FourthBrain

GLG Project Proposal

Name of the Organization	Gerson Lehrman Group
Project location (city or timezone)	Austin, TX

A. Project Description

Problem definition
[50-100 word description of the problem which the candidates need to solve]

Automated Meta-data Tagging and Topic Modeling

GLG powers great decisions through our network of experts. Our company receives hundreds of requests a day from clients seeking insights on topics ranging from the airline industry's ability to cope with COVID-19 to the zebra mussel infestations in North America. The goal is to match each request to a topic specialist in our database. This project on Natural Language Processing (NLP) is aimed at improving topic/keyword detection process from the client submitted reports and identifying the underlying patterns in submitted requests over time. The primary challenges include Named Entity Recognition (NER) and Pattern Recognition for Hierarchical Clustering of Topics.

Typically, the client requests we receive comprises a form with unstructured free text with screening questions. Thus, we have a need to group these requests into common topics – to better understand and service demand. This project is aimed to increase the resourcefulness of the current data pipelines for efficient data storage and retrieval.

Key Research Questions/ Technological constraints that the Project will Answer

1. Can we group similar client requests together? (Eg. Google News)
2. Can we perform NER for unstructured data geared towards the Tech Industry or Healthcare industry with reasonable accuracy?
3. Can we find hierarchical patterns in the topics for requests to identify temporal directions of the requests?

Final deliverables at the end of the project
[Please list the desired technical deliverables from the project team in as much detail as possible]

1. A deployable ML model that performs NER with reasonable accuracy.
2. A clustering mechanism to find patterns from submitted topics or requests.
3. A hierarchical clustering method that can produce a hierarchical dendrogram of topics submitted over a period of time.

Key activities/ technologies the project team may be expected to undertake/ work with
[E.g. What kind of technology stack they will work with, the datasets they may need to work on, what kind of analysis they may be expected to undertake, etc.]

1. Named Entity Recognition for short text paragraphs geared towards identification of topics that imply technological or healthcare terms.
2. Unsupervised clustering of time-stamped topics.
3. Hierarchical clustering of topics or temporal sequence learning for identified topics.

There will be two public datasets for this project:

1. https://components.one/datasets/all-the-news-2-news-articles-dataset/ (Effort can be limited to mining a single paragraph of text from these articles only).
2. https://www.kaggle.com/abhinavwalia95/entity-annotated-corpus

Expected learning outcomes

[What do you expect the candidates to learn from the project. Please mention the technical skills they will imbibe over the project.]

1. Data engineering to shape real world data suitable for machine learning
2. Experience into efficient Information storage and retrieval methods.
3. Experience in unsupervised hierarchical clustering of text
4. Semantic similarity modelling
5. Building a machine learning enabled pipeline from raw data to useable insight

Desired Team Size (if any):	Quality not quantity!