# 

# Student Project Proposal

| Project Title | Topic detection model (Information Retrieval) |
| --- | --- |
| Industry Sponsorship (if Any) | Gerson Lehrman Group |

## Project Description

**Problem definition**

| This project is aimed to increase the resourcefulness of the current data pipelines for efficient data storage and retrieval. Build a topic/keyword detection process from unstructured free text with screening questions through Named Entity Recognition (NER). Identify patterns in submitted requests over time for Hierarchical Clustering of Topics (scope: Tech and Healthcare industries). |
| --- |

**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**

| 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**

| 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. Training set : ~14.5 MB <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).   1. Test set : ~8.2 GB, <https://www.kaggle.com/abhinavwalia95/entity-annotated-corpus> |
| --- |

**Expected learning outcomes**

| 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 (and/or semi-supervised) hierarchical clustering of text 4. Semantic similarity modelling 5. Building a machine learning enabled pipeline from raw data to usable insight 6. Explore Human-in-the-loop implementation possibilities |
| --- |

| Team Size: | 2 |
| --- | --- |
| Member names: | Bryan Kim, Daniel S. Lee |

## 

## Tentative Time Plan

Week 6

1. project selection
2. team creation
3. initial exploratory data analysis

Week 7

1. continuation of exploratory data analysis
2. review and replicate similar projects from Kaggle

Week 8

1. submit project proposal
2. test from 2+ subsets of larger dataset
3. research potential use of 2+ additional algorithms (e.g. mini-BERT)

Week 9

1. project baselining and code
2. short slide deck for class
3. begin narrowing down algorithm(s) for use in implementation

Week 10

1. communications with domain expert(s) using summary slides prepared from logging activities and observations
2. communication with other engineers in weekly standup during live session

Week 11

1. share out quantitative results from benchmarking and any additional progress in the project
2. retrain models, identify (and quantify, if possible) limiting conditions, devise plans to overcome challenges
3. start AWS-based data processing
4. refine vision for final deployment
5. prepare for Week 12 review of project v1
6. prepare updated slides with quantitative results and time plan

Week 12

1. Submit either a 10 minute formal presentation or a Technical Report (we intend to be prepared for both)
2. begin project v2

Week 13

1. at least one version of a model on at least one data set deployed on AWS
2. continue documentation of: limitations, observations, solutions, alternatives, etc.
3. implement AutoML tools to benchmark our model(s).
4. experiment with visualization using Slapley, Lime, and/or other options

Week 14

1. continue project v2, with a focus shifting more towards outcomes (contra data and process)
2. explore and record possible extensions/modifications to v2 for future versions
3. fine tuning project v2 in preparation for Final Presentation
4. try to incorporate some concepts from ML Ops, record observations
5. finalize ideas for the Final Presentation

Week 15

1. finish all training runs, generate final deployed outcome
2. finalize all documentation
3. initialize demo presentation to run no longer than 10 minutes

## System Design

Data

* available public datasets for training and news articles for testing

Process

* testing LSTM with word2vec and BERT as initial models
* model deployment in batch (we may explore event-based api deployment, if time permits)
* iterations not yet determined (we will evaluate several ranges of iterations to balance out:
  1. computational resources
  2. accuracy in predictions
  3. adaptability to additional ongoing data inputs

Outcomes

* effective NER with comparable accuracy to existing solutions
* improvement in data pipeline management
* find hierarchical patterns in the topics for requests (extensible to topics beyond the current project scope)
* recommendations based on observations from training models with different algorithms and tuning hyperparameters

Potential Limitations

* improperly formatted user generated requests (e.g. missing punctuation, incorrect capitalization, etc.)
* bad word filtering mechanism failing
* handling extremely large datasets
* managing different data sources and available inputs (e.g. numbers and types of columns, sparsity of data, inaccuracies in data, skewed nature of data, etc.)

Solutions to Potential Limitations

* perform basic root cause analysis when/where possible (and if efficient)
* apply different filtering mechanisms
* contract/expand algorithm to exclude/include more or fewer words
* extract problematic inputs for human evaluation

## Ethical Considerations

Our plan currently requires little modification based on ethical conditions. User input requests will not contain personally identifiable information (PII). In case requests require that, we will need to have a process to handle those inputs, but those are out of the scope of this project.

Anticipated issues that might arise during the process are noted earlier under “Potential Limitations.”