IIT BHILAI

MACHINE LEARNING

Company360: A Deep Learning Approach for a complete analysis of companies

Final Project Proposal: Phase 1 Report
Submitted to: Dr. Gagan Raj Gupta
Associate Professor

Team Members:

Moyank Giri	12310830	M.Tech DSAI
Aishika Nandi	12140120	B.Tech CSE

Contents

Data Collection

Various data sources were used for scraping information out of the sites. The details are as follows

- Company Reviews Analysis
  - Aspect Based Sentiment Analysis (ABSA) Dataset [Link].
  - Google News Dataset [Link].
  - Company Reviews from employers [Link]
  - Company Reviews from Blind App [Link]
- Company Reputation Analysis
  - Web Scraping for news from Reputed sites such as MoneyControl (Shown in Code)
  - Extreme Summarization Dataset [Link]
- Company Work-life Analysis
  - Web Scraping employee blogs from renowned sites like ”Quora” (Shown in Code)
  - Multi-News Dataset [Link].

Pre-processing

The pre-processing has currently been applied to a subset of the collected data. Pre-processing majorly includes some commonly used techniques and these are detailed below

- Company Reviews Analysis
  - Punctuation Removal, Stop Word Removal, Stemming, Lemmatization
- Company Reputation Analysis
  - HTML Extraction, Data cleaning, Prefixing, Truncation (if needed), Tokenization
- Company Work-Life Analysis
  - HTML Extraction, Stop word removal, special characters removal, white spaces removal, Tokenisation

Model Training and Pipeline

- Company Reviews Analysis
  - Currently the models are trained and tested on the existing reviews collected from open-source data repositories such as Github.
  - The models trained are Naive Bayes Classifier, Support Vector Machine and Random Forest Classifier

  - The accuracies obtained are NB-classifier: 0.5822660098522168, SVM 0.9894909688013136, Decision tree classifier 0.9973727422003285

- Company Reputation Analysis
  - Currently for news summarization we have trained/ applied 2 models namely GPT2 Model and Transformers
  - The transformer model is tested using the RougeL metric whose value is 0.2092 after 2 epochs

- Company Work-Life Analysis
  - We have applied 1 pre trained model (Bart Large CNN) and trained the transformer model for summarisation similar to Company Reputaion Analysis.

Challenges Faced and Deliverables for Phase 2

- Challenges faced till now are as follows:
  - Every news site such as CrunchBase, Google News, Inc42 etc has its own implementation of tags for frontend WebView which makes it difficult to create a generalized web scraper
  - Various News sites also restrict the direct use of web scraping bots and therefore require additional support via libraries such as selenium.
  - Company Review sites like Glassdoor, AmbitionBox requires sign up before scraping the reviews. This needs to be automated in the next phase.

- Final Deliverables
  - Improvement in the performance of models and testing with other models
  - A fully integrated system which would be able to provide insights into companies via Company reviews, Company News and Blogs/Videos available on the internet.
  - Automated authentication to sites such as Glassdoor, AmbitionBox for insights into company reviews
  - Testing Summarization and Aspect Based Sentiment Analysis models on real-world data scraped from the internet.

Github Repo Link

- Our project can be found here: [Link]

Work Distribution

We have followed the proposed work distribution:

Task	Done by
Company Review Analysis	Moyank Giri, Aishika Nandi
Company Reputation Analysis	Moyank Giri
Exployee Work Life Analysis	Aishika Nandi