkedin/companies** datasets for your project.

Additional Data Collection:

Discuss any supplementary data you plan to incorporate, why it's relevant, and your methodology for collecting it. Define what you consider as an item and your estimate for enrichment size (rows and columns), and $\alpha(group)$.

Project End Goals:

Define the ultimate goals of the project and the specific AI solution, and how you plan to implement it.

Submission Checklist:

- ✓ Proposal PDF with the name Proposal_ID1_ID2_ID3.

Poster Day

Deadline: 02.04.24

Poster day is a day where you present your project main ideas. The project should be ~80% done.

We prefer in-person presentations, but we'll also organize a Zoom session later in the day to accommodate groups with multiple students at Miluim.

The poster should be submitted to Moodle by 01.04.24 at 23:59.

What is a Research Poster?

Posters are widely used in the academic community, and most conferences include poster presentations in their program. Research posters summarize information or research concisely and attractively to help publicize it and generate discussion.

The poster is usually a mixture of a brief text mixed with tables, graphs, pictures, and other presentation formats. At a conference, the researcher stands by the poster display while other participants can come and view the presentation and interact with the author.

What Makes a Good Poster?

- Important information should be readable from about 2 meters away
- Title is short and draws interest
- Text is clear and to the point
- No more than 500 words. The less the better.
- Use of bullets, numbering, and headlines make it easy to read
- Effective use of graphics, color and fonts
- Consistent and clean layout

Where do I begin?

Answer these questions:

1. What are my project objectives?
2. What data have I used? What data have I collected?
3. What are the key findings or most interesting results from my project?
4. How can I visually share my project information? Should I use charts, graphs, photos, images?
5. What kind of information can I convey during my talk that will complement my poster?

Examples:

- Good examples
    - https://neurips.cc/media/PosterPDFs/NeurIPS%202022/54720.png?t=1668992853.1167307
    - https://nips.cc/media/PosterPDFs/NeurIPS%202023/71081.png?t=1698653842.1257799
    - https://nips.cc/media/PosterPDFs/NeurIPS%202022/52940.png?t=1669223398.122751
- Bad examples
    - https://betterposters.blogspot.com/2011/04/critique-breast-cancer-inhibition.html

Resources:

- https://365datascience.com/trending/how-to-make-a-poster-for-your-data-science-project/

- NIPS 2022 posters: https://nips.cc/virtual/2022/paper_vis.html

Poster Submission:

- 1 **page** pptx slide or similar (e.g. canva)

Project Report

Deadline: 09.04.24 at 23:59 (miluim have 2 days extension – 11.04.24 at 23:59)

For your final submission, you must submit a final report of a maximum of 5 pages detailing your work. The report should have the following sections:

Project Introduction:

Provide a concise introduction to the concept behind your developed project, emphasizing its necessity and the reasons for its creation.

Data Collection and Integration:

- Describe the data you have used from the original datasets.
- Describe the additional data you have collected; explain the methods you employed to gather and incorporate the data into your project.
- Describe how the additional data integrates in your solution.
- Define what you consider as an item and the *actual* enrichment size (rows and columns), and $\alpha(group)$.
- Include images of datasets and relevant feature names in the appendix section of your document.

Data Analysis:

- Analysis Techniques: Outline the methods used for data analysis, such as statistical models or machine learning algorithms.
- Feature Selection: Highlight the key features chosen based on statistical analysis and domain knowledge.
- Visualizations: Showcase your analysis through engaging visual representations. In the 5-page document, include only the most critical visualizations.

AI Methodologies:

Explain the methodologies, algorithms and evaluations you implemented in your project.

Evaluation and Results:

Discuss the evaluation process for both your project and algorithm.
Subsequently, present the results and key findings derived from your project.

Limitations and Reflection

This section should detail the constraints and challenges you encountered during the project. Discuss any factors that limited your approach, such as resource availability, technological constraints, time restrictions, or data limitations. Reflect on how these limitations may have influenced the project's outcomes.

Conclusions

In conclusion, summarize the achievements of your project and reiterate the key outcomes.

APPENDIX

The appendix is not counted towards the five-page document, and should be at the end of the PDF.

Images, Graphs, Plots:

You may include relevant figures from the project.

References:

This section should list all the sources you consulted or cited throughout your project. This can include blog posts, academic papers, open-source software packages, and GitHub repositories.

Project Submission Checklist:

- ✓ Final Report PDF with the name Project_ID1_ID2_ID3.
- ✓ Link to the collected dataset (google drive/OneDrive).
- ✓ Link to public **GitHub** repo which contains readable and runnable code files for all the project steps.
    - The **GitHub** repo should include instructions on how to run your project as a README.md.
    - Bonus (up to 3 points):
        - Create a pretty and appealing GitHub page with all the project files and a well written documentation.
        - A good example is – link to example git

Grading scheme:

<u>Important note:</u> the following grading scheme outlines our general approach to grading the project (report + presentation). However, we reserve the right to make exceptions/changes in some circumstances.

Verticals	Excellent	Reasonable	Could be better	Unacceptable
Problem Understanding and Motivation	Deep insight and clear alignment with Project goals; comprehensive and well-researched.	Good understanding and clear motivation; solid work with room for deeper insight.	Basic understanding; lacks depth and full alignment with the project goals.	Misunderstood or irrelevant problem area; unclear motivation.
Data Collection and Enrichment	Sophisticated data collection strategy which uses Bright Data or similar tools; highly relevant and at scale.	Adequate data collection strategy that used some tools; data is relevant but could be higher quality and more at scale.	Minimal data collection strategy; using manual collection or no code-based solutions; enriched data doesn't contribute to the solution.	Poor data quality; irrelevant to the project; data size is smaller than required.
Technical Implementation	High technical proficiency; complex solution using advanced data science tools from multiple courses or areas of research; machine learning is an integral part of the solution.	Functional solution showing competence; lacks complexity; machine learning is partly integrated in the solution.	Meets basic requirements; simple or flawed methodologies; machine learning is barely used or doesn't contribute to the solution.	Poor or non-functional; lacks data science understanding; There is no machine learning in the solution or it is used without real impact.
Creativity and Innovation	Highly original and innovative.	Some creativity and innovation.	Limited originality; minimal innovation beyond basic requirements.	Lacks originality and innovation; offers nothing new or creative.
Analysis and Visualization	Insightful analysis with clear, impactful visualizations.	Solid analysis; visualizations support findings but lack impact.	Basic analysis; visualizations lack clarity or relevance; contains grammatical mistakes.	Inadequate analysis; poor or irrelevant visualizations.
Project Report quality	Exceptionally well-written and organized; comprehensive analysis, precise language, and professional visuals.	Clearly written and structured; solid analysis with effective use of visuals.	Some organization and clarity issues; analysis and visuals need refinement.	Poorly written and disorganized; lacks clarity, analysis, and professional presentation; lacks required parts.