End-2-End System for Document Categorization

Project Name: End-to-End System for Document Categorization
Mentor(s) : Amit Vhatkar, Atul S
Interns Assigned: 3

Abstract:

We often read text coming from different genres like technical, comical, news articles etc. GIven a set of documents it becomes very hard for a person or even a group of people to categorize a bunch of documents and putting time constraints will make life worse. Main aim of this project is to categorize eYIC proposals with predefined categories and with the categories generated by ML algorithm. Going one step ahead one can assign each eYIC proposal to the person having the same set of skills as that of the categories of respective proposal.

Task List:

Task No.	Tasks	No of Days to Complete.
1	Week 1: ● Manually labelling 400+ proposals (2 Days) ● Understanding document categorization (2 Days) ● Starting/Brushing up NLP, RNN, KG (2 Days)	6 Days
2	Week 2: ● Modelling ML based algo on standard dataset (3 Days) ● Modelling KG based algo on standard dataset (3 Days)	6 Days
3	Week 3: ● Implementing DNN based approach on standard dataset (3 Days) ● Fitting KG and ML based model for eYIC proposals (2 Days) ● Categorizing eYIC proposals based on predefined categories (1 Day)	6 Days
4	Week 4: ● Categorizing eYIC proposals based on predefined categories (2 Days) ● Making APIs of all implemented models (1 Day) ● Making UI (3 Days)	6 Days
5	Week 5: ● Backend integration (3 Days) ● Testing of end-to-end system (2 Days)	5 Days
6	Week 6:	5 Days

	● Tuning model parameters for performance improvement (~3 Days) ● Documentation (~ 2 Days)	

Prerequisite:

- Knowledge of neural networks
- Tensorflow and tensorboard.
- Knowledge of NLP
- nltk library
- Django

Software Required:

1. Tensorflow
2. CUDA (if GPU available)
3. Other standard NLP libraries
4. Django

Hardware Required:

1. Laptop (8GB RAM, Descent Graphics Card)
2. Good Internet Connectivity

Deliverables:

1. UI based application which accepts bunch of documents and expected categories
    a. It will output two categories for a document
        i. Programmatically defined category
        ii. Manually defined category
2. Manually labelling category to e-YIC proposals
3. Analysis of the scores obtained over different models

Acceptance Criterion:

1. If and only if all deliverables are completed

Note: Project will be considered successful only after all deliverables are met and all acceptance criterias are met.

References: Nil

Videos (if any):

- NLP
- Django
- KMeans by Andrew Ng
- RNN by Andrew Ng

Papers (if any):

This is just a heads up paper, Interns has to search and present papers on routine basis

- Enriching BERT with Knowledge Graph Embeddings for Document Classification

Internet links/datasheets (if any) :

- Responses categorization blog