Project

The 2021 Data Engineering course features a project.

Goal of the Project: Design and implement a data pipeline using Apache Airflow

2021 Project Theme: **Internet Memes**

Project Structure and Assignments

Project will focus on the Processing Phase, i.e., ingestion, cleansing, and then query-based analyses

During the course, you'll be assigned simple tasks in Airflow, which will be the base for your project!

You can complete the project points step by step and ASK FOR FEEDBACKS early!

Project Structure Details

Every project must feature:

1. Submit your task plan via model using the assignment link.
2. An initial pipeline where data are loaded from file (provided) and cleaned

3. A second pipeline where data are loaded into an ingestion system and then cleaned and processed
4. A third pipeline where a relational view is built on the data to perform some analysis
5. A fourth pipeline where data are enriched (use your creativity)
6. A fifth pipeline where a graph view is built on the data to facilitate some analysis

Natural language analyses will be provided to be implemented at point 3 and 5, a base example using the images (which are not stored) will be included in 4.

Project Requirements and Grades

- 1-3: Up to 71 Points
- 4: Up to 61 Points
- Completing the base examples at 3,4,5 up to 91
- Re-implementing the pipeline as a streaming pipeline: up to 100
- Presenting original work: new analysis, innovative solution up to 100

Project Submission Instruction

The Project submission consists of:

- a GitHub repository with the Airflow Code
- a report (written preferably in latex or markdown) that describes the design decision behind every pipeline step
- a presentation in google slide to share the day before the project presentation
- Every group shall submit the repository link via Moodle ASAP
- at submission time, a zip file containing the code + report should be uploaded to Moodle.
- You can complete the project points step by step and ASK FOR FEEDBACKS early

Examples of Tasks

Below you can find examples of valid tasks for each for the various project phases. You can use them, or you can invent new ones (appreciated).

NOTE: we are not interested in the actual result but in the design process. We recommend the group to work together and brainstorm, try different things and carry on a discussion for each tasks

- Data Cleansing
    - Removing Non-Memes entry from Know your meme.
    - removing data non forming to the schema
    - removing memes with bad-words or sensitive content
    - uniforming content structure, e.g, clustering similar tags
- Data Augmentation/Enrichment
    - Process the memes text fields (about, origin) to include bag of words
    - Download instances related to each meme template
    - Link meme templates to ImgFlip meme templates
    - Extract temporal information from text fields
    - Link the memes to Knowledge Graphs: DBPedia, Wikidata, Yago
    - Include new data by using extra APIs: DBPrdia Spotlight or google vision (we have credits if you need some)
- Data Transformations
    - Covert the memes into RDF/Labelled Property Graph
    - Create a relational Model (ER) for any of the available dataset
- Analysis
    - the queries should be written in two different query languages, e.g., SQL and Cypher (but also mongodb and SPARQL are ok).
    - What the most popular memes across country/website/community?
    - how many memes include a parent relation?
        - design issues
            - define the grouping criteria

        - design the relational schema
- given a meme m, what are the related memes m' in the dataset
- return all the memes pairs that are not related to each other directly but they are related to at least two of the same memes memes
- return all the meme that share an entity in common, but they are not related directly
    - design issues
        - define what "related" means, e.g., it exists a link of any kind
        - identity nodes identity
        - encode the properties adequately