The Talash Search Engine:

Instructor name: Engr. Tayyaba Khursheed
Members:
- Burhan Arif 24F-CY-001
- Muhammad Saad Rizwan 24F-CY-037

Abstract: Modern search engines are quite complex and well versed in find the relevant data based on the queries of end users. This project is aimed to mimic just that but with the primary objective being enabling search of different languages to enable accessibility for people with lingual barriers. Urdu speakers face the dreaded english-urdu language barrier everyday. This project sets sail to overcome that problem for an efficient workflow in daily life of individuals who primarily speak **Urdu**. Though this project will not be deployment ready due to resource limitation, it is aimed to encourage the community to improve upon the technology to establish it properly in future.

Objective:
- To create a basic search system that is focused on search based on the **UTF-8 character-set**. Purpose of using UTF-8 characters is to enable search in Other languages Like **Urdu**.
- Run indexing on a predefined dataset, store the indices in an efficient manner so they can be accessed quickly.
- Apply searching techniques like string similarity match and binary search to allow fast and efficient search.
- Display a Google-like interface to allow intuitive UX (User Experience)

Technologies:
- Python Flask application framework
- Web Platform (HTML, CSS, JS)
- SQL database (MySQL/SQLite)

Rules and Constraints:
- At-most 10% of the project shall be allowed to be developed/coded/managed by LLMs/Chatbots to enforce learning and pondering over the concepts.

References:
- Documentations of technologies used:
  Flask: https://flask.palletsprojects.com/
  SQL: https://sqlite.org/docs.html , https://docs.python.org/3/library/sqlite3.html
  MDN for Web Docs: https://developer.mozilla.org/en-US/
- Indexing data sources:
  - https://github.com/zeerakahmed/makhzan
- Third-party libraries:
  LughaatNLP: https://lughaatnlp.blogspot.com/2024/04/mastering-urdu-text-processing.html

Methodology:
Aquiring the data:

- The corpus (Data Collection):
  - As all search engines require an enormous amount of data, we lack the resources for aquiring the data by our selves. It would require us to crawl many sites on the internet. Hence we plan to take a small already scraped dataset from a predefined index of websites. The github repository from which data was taken is given: https://github.com/zeerakahmed/makhzan
  - More data may be taken or scraped from a number of sources.

- Data Processing and Indexing:
  English is easy: Words are separated by spaces.

  Urdu is hard: Words can be connected, and the space rules are more complex. You cannot simply split on space.

  - We will use **LughaatNLP** to tokenize the urdu words correctly to associate words and interconnect them.

- Building The Index:
  - Search engine indexing is the collecting, parsing, and storing of data to facilitate fast and accurate information retrieval.

    `Eg: index["کتاب"] = [doc_id_1, doc_id_45, doc_id_102]` (a list of documents containing "کتاب")
  - This will allow us to efficiently run searches on the data to allow fast data lookup.

- Query Processing:
  - In this step, we will tokenize the user query and look for relavant indices in our system.
  - We will then rank the token results and evaluate what result will be displayed on top of the search results.
    TF (Term Frequency): How often does the query term appear in the document? More is better.
    IDF (Inverse Document Frequency): How rare is the query term across all documents? Common terms (like "میں" - "I") are less important than rare ones.

- User Interface:
  - A simple face for this proof of concept of search engine will be made in the Flask framework.

Expected Final Outcomes:
The final product will be an urdu language search engine that allows fast look-up for sites in a **Pre-defined index list.** The UI will be intuitive with pagination to allow easy browsing.

--Thank You--