# Team ZeRoS

Reeka Hazarika (UAE) – Free Lancer hazarika.reeka@gmail.com

Swapnil Vishwakarma (India) - SIES Graduate School of Technology

swapnilvishwakarma7@gmail.com

Zyad Al-Azazi (Lebanon) – Lebanese American University

zyadazazi@gmail.com

**Specialization**: Natural Language Processing

## Problem Description:

When companies recruit for any position, they usually end up receiving thousands, if not millions, of resumes. Such a huge number of resumes makes the task of going over all these resumes an extremely difficult and tedious job for HR employees. This made a lot of companies opt for systems that take the necessary information from the candidate after they fill an application with all the required fields. The solution worked greatly for employers; nevertheless, candidates have always found it very illogical to spend tens of hours sharpening their CVs and cover letters only to find out that they must spend another hour or so re-entering all the information they have on their CVs in the designated fields.

### Business Understanding:

It is often observed by HR that the manual process of evaluation of Resumes in bulk which are populated with excess information often becomes tedious and hectic. Therefore, we could automate this process by reading several formats of files (CV). Then using some basic techniques of Natural Language Processing like word parsing, chunking, regex parser and/or Named Entity Recognition to easily capture information like name, email id, address, educational qualification, experience in seconds from a large number of documents.

### Project life cycle along with deadline:

1. Data Understanding and Analysis and Data Cleaning and Transformation (19th of April 2021 – 26th of April).
2. Performing Exploratory Data Analysis and EDA Presentation and Model Choice (27th of May – 3rd of April).
3. Model Building and Final Project Presentation (4th of May – 15th of May).

### Data Intake Report:

A separate document is provided.

### GitHub Repo Link:

zyadalazazi/resume_extraction_team_zeros (github.com)

## Week 8 Deliverables:

### Data Understanding:

Our dataset is unstructured data of text in json format. When the data was imported using the pandas data frame we were able to know more details. The dataset contains two main columns: content and annotation. Content is the main text of the resume, whereas the annotation is the labeling of the information provided in the content. It represents resumes of 200 different people. These resumes include information about the applicants; this information is labeled into different categories: name, location, contact information (indeed account), university/college name, degree, graduation year, years of experience, companies of previous experience, designation and skills.

### Type of Data:

Unstructured data (text) in json format.

### Data Problems:

Since the type of our data is text, there are no outlier data points. Also, many of the issues related to quantitative data distribution, such as skew, or the need to statistically normalize the data are not applicable to our case.

### Solution Approaches:

No need for solutions.

### GitHub Repo Link:

zyadalazazi/resume_extraction_team_zeros (github.com)

## Week 9 Deliverables:

### Problem Description:

When companies recruit for any position, they usually end up receiving thousands, if not millions, of resumes. Such a huge number of resumes makes the task of going over all these resumes an extremely difficult and tedious job for HR employees. This made a lot of companies opt for systems that take the necessary information from the candidate after they fill an application with all the required fields. The solution worked greatly for employers; nevertheless, candidates have always found it very illogical to spend tens of hours sharpening their CVs and cover letters only to find out that they must spend another hour or so re-entering all the information they have on their CVs in the designated fields.

### Data cleansing and transformation done on the data.

## Swapnil Vishwakarma:

Data Cleaning Methods: Worked on expanding contractions, checking null values, removed stop words & lemmatized it.

## Reeka Hazarika:

Data Cleaning Methods: Used Tokenization, POS Tag, Stopwords, Lemmatization , TF-IDF & Word2Vec.

## Zyad Al Azazi:

Data Cleaning Methods: Creating the Functions for Text Cleaning and Word Tagging, cleaning the 'content' column, using the Tf-Idf vectorizer to gain insights about the probabilities of all the possible words and collocations we can find in these resumes, POS, displaying the features table, where columns are the possible mono-, bi- and tri-grams in all of the resumes and trying to identify the most common words and collocations to use in our NER model, later, Tokenization and spacy tagging.

## GitHub Repo Link:

zyadalazazi/resume_extraction_team_zeros (github.com)