MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING FOR INTELLIGENT HIRING WITH RESUME PARSER AND RANKING

Akash Sharma, Saurabh Sharma
Global Nature Care Sangathans Group of Institutions

Abstract

Using Natural Language Processing (NLP) and Machine Learning (ML), this intelligent system ranks resumes of any format in accordance with the provided limitations or the following requirements provided by the client company. The majority of the input resumes will come from the client organization, which will also provide the specifications and constraints used by our system to rank the resumes. Our algorithm will also look at the candidates' social media accounts (such as those on LinkedIn, Github, etc.), which will provide more precise information about the candidate, in addition to the data it gathers from resumes.

Keywords: Resume parser; Indexer; Social Profiles; JSON resume; data-dictionary; chunkers; Segmentation; Semantic Analysis.

Problem Definition

Developing an automated process to extract information from haphazardly organized resumes and transform it into a more organized one. such resumes are graded based on the data gathered, the candidate's skill set, and the organization's job description.

Introduction

Large organizations and employment agencies receive, process, and manage tens of thousands of applications from job applicants. These resumes will be processed automatically by the information extraction system. Names, phone numbers, email addresses, qualifications, experience, and skill sets are just a few examples of the types of extracted data that can be stored as structured data in a database and used in numerous various areas and categories.

Contrary to many non-structured document formats, resume content is relatively organized and is comprised in distinct parts. Each block contains crucial information on a person's contacts, qualifications, and employment background. Even in the confined area and with some structure, resume papers are extremely difficult to mechanically parse. They frequently vary in terms of the information types, order, etc., and may have phrases that are entirely complete or partially incomplete. Additionally, when converting from other document formats like pdf, doc, docx, etc. to text, information in unexpected formats is created. For the system to evaluate resumes accurately and rapidly, it should be unaffected by their order or type of content. With segments at the top level, we assume that resumes have a three-level hierarchical structure. These bits of related informational content make up these portions. A block may include several named objects known as chunks.

Literature Review

First-generation hiring practices: In this approach, job openings were advertised and application invitations were distributed. Publication techniques include word of mouth, television, and newspapers. Those that remained interested would then send in their resumes. After receiving and sorting these resumes, the hiring team called the candidates who made the short list for the next round of interviews. It will take a lot of time and labor to find the finest person for each job function.

Second Generation employment Systems: As business sectors have grown, so have the employment needs within them. To address these job demands, specific consulting units were created. They offered a system via which applicants could send their application materials to the organization by uploading them in a particular format. These agencies would then use particular keywords to find the prospects. These agencies were mid-level organizations that stood between the candidate and the business. These systems were rigid, requiring candidates to submit their resumes in a specific format that changed from system to system.

We advise using the third generation hiring system, which enables job seekers to upload their resumes in a number of different formats. Then, in a certain format, our system assesses, indexes, and stores these resumes. This makes our search process easier. Natural language processing is a component of the artificial intelligence (AI) algorithm utilized by the analysis system. By understanding the candidate's natural language/format, it reads resumes and transforms them into a certain format. The knowledge base contains the learned data. After acquiring further information about the candidate from his social media accounts, including LinkedIn and Github, the algorithm refreshes the knowledge base.

Ranking Attributes are:

1. Current Compensation
2. Expected Compensation
3. Education
4. Specialization
5. Location
6. Earliest Start Date

7. Total Experience
8. Relevant Experience
9. Communication
10. Current Employer
11. Stability
12. Work Gap.

System Architecture

The System Architecture consists of two modules:

1. Outer World System
2. Resume Ranking System

Outer World System Consists of:

1. Client Company.
2. System C.V"s Data base.
3. Social Profile.

Resume Ranking System Consists Of:

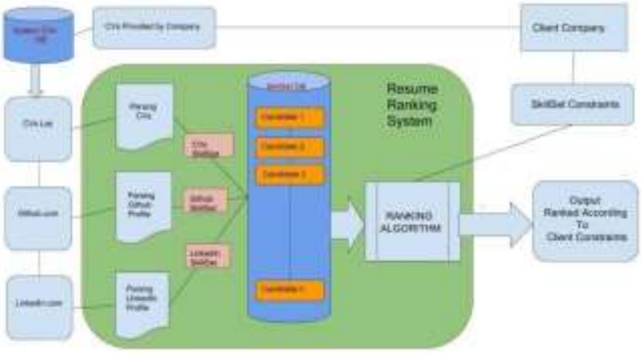
1. Parser System.
2. Candidate Skill-set Database.
3. Resume Ranking algorithm.

Outer World System Consists Of:

Client Company:
This is the client company who will provide us the bulk of the resume or C.V"s with the specific requirement and constraints, according to which it should be ranked.

System C. V's Database:
This is the large database which is used to store the bulk of resumes provided by the client company in a distributed environment.

Social Profiles:
Social Profiles include LinkedIn Profile of the candidate, Github Profile of the Candidate. This social profile module can be extended to different community too.



Resume Ranking System Consists Of: Parser System:

Parsing system includes the parsing of the following candidate resume and their social profiles using NLP. That is without any manual interaction. Here, using Natural Language Processing the this is how we are going to parse the resume one at a time. NLP (Natural Language Processing) requires following constraint for parsing:

- Lexical Analysis
- Syntactic Analysis
- Semantic Analysis

Lexical Analysis:

Text Segmentation stage do work on the fact that each heading in a resume contains a block of related information following it. So in that case our resume will segregate out into segments named as contact information, education information, professional details and personal information segment.
A data-dictionary is used to store common headings in a resume which are definitely occurring in a resume. These headings are then searched in a given resume to find segments of related matching information. All of the text information available between the start and the end of the heading is then accepted as a segment. One exception that will possibly or may occur, is the first segment which holds the name of the person and generally the contact information. It is found by extracting the text between the top and the first heading of the document. For each segment there is a group of Named-Entity Recognizers, called chunkers, that works only for that segment. This improves the performance and reduce the complexity of the system, since a certain group of chunkers only works for a given segment. If there is an error in the segmentation module, chunkers will run on a wrong context. This will produce unexpected results.

Syntactic Analysis:

The objective of the syntactic analysis is to find the syntactic form of structure of the sentence. It is also called as Hierarchical analysis/Parsing, used to recognize a sentence, to allocate token groups into grammatical sentences and to assign a syntactic structure to it.

Parse tree:

Parser generates the parse tree with the help of syntactic analysis. A parse tree or parsing tree is an ordered, rooted tree that represents the syntactic structure of a string according to some context free grammar.

Semantic Analysis:

Semantic Analysis is related to create the representations presentations for meaning of linguistics inputs. It deals with how to determine the meaning of the sentence from the meaning of its parts. Some phrases contains multiple meaning For example, 'University of ABC' is converted to 'ABC University', "Go ahead I am holding your back". The focus in Information Retrieval research lays on text classification systems which make binary decisions for text document as either relevant or non-relevant with respect to a user's information need. Capturing the user information need is not a trivial task.

Candidate Skill-set Databases:

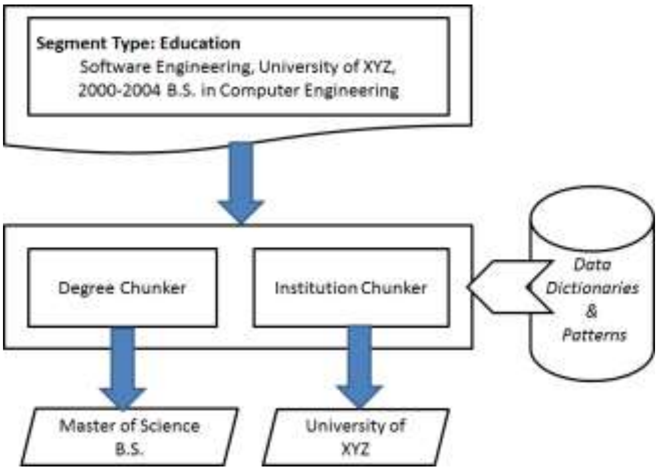
$$F\text{-measure}=2^*\left[\frac{(Precision * Recall)}{(Precision + Recall)}\right]$$

#(relevant items)

We are extracting the information from the candidate resume and storing it in the JSON format. As a database constraint, we are using the PostgresSQL to store the information extracted from the candidate's resume.

Ranking Algorithm:

Each candidate will be scored based on the skillset, experience and project. Scoring will also be influenced by his Github and linkedIn profile.


Segment Type: Education
Software Engineering, University of XYZ,
2000-2004 B.S. in Computer Engineering
Degree Chunker
Institution Chunker
Data Dictionaries & Patterns
Master of Science B.S.
University of XYZ

The focus in Information Retrieval research lays on text classification systems which make binary decisions for text document as either relevant or non-relevant with respect to a user's information need. We used precision, recall and F-measure metrics for performance evaluation.

Performance Measures:

Precision measures the number of relevant items retrieved as a percentage of the total number of items retrieved.

$$Precision=\frac{\#(relevant\ items\ retrieved)}{\#(retrieved\ items)}$$

Recall measures the number of relevant items retrieved percentage of the number of relevant items in the collection.

The F-measure is the harmonic mean of precision and recall.

Expected Outcome

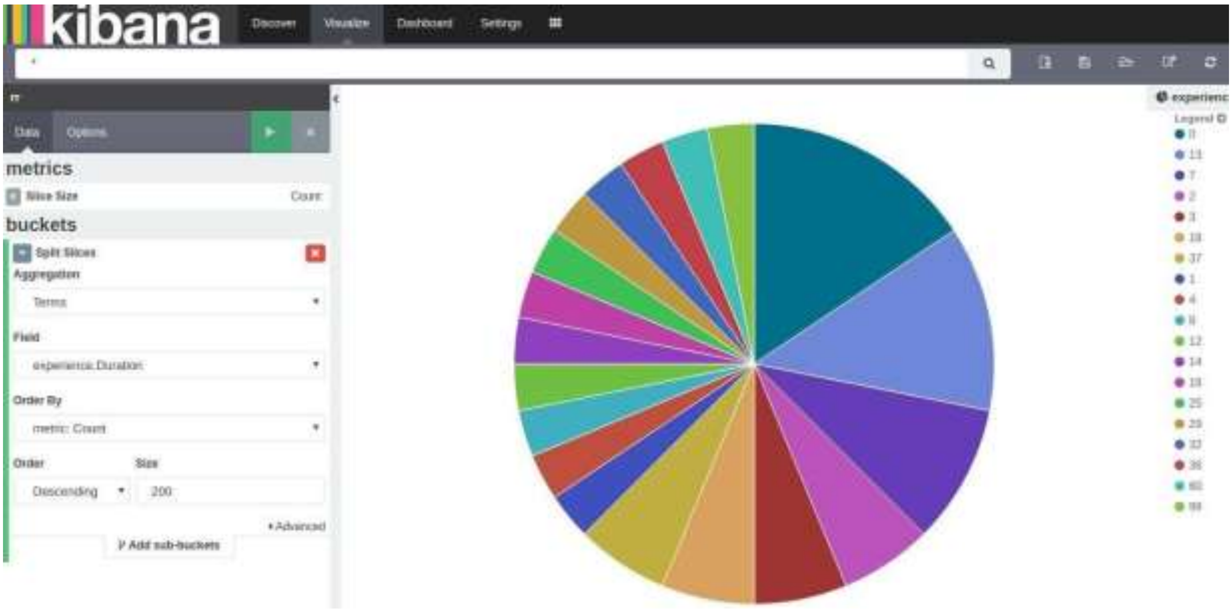
Our system will satisfy both employer and candidate. This online tool has been able to reduce lots of burden on the head of candidate/employee in Online Recruitment System(ORS). Maintain the basic information of employees in the Company/Organization. Put simply, Artificial Intelligence or "AI" is an add-on to system, complementing to provide the online recruitment solution . As the name suggests, AI enables a combination of an applicant-tracking system(ATS) and an artificial intelligence resume parsing, searching and ranking engine. The result is a super charged tool giving incredibly accurate and potential candidate matching to job description, and „talent pool" searching that makes other systems look like they"re from the stone-age.

Our System output in JSON format:

{'education': [{'College': u'Dawoodbhoy Fazalbhoy high school',
'Degree': u'SSC, 2000 - 2010',
'Duration': None, 'Grade': None},
{'College': u'Anjuman-I-Islam Kalsekar Technical Campus',
'Degree': u"Bachelor's degree, Computer Engineering, 2012 - 2016", 'Duration': None,
'Grade': u'6.89cgpa'}],
'experience': [{'Company': u'Vizista Technologies',
'Duration': u'2 months',
'Role': u'PHP Developer ',
'job': u'In Vizista, I worked as an intern for trainee PHP developer. During internship I worked on Hospital Management Sys. And Within a period of one and a half month, me and my friend have successfully completed their project. With appreciation'}],
'personal': {'current_designation': None, 'email': None,
'first_name': u'Afzal', 'last_name': u'Juneja'},
'project': [{'Description': u"Here I've created a decision tree using id3 algorithm. A prediction algorithm for artificial intelligence. Basically, it was an application of id3 algorithm and it was working quite well for n number of datasets. ",
'Duration': u'November 2015 to Present', 'Members': u'Members:Afzal Juneja',
'Name': u'Decision Tree Implementation(id3 algorithm)'},
{'Description': u"I've written some of the artificial intelligence program using python scripting. Like hill climbing, Alpha beta pruning. ",
'Duration': u'November 2015 to Present', 'Members': u'Members:Afzal Juneja',
'Name': u'Python Programs for artificial Intelligence'},
],
'skills': {
u'Natural Language Processing': 30, u'Python': 50,
u'django': 90}, 'summary': None}

Simulation Results

The results involves parsing a resume in /pdf/doc/docx/rtf format into plain json, Github and StackOverflow information is also extracted which will influence the result of individual skills. Info graphics Resume is generated with all the stats and query can be made to visualize the data.



metrics

Slice Size Count

buckets

Split Slices

Aggregation: Terms

Field: experience.Duration

Order By: metric: Count

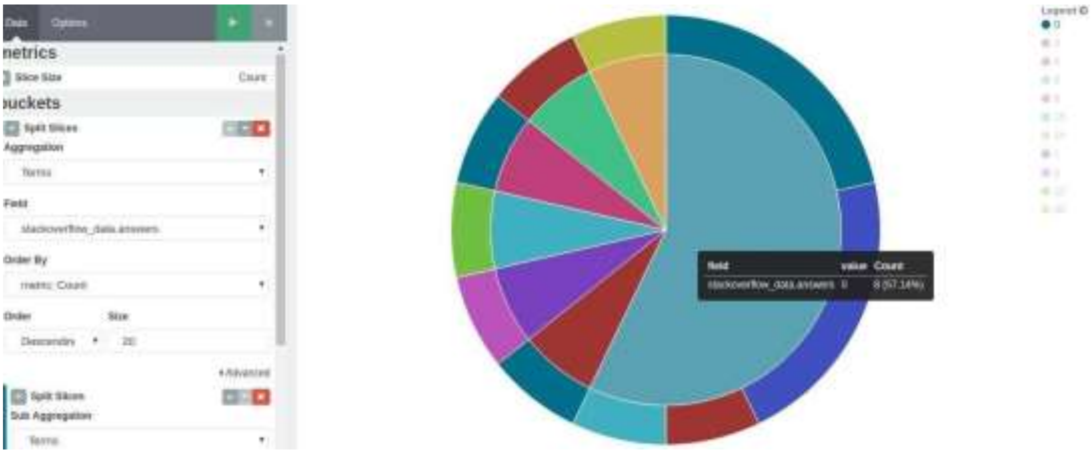
Order: Descending

Size: 200

Advanced

Add sub-buckets

Work Experience distribution of various candidates.



metrics

Slice Size Count

buckets

Split Slices

Aggregation: Terms

Field: stackoverflow_data.answers

Order By: metric: Count

Order: Descending

Size: 20

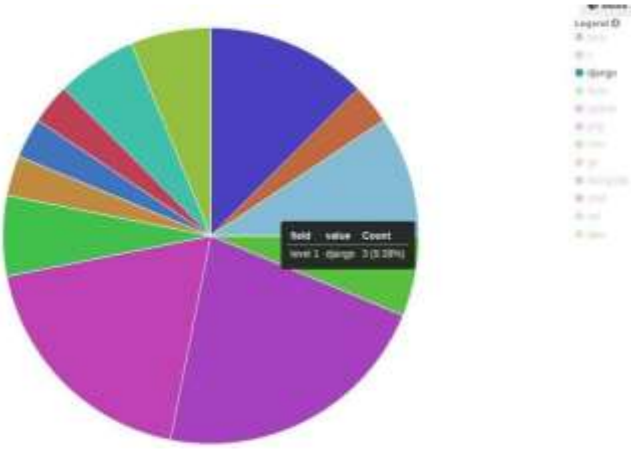
Advanced

Split Slices

Sub Aggregation: Terms

field	value	Count
stackoverflow_data.answers	0	8 (57.14%)

StackoverFlow Question and Answer pie chart.



field	value	Count
level 1	django	3 (5.26%)

Candidates skills aggregated

Conclusion and Future Work

We were able to extract user information from github and stackoverflow and parse the resume. We ranked each user's specific skills based on the available data. The parser's precision might be increased, and information from Twitter and Facebook could provide psychometric information about the individual. A competitive programming website will aid in finding the ideal applicant for a given job profile.

References

[1]. (Jan 2019). Talent Shortage of Software Developers. https://fullscale.io/talent- shortage- software-developers/

[2]. Glassdoor Team . (January 20, 2015). 50 HR & Recruiting Stats That Make You Think. https://www.glassdoor.com/employers/blog/50-hr-recruiting-stats-make-think/

[3]. Bika, N. Recruiting costs FAQ: Budget and cost per hire. https://resources.workable.com/tutorial/faq-recruitment-budget-metrics

[4]. Shehu, M. A., & Saeed, F. (2016). An adaptive personnel selection model for recruitment using domain-driven data mining. Journal of Theoretical and Applied Information Technology, 91(1), 117. ISSN 1992-8645

[5]. Zaman, E. A. K., Kamal, A. F. A., Mohamed, A., Ahmad, A., & Zamri, R. A. Z. R. M. (2018, August). Staff Employment Platform (StEP) Using Job Profiling Analytics. In International Conference on Soft Computing in Data Science (pp. 387-401). Springer,Singapore. http://doi- org-443.webvpn.fjmu.edu.cn/10.1007/978-981-13-3441-2_30

[6]. Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. Journal of machine learning research, 3(Feb), 1137-1155.

[7]. Fernández-Reyes, F. C., & Shinde, S. (2019). CV Retrieval System based on job description matching using hybrid word embeddings. Computer Speech & Language, 56, 73-79. https://doi.org/10.1016/j.csl.2019.01.003

[8]. Keenan, P., McGarraghy, S., McNamara, C., Phelan, M., & Schools, U. B. (2004). Human resource management DSS. In International Conference DSS2004 (pp. 525- 534). Corpus ID: 1638639

[9]. Faliagka, E., Ramantas, K., Tsakalidis, A., & Tzimas, G. (2012, May). Application of machine learning algorithms to an online recruitment system. In Proc. International Conference on Internet and Web Applications and Services. ISBN: 978-1-61208-200-4