# Smart Text Analytics for Information Visualization

**Supervisor**: Dr. Rajendra Prasath

**Team Members :-**
Prashant Kumar Mahanta(S20160010066)
Deepak Kumar (S20160010022)

# Need of work in field of Text Analytics for Information Visualization?

Why read a whole lot of text when you can get the same amount of information by some awesome/interactive visualization.

# Work already been done in this area

Companies such as

- Gramener - A Data Science based company helps Organizations and clients to analyse data and provide insights using information visualization.
- Lexalytics - Uncover insights hidden within mountains of social comments, surveys, reviews, and any other text documents and visualize them

# Task Division

## Task 1 (This semester)

Creating tools for text understanding and analysis.

**Implications:**

- Analyse each part of sentence.
- Import data in NoSQL database and perform query.
- Extract **Attribute - Value(s)** pairs.

## Task 2 (Next Semester)

Developing a framework for Information Visualization.

**Implications:**

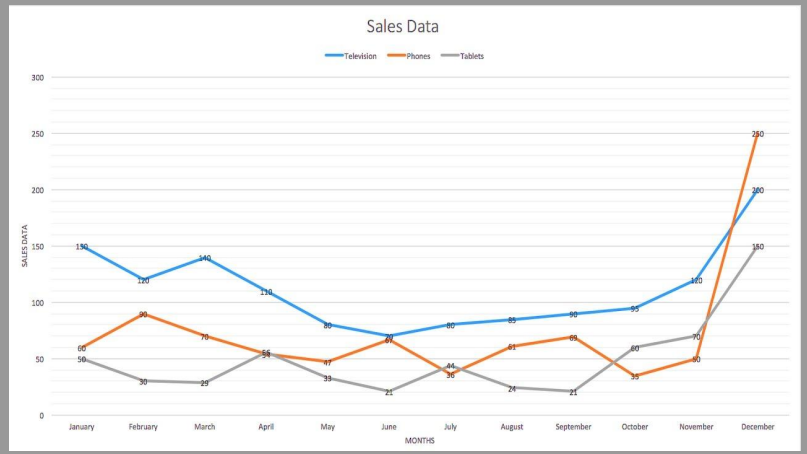
- Visualisation of the data.
- Complete application with interactive graphs and plots.

Text (Unstructured)

Extract (A-V) pairs

Structure the Data

Visualize based on specific features



# Steps involved

# [Working]

Step 1

We break each document into sentences and then use Stanford NER Tagger on individual sentence to get attribute-values pairs.

Step 2

We then store the data in MongoDB ( a NoSQL database) and use Express JS framework to create a web app to perform search queries on data.

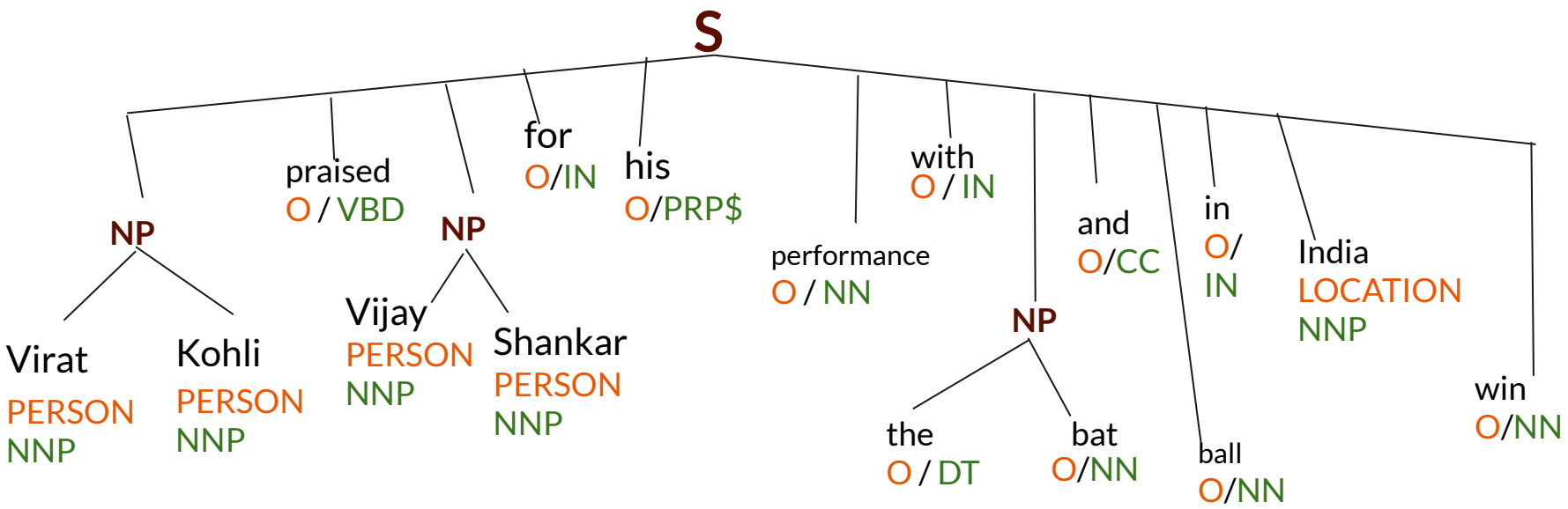
Step 3

We use Stanford NER Tagger on the query and then search in database for that query in a particular attribute section.

# Stanford NER (Named Entity Recognizer) Tagger

- We are using Stanford NER tagger and extracting the POLD (Person Organisation Location Date) for each sentence from an article.
- NER labels sequences of words in a text which are the names of things, such as person, place or company. It comes with well-engineered feature extractors for Named Entity Recognition, and many options for defining feature extractors.

# Virat Kohli praised Vijay Shankar for his performance with the bat and ball in India win




**NP: Noun Phrase**
**NNP: proper noun, singular**

# NER RESULTs

KXIP Team 2019 players list: Complete squad of Kings XI Punjab in IPL 2019. In a bid to revive their fortunes in the Indian Premier League Kings XI Punjab strengthened their squad by snapping up 13 players in the IPL auction - highest amongst eight teams. The Punjab-based team surprised one and all as they shell out Rs 8.4 crore to buy unheralded Virat Kohli after a bidding war that also involved Royal Challengers Banglore, Chennai Super Kings and Rajasthan Royals. KXIP, who had the maximum purse of Rs 36.2 crore going into the auction, also spent big on England all-rounder Sam Curran, buying him for Rs 7.2 crore, the most for an overseas player on Tuesday.

Potential tags:

LOCATION
ORGANIZATION
DATE
MONEY
PERSON
PERCENT
TIME


```
{
  "fileName": "newscrap/article1.txt",
  "Image": "newscrap/Image: 1.jpg",
  "sentences": {
    "PERSON": [
      "Virat Kohli",
      "Sam Curran"
    ],
    "ORGANIZATION": [
      "Kings XI Punjab",
      "IPL",
      "Indian Premier League",
      "Royal Challengers Banglore",
      "Chennai Super Kings",
      "Rajasthan Royals"
    ],
    "LOCATION": [
      "England"
    ],
    "DATE": [
      "2019",
      "Tuesday"
    ],
    "O": [
      "KXIP",
      "Team",
      "players",
      "Complete",
      "squad",...
    ]
  }
}
```

# NEXT

- Finding dependencies in POLD
- Mapping activities (verbs) with corresponding names (nouns)
- Understanding relationships between entities and fixing errors.
- Plotting relationships using graph where each node is some entity.

| | |
|---|---|
| Deliverable 1 | • Parsing news articles and performing NER, identifying phrases. (Stanford NER Tagger).<br>• Export data to NoSQL framework (MongoDB). |
| Deliverable 2 | • Extract information from data.<br>• Text analytic tool (a web app). |
| Deliverable 3 | • Creating graphs and plots using some framework.<br>• Case studies. |
| Deliverable 4 | • Integrating graphs with main web app to get the visualization. |

# Summary

- Extracting correct attributes and value(s) pairs.
- Modifying or correcting values based on our analysis.
- Finding dependency in the text data. Understanding relations between entities of attributes.
- Proper structuring of unstructured text for visualization.

# References


- Stanford NER documentation.
- NLP blogs - https://pythonprogramming.net
- Named Entity Recognition for Unstructured Documents - https://medium.com/@dudsdu/named-entity-recognition-for-unstructured-documents-c325d47c7e3a

# THANK YOU