**Beauty of Unstructured Text**

**Talk will be About**

1. **Introduction of Unstructured Text:** We will start with the introduction of unstructured text, explaining what it means, motivation or solid reason I.e. why to start with unstructured text by discussing applications and if we solve it, what it brings to our table. During motivation, we will also show over role of simple counting/ratios mathematics in developing some naive solutions which we use in our daily life while using internet. Some of sample examples we will talk about

* Google’s Did you mean by, showing results for and next word prediction
* Word of Mouth (Topic Modelling)
* Sentiment Analysis
* Document based
* Aspect/opinion based
* Word vector

1. **Problem, Business case, Solution and Demo:** Now, we will go in depth for one problem, talk about problem statement, business case, solution and further scope. We made a capability for **“Aspect/Opinion Mining”** to analyze the reviews regarding service provided by the company.

* **Problem Statement:** Given a review, containing multiple opinions or multiple features of a product/service, need to come up with opinion of user for each feature/service.

Example: “Battery is good but camera is bad.”

Output**:** {‘battery’:+ve,’camera’:-ve}

* **Business case:**
* Aspect identification has grown to be one of the most active research areas in NLP because of increase in volume of opinionated data recorded in social digital platforms like Facebook, amazon, twitter etc. And companies are much interested to understand the opinionated information about their services/products for decision making. Most of the sentiment analysis algorithms provides opinion of the user at an overall product level but most companies are more concerned about the opinion of the user at product feature level and this is more helpful to the companies to improve their product design and target the audience with better marketing schemes.
* Granularity matters in NLP and it means capturing information at each level of segment of review. For example, A review from hotel industry consisting multiple aspects like food, service, price, ambience etc., system can label the entire review and say that it mentions both food and price. At this level, you would have the most information about the context to make an accurate prediction, but it may require extra steps if you wanted to find out which particular sentence or word is referring to the specific aspect. Complexity of language pattern can be solved based on granularity. At the individual word level, you have the most specificity; maybe a person was dissatisfied particularly with the **music**, which would be a feature of **ambience**.
* **Solution:**
* The real challenge then, is for businesses to parse and organize this amount of data into a more digestible and actionable insight.
* Two fundamental problems in opinion mining are opinion lexicon expansion and opinion target extraction.
* An opinion lexicon is a list of opinion words such as good, excellent, poor, and bad which are used to indicate positive or negative sentiments.
* Opinion targets are topics on which opinions are expressed.
* To achieve the aforementioned tasks, we observed that there are different syntactic relationships between opinion words and targets (features of product/service). Dependency parser helps to identify these syntactic relations and can be used to expand the initial opinion lexicon and to extract targets. This proposed methodology is based on bootstrapping. We named it as double propagation as it propagates relationship information between opinion words and target.
* In this session we summarize the Double propagation algorithm which uses dependency parser to understand syntactic relations and discuss about how this helps in self-learning of opinions of user in a text corpora.
* **Demo/Tools:**
* Designed system accept set of reviews and assume that we have lexicon of target domain, system generates sentiment for each aspect present in review.
* [spaCy](https://spacy.io/usage/) (tokenization, sentence boundary detection, dependency parser, etc.)
* Using NLTK, Wordnet to perform NLP operations.

**Session Structure**

1. Introduction of Unstructured Text – 5 minute
2. Problem and Solution – 20-25 minute
3. Sample Demo – 10 minute
4. Doubts/Discussion – 5-10 minute

**Key Takeaways**

1. How to leverage the syntactical relations present in language to understand the opinionated information.
2. Why unsupervised is better in this case as compared to supervised, as labelled data collection is tedious task for specific domain.