Read the document carefully and follow as instructed. The behavioral survey (Link at the end of the document) is to be filled once the task is completed.

About the Task

Machine Learning Roles / Business Analyst

Important Points:
(1) Please use separate .py files to harvest ONLY the data from API or scrap from websites.
(2) Firstly, collect data from diverse and large datasets listed below.
(3) Use a combination of models of your choice to benchmark the accuracy, Tagging for Keyword Extraction or Named Entity Recognition (NER) using models such as hugging face, YAKE, BERT-derived models, spaCy, most popular language models. You are free to use other models
(4) Provide a report with succinct visualization of results and all your different .py scripts (class object oriented good scripting practices) and final Python notebook.

Please submit a PDF of google slides or a document presenting your findings. Upload the PDF onto your Google Drive and share the link as follows:

Google drive link with general access set as 'Anyone with a link' and role set as 'Editor'

Share the link in the Behavioral Survey.

Your evaluation criteria is partially technical and partially the ability to explain meaningful results in a presentable manner.

Natural Language Understanding Trial Task

- Data Sources: Combine different data sources from the below list to (i) get data and (ii) use Python harvester to scrape or download data from the sources. Evaluation will also be based on the diversity of the data chosen.

Open Contracting
Tenders Electronic Daily
Open Tenders
GI Hub Pipelines
California Tenders
Florida Tenders
Texas Tenders
FDOT
TxDOT
Caltran Tenders
Asian Development Bank (ADB) Projects

African Development Bank (AfDB) Projects
Asian Infrastructure Investment Bank (AIIB) Projects

- Learn more about Projects & Tenders standard data structure
    - https://developer.taiyo.ai/api-doc/ProjectsandTenders/
    - https://www.open-contracting.org/data-standard/

Modeling and Report

Use open source Natural Language Models with the above data sources to synthesize your findings addressing the three points mentioned below

1. **Extract entities**. Use Named Entity Recognition (NER) to identify and extract sector, sub-sector, location, or entities like Government Agency, Company Name, Contractors, Investor, or unit measurements such as cost per square kilometer. Ideally using the projects / tenders description and the original PDF document.
2. **Similar projects**. Word2vec and / or cosine similarity for semantic and syntactic for identifying similar projects. For example: For a given project identify all similar projects within the past 10 years within 500 miles
3. **Trends**. Show data visualizations for aggregated time series, bar chart / line chart, where the X-axis is Time and Y-axis is 'Total number of records' and/or 'Total budget/cost'. These charts can be segmented to show different countries or sectors within a country over time.

Bonus question: Find data sources for road projects and tenders in the state of California.

Evaluation is based on the following parameters:

- Extensiveness of the dataset and understanding of projects and tenders data structure
- Modular, DRY Code
- Config Params, Unit Tests & Logging Standards
- Presentation of results and understanding of the problem statement

Behavioral Survey