**INDUSTRIAL TRAINING REPORT**

# NATURAL LANGUAGE PROCESSING

at


Samar Dikshit
150953058
CCE - B

# iEnabler

- Specialises in guiding organisations to realise their transformational possibilities
- Helps companies innovate by growing profitable revenue, discovering new products and generating IP, and build new business models
- Provides a product discovery platform for these tasks that uses a set of ‘Think Models’, along with a web search API, local database, and trends database
- Founded by Sridhar DP and Dr. Shankar Venugopal in 2011

# Technical Background

## 1. Artificial Intelligence

- The technology that enables machines to demonstrate human-like intelligence
- Conceptually been around for centuries, formalised as a field of research in 1956
- AI is used for problem solving and optimisation, computer vision, text processing, speech recognition
- Narrow AI: Systems that have learnt to carry out tasks without being explicitly programmed to do so
- General AI: Adaptable intellect systems

# Technical Background

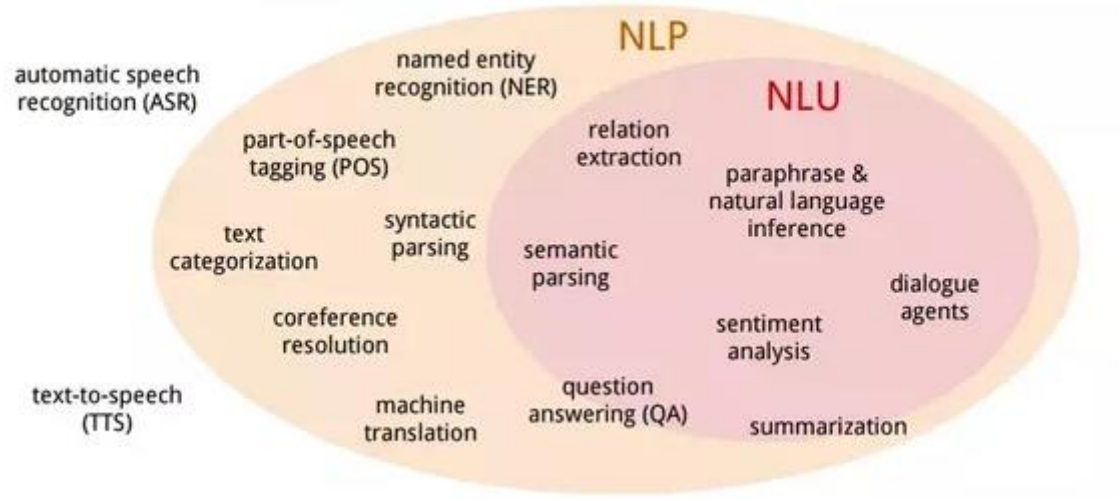
## 2. Machine Learning

- A practical application of AI based on the idea that machines can learn from a given set of data
- Types of ML: supervised, unsupervised, semi-supervised, and reinforcement learning
- Used in situations where designing explicit programming algorithms is infeasible or too difficult
- ML algorithms can fail due to lack of suitable data, improper algorithm selection, and data bias

# Technical Background

## 3. Natural Language Processing

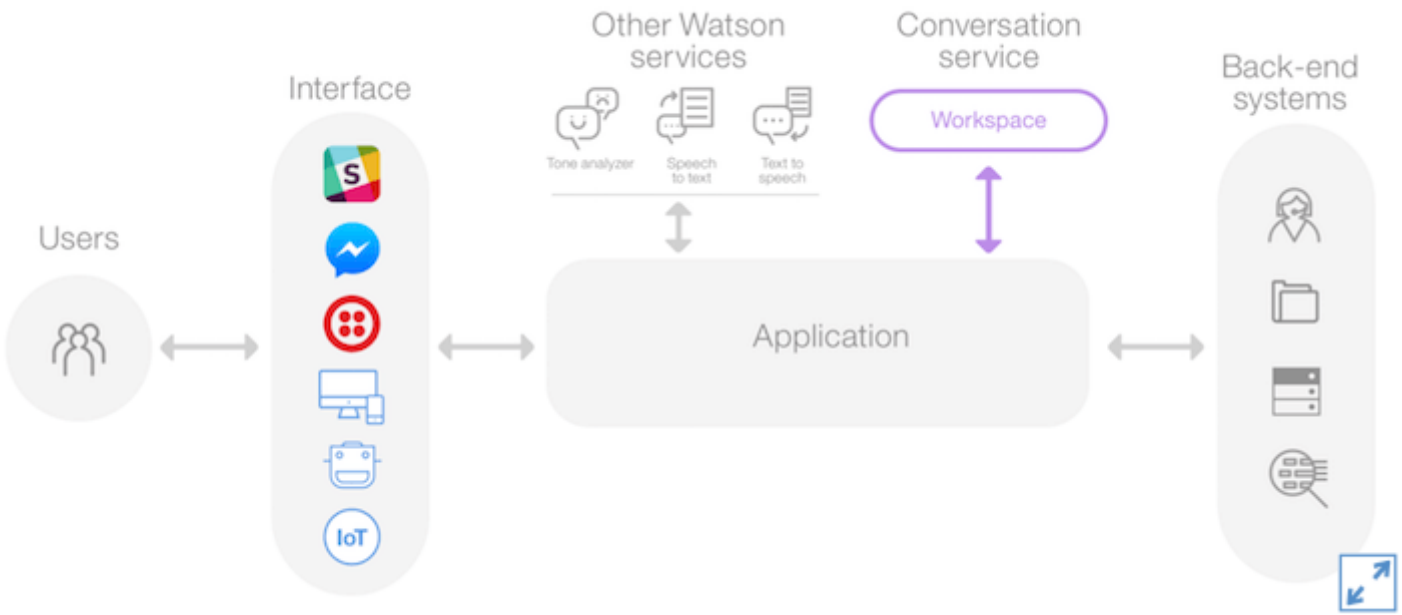
- Linguistics application of ML
- NLP tasks: POS tagging, lemmatisation, semantic parsing, NER, question answering, sentiment analysis
- Inherently non deterministic
- NLU is a subset of NLP that deals with machine reading comprehension
- Implemented in Python using NLTK and TextBlob



# Technical Background

## 4. IBM Watson and Chatbots

- Watson is a natural language system capable of question answering
- Services and APIs offered: Assistant, Discovery, Speech to Text, Text to Speech, Tone Analyser
- Chatbots are used to mimic human speech and simulate conversations
- Based on NLP concepts for conversation, question answering, and reply generation
- Chatbots consist of intents, entities, and dialogs
- Assistant is easy to use, and can be integrated with other Watson services


Users
Interface
Other Watson services
Tone analyzer
Speech to text
Text to speech
Conversation service
Workspace
Application
Back-end systems
IoT

# Methodology

Problem statement: Implement NLP systems on the iEnabler platform to make it easier to use, and more optimised for each user's project.

Solution:

- Use the project title to generate tags which are used while storing details in the local database to optimise local search
- Use the project abstract to generate phrases and keywords which are used by the web search API to return relevant results to the user
- Implement a chatbot to make the platform user friendly
- Create a timeline for further implementation of ML based applications for the platform

# Methodology

## 1. Keywords and Phrases Generation

- NLP system based on Rapid Automatic Keyword Extraction (RAKE)
- RAKE is domain independent, ranks phrases and words in text by analysing their frequency of appearance and its co-occurrence with other words
- The top $1/5^{th}$ of the words and phrases are selected, and are later passed on to the Bing Web Search API
- TextBlob is used to calculate polarity

Input:

```
Enter the text: Geographic Information System is a powerful tool that can be used to locate springs sourced in ophiolites. The unique features
    associated with these springs include a reducing subsurface environment reacting at low temperatures producing high pH, Ca-rich formation
    fluids with high dissolved hydrogen and methane. Because of their unique chemical characteristics, these areas are often associated with
    microbes and are thought to be similar to the features that enabled life to evolve on Earth. Locating and sampling these springs could offer a
     deeper look into Earth's deep biosphere and the history of life on Earth. Springs have traditionally been located using expensive and time
    consuming field techniques. Field work can be dangerous. The goal of this study was to develop a model that could locate these unique
    geological features without first going into the field, thus saving time, money and reducing the risks associated with remote field
    localities. A GIS site suitability analysis works by overlaying existing geo-referenced data into a computer program and adding the different
    data sets after assigning a numerical value to the important fields. For this project, I used surface and ground water maps, geologic maps, a
    soil map, and a fault map for four counties in Northern California. The model has demonstrated that it is possible to use this time of model
    and apply it to a complex geologic area to produce a usable field map for future field work.
```

RAKE and polarity output:

```
['unique geological features without first going', 'gis site suitability analysis works', 'low temperatures producing high ph', 'reducing
    subsurface environment reacting', 'time consuming field techniques', 'unique chemical characteristics', 'high dissolved hydrogen', 'unique
    features associated', 'rich formation fluids', 'overlaying existing geo', 'located using expensive', 'geographic information system', 'thus
    saving time', 'remote field localities']
0.05
```

# Methodology

## 2. Tag Generation

- Done using NLTK’s word tokeniser, POS tagger, and lemmatizer. TextBlob’s noun phrase generator is also used
- Certain types of words, such as coordinating conjunctions and determiners, are removed excluded as they do no provide any real information
- Lemmatized words must also satisfy a minimum length condition
- The tags are not ranked
- TextBlob is used to calculate the polarity

Title:

```
Enter the title: Using GIS Site Suitability Analysis to Study Adaptability and Evolution of
Life: Locating Springs in Mantle Units of Ophiolites
```

Tags and polarity:

```
['site', 'suitability', 'analysis', 'adaptability', 'evolution', 'gis site suitability analysis', 'adaptability', 'evolution']
0.0
```

# Methodology

## 3. Chatbot

- A chatbot can help someone using the iEnabler platform understand the services provided, and clear any doubts regarding the same
- Developed on Watson Assistant
- Demonstrative only, to show how the platform can be made interactive and easier to user
- Sample intents: ‘explain competitor analysis’, ‘how to I perform SWOT analysis’

# Methodology

## 4. Machine Learning Applications Implementation Roadmap

Roadmap lists out the following:

- Potential AI based services that can be offered
- Potential machine learning models for the same
- A timeline for the implementation of these systems