Resume Extraction

Name: Munira saud bin munif

Email: muniaralmunif1@gmail.com

City: Saudi Arabia

College: Princess Norah university

Specialization: NLP

**Introduction**

Corporate companies and recruitment agencies process numerous resumes daily. This is no task for humans. An automated intelligent system is required which can take out all the vital information from the unstructured resumes and transform all of them to a common structured format which can then be ranked for a specific job position. Parsed information include name, email address, social profiles, personal websites, years of work experience, work experiences, years of education, education experiences, publications, certifications, volunteer experiences, keywords and finally the cluster of the resume (ex: computer science, human resource, etc.). The parsed information is then stored in a database (NoSQL in this case) for later use. Unlike other unstructured data (ex: email body, web page contents, etc.), resumes are a bit structured. Information is stored in discrete sets. Each set contains data about the person's contact, work experience or education details. In spite of this resumes are difficult to parse. This is because they vary in types of information, their order, writing style, etc. Moreover, they can be written in various formats. Some of the common ones include '.txt', '.pdf', '.doc', '.docx', '.odt', '.rtf' etc. To parse the data from different kinds of resumes effectively and efficiently, the model must not rely on the order or type of data.

**PREPROCESSING**

Data preprocessing is the first and foremost step of natural language processing. Data preprocessing is a technique of data mining which transforms raw data into a comprehensible format. Data from the real world is mostly inadequate, conflicting and contains innumerable errors. The method of Data preprocessing has proven to resolve such issues. Data preprocessing thus further processes the raw data. Data is made to pass through a series of steps in the time of preprocessing:

**Data Cleaning**: Processes, like filling in missing values, smoothing noisy data or resolving inconsistencies, cleanses the data.

**Data Integration**: Data consisting of various representations are clustered together and the clashes between the data are taken care of. **Data Transformation**: Data is distributed, assembled, and theorized. **Data Reduction**: The objective of this step is to present a contracted model in a data warehouse. Data Discretization: In this step, the number of values of an uninterrupted characteristic is reduced by division of the range of intervals of characteristics.

**Natural Language Processing**:

Natural language processing is a branch of artificial intelligence and computational linguistics. It can be defined as the process which is involved in the interaction between a computer and natural language i.e the language, spoken by humans. It is directly related to the field of human-computer interaction. Now that natural language processing is properly defined, we will be using the following constraints of NLP to parse the information from the resumes: I. Lexical Analysis II. Syntactic Analysis III. Semantic Analysis

I**. LEXICAL ANALYS**

IS The pilot stage of the compiler is lexical analysis. The altered source code is taken from the language preprocessor which writes in the form of sentences. The analyzer removes any comments or whitespace from the source code, breaking these syntaxes into a chain of tokens.

Considering our case, the resume is discriminated onto various segments including contact information, educational experiences, work experiences and more. We use a database or a data dictionary to hold the keywords or headings we find common in most of the resumes. Now when a new resume is taken, the parser searches for the keywords and extracts all the

data between the starting and the ending of them, which we call as segments. Out of the many exceptions which might occur, one which is common is that the first segment generally contains the name as well as the contact information of the person. Now we program chunkers or Named Entity Recognizer to extract data from each segment specifically. This method makes the system efficient and reduces its complexity. Now if due to some reason the recognizer runs on a wrong piece of data, the system will produce unexpected results.

**II. SYNTACTIC ANALYSIS**

The syntactic analysis determines the structure of data. The architecture comprises of a hierarchy of expressions, the smallest being a basic symbol and the largest being sentences. We can visualise the architecture as a tree whose nodes represents the expressions. Values stored in the nodes represent the basic symbols. The root represents the sentence

Parse Tree: The parse tree is generated by the parser with syntactic analysis. A parse tree or a parsing tree is an organised, entrenched structure which we use to represent the syntactic analysis of a string. They categorically reflect the syntax of the input data, making them noticeable from the abstract syntax trees used in programming.

II**. S EMANTIC ANALYS IS**

Semantic analysis can be defined as the study of semantics i.e the structure and meaning of speech. This process relates syntactic structure to the level of the writing as a whole from the levels of clauses, phrases, paragraph and sentences. It relates to their language-independent meanings. Let's take an example. Person A has a resume which states he has graduated from the "University of Calcutta" and person B has a resume which says he has graduated from "Calcutta University". Essentially they both graduated from the same place. So what semantic analyzer does is convert "University of Calcutta" to "Calcutta University". In Information Retrieval research, text classification system is given the utmost focus which bounds the decisions to either relevant or non-relevant depending upon the information need of the user. It is not a hard task to get the user information need.

Project lifecycle

The planning phase:

Objective is allowing for the automated storage and analysis of resume data.

## **The execution phase:**

## Package PyPDF2.

## **Language**: Python

## **Data Visualization**: Matplotlib

The closing phase:

Get the resume extraction works.

To Get data

# **Named Entity Recognition: Places**

Next, we want to enrich our data. Ultimately, the goal of structuring data is typically to perform some kind of analysis or visualization — in the case of this international conflict information, it would be valuable to plot the information geographically. To do this, we need coordinates corresponding to the documents.

First, we will use **natural language processing (NLP)** and **named entity recognition (NER)** to extract place-names from the text.

# **Named Entity Recognition: People**

Next, we will extract the names of people mentioned in the document. To do this, we will again use the NER algorithms from the **NER-D python library**.

## Get the Full Names

In the final structured data, I only want full names. Wouldn’t it be confusing to find a data entry with a ‘mentioned person’ of “Jack” or “John”? To accomplish this, we will once again employ some rudimentary statistics. The function will track full names when they are mentioned, usually in the beginning of the text.

When a partial name is mentioned later, it will reference the list of full names to identify who the partial name is referencing. For example, if a news article read as follows: ‘Joe Biden is running for President. Joe is best known as the Vice President for former President Barrack Obama.’ We know that **Joe** is referencing **Joe Biden**, because his full name was given earlier in the text. This function will operate in that same way.

## De-Conflict Similar Names

In the case of duplicates, the function will use the same statistics used earlier for the country function. It will measure a count of how many times a name was mentioned, and use that as the most likely identifier. Example: ‘Joe Biden and his son, Hunter Biden, are popular US politicians. Joe Biden is the former VP. Biden is now making a run for president against incumbent Donald Trump’ We know that **‘Biden’** is referencing **‘Joe Biden’** from context. The passage is clearly about Joe Biden, not Hunter Biden, based on the statistical focus of the text.