REVA UNIVERSITY

Bengaluru, India

A Project Report on

RESUME CLASSIFICATION AND SCORING USING NLP

Submitted in partial fulfilment for award of degree of

Master of Business Administration

In **Business Analytics**

Submitted by

Parimala Mudimela

R17DM014

Under the Guidance of

Dr. J.B. Simha

Mentor

CTO, ABIBA Systems

REVA Academy for Corporate Excellence

REVA University

Rukmini Knowledge Park, Kattigenahalli,

Yelahanka, Bangalore – 560064

October, 2020

REVA
UNIVERSITY
Bengaluru, India

Candidate’s Declaration

I, Parimala Mudimela hereby declare that I have completed the project work towards the Master of Business Administration in Business Analytics at, REVA University on the topic entitled Resume Classification and Scoring using NLP under the supervision of Jay Bharatheesh Simha. This report embodies the original work done by me in partial fulfilment of the requirements for the award of degree for the academic year 2020.

Place: Bengaluru
Name of the Student: Parimala M

Date:07-Oct-2020
Signature: M.Parimela

REVA
UNIVERSITY
Bengaluru, India

Certificate

This is to Certify that the PROJECT work entitled Resume classification and scoring using NLP carried out by Parimala Mudimela with SRN R17DM014, is a bonafide student of REVA University, is submitting the project report in fulfilment for the award of Master of Business Administration in Business Analytics during the academic year 2020. The Project report has been tested for plagiarism, and has passed the plagiarism test with the similarity score less than 15%. The project report has been approved as it satisfies the academic requirements in respect of PROJECT work prescribed for the said Degree.

<u>Signature of the Guide</u>

ಜೆ. ಬಿ. ಸಿಂಹ

Dr. J B Simha
Guide
CTO, ABIBA Systems

<u>Signature of the Director</u>

Name of the Director

Director

External Viva(Taken Virtually)

Names of the Examiners

1. Ravi Shukla ,Sr. Advisor and Data Scientist, Dell.
2. Krishna Kumar Tiwari , Senior Data Scientist ,CEO, AI/ML, Jio.

Place: Bengaluru
Date:07-10-2020

REVA
UNIVERSITY
Bengaluru, India

Acknowledgement

I would like to take this opportunity to thank my mentor and guide Dr. JB Simha who has supported me and guided me in successful completion of my project. I am extremely grateful to him who had always helped me in overcoming the obstacles and guided me in the right direction.

I would also like to thank and express my gratitude towards Dr. Shinu Abhi who had been a great support and for the continuous encouragement provided throughout the program.

I would like to express my special thanks to the support provided by Chancellor ,
Dr. P.ShyamaRaju, Dr. S.Y.Kulkarni, Ex Vice Chancellor, Dr. K. Mallikharjuna Babu, Vice Chancellor and Dr. M. Dhanamjaya, Registrar.

I would like to thank my family and friends who were constant support to me through this entire journey.

Place: Bengaluru
Date:07-10-2020

REVA UNIVERSITY
Bengaluru, India

Similarity Index Report

Title of the Thesis: Resume Classification and Scoring using NLP

Total No. of Pages:28

Name of the Student: Parimala Mudimela

Name of the Guide(s):Dr. JB Simha

This is to certify that the above thesis was scanned for similarity detection. Process and outcome is given below.

Software Used: Turnitin
Date of Report Generation: 18-09-2020
Similarity Index in %: 14%
Total word count: 3672

Place: Bengaluru

Name of the Student: Parimala M

Date: 07-09-2020

Signature : M.Parimela

Verified By:

Signature

Dr. Shinu Abhi, Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	NLP	Natural Language Processing
2	NLTK	Natural Language Tool Kit
3	NER	Named Entity Recognition

List of Figures

List of Tables

Abstract

Resume screening is the system of locating if a candidate is certified for a function primarily based totally on his or her education, experience, and different facts cited on their resume.

In different words, it’s a shape of sample matching among a process description and the qualifications of a candidate primarily based totally on their CV. (Amin et al., 2019)

Resume screening is as yet the most tedious piece of enlisting: screening resumes is assessed to take as long as 23 hours for only one recruit. At the point when an employment opportunity gets 250 continues all things considered and 75% to 88% of them are inadequate, it's no big surprise most of ability procurement pioneers despite everything locate the hardest some portion of enlistment is screening the correct competitors from an enormous candidate pool. The time spent on screening resumes often the maximum time of the metric time-to-fill.

The web application developed in this project for resume parsing and scoring helps the recruiters in saving the time and budget of the company. In whole process of talent acquisition, finding the suitable candidates for the job requisition and taking them to the next level in the hiring process consumes more time. The designed system helps the recruiters to achieve this task easily. It will help them to find the suitable candidates from an enormous resume database. The resume scoring system uses techniques like machine learning and NLP in the backend and for the frontend user interface is developed using HTML, CSS and java script. Python flask framework is used to deploy the application and make it up and running.

Keywords: Renege, Classification Model, Logistic Regression

Contents

Chapter 1: Introduction

Sourcing and screening the resumes is the initial step in any hiring process. Often recruiters end up spending most of their time in this activity. After a recruiter posts job in the internet, usually hundreds or thousands of resumes are received. Picking up the right candidates from a huge number of resumes received who have relevant experience and skill sets that match the job post is a very tedious task for the recruiter. Every candidate writes resume in their own format and style. They do not showcase their skill sets and experience properly. This poses a huge challenge for the recruiters to screen all the received resumes and find the suitable candidate. At the minimum, recruiter usually spends 3 to 5 minutes per resume and at that rate he spends he amount of time to review hundreds of resumes. At times he/she can miss potential cnadidates and might select wrong candidates. The recruiters may be biased at times. All these problems can be solved developing a web application that automatically screens the resumes and rank them.

In the recent years, many recruitement websites have been built to address this issue. These on-line applications have used different methods to shortlist the candidates suitable to a job description. (Amin et al., 2019). Few of them applied different selection category strategies that allows you to categorize the candidate profiles into numerous classes for a particular job requisition. In those processes, each candidate CV is attempted to fit with each given process posting at the recruitment site. The intention recruitment web sites is to throw up the effects to the candidate to which they may be pleasant suit into.(Amin et al., 2019) The strategies utilized by those websites has led to excessive accuracy and precision, however the important risk is the issue of time. If each candidate resume is matched with each different process posting given on the net company career site, the time multiciplity for obtaining the effects could be very excessive.

Chapter 2: Literature Review

Hiring is a vital, complex, and effort-intensive function within Human Resources department. According to a survey,TCS currently hires about 400,000 employees and as per the Annual Reportof TCS 2015-16, in the fiscal year 2015-2016, TCS employed 90,182 people (about 97% with IT background), 74,009 in India and 16,173 abroad. Taking as an assumption of 20% selection rate, on an average 5 people were interviewed. For shortlisting 5 persons per post, HR executives have to screen manually at least 10 resumes per requirement. Taking in to account an average effort of 350 seconds per profile, this interprets 70,000 man hours anually., spent solely for resume screeningand shortlisting the prospective candiates for the interview. After going through so may a much tedious task, the manually evaluating the resumes is often error-prone, opaque, biased, does not facilitate comparisons, and opportunities for improving the quality of recruitment are difficult.(Palshikar et al., 2018)

Multiple trials were performed to automate several operations of talent acquisition. (Lee, 2007) For example, (Laumer & Eckhardt, 2009) recommends methods such as “collaborative filtering”(Singh et al., 2010) to shortlist the resumes suitable for a job a requirement. (Yi et al., 2007) defines a technique which uses “relevance models” to match the text content to match the resumes with the job requisition. After that, candidate profiles that match with the job postings are used to capture semantics that is not clearly cited in the job profile requirements. The approaches are based on assumption that the resumes are having manually labeled rankings.

In (Faerber et al., 2003) Collaborative Filtering and content-based similarity measures are combined for more accurate results in ranking. All the past research have been done on artificially generated data and di dnot use unstructed resume documents and job requisition document.(Singh et al., 2010) (Amin et al., 2019)

Chapter 3: Problem Statement

“Resume screening is as yet considered the most tedious piece of enrolling: screening resumes is assessed to take as long as 23 hours for only one employment profile. At the point when a job post gets 250 continues all things taken in to account and 75% to 88% of them are unfit, it's no big surprise most of ability securing pioneers despite everything locate the hardest some portion of enrolment is screening the correct up-and-comers from an enormous candidate pool. Exacerbating the issue, an ongoing review of ability procurement pioneers found that 56% will expand their employing volume one year from now, however 66% of selecting groups will either remain a similar size or shrink."(Resume Screening, n.d.)
"The time spent on screening resumes regularly takes up the biggest part of time-to-fill. With the present serious applicant driven ability market, top ability just remains available for 10 days on average"

Chapter 4: Objectives of the Study

The objective of this project is to develop an online software that will give the output as list of ranked resumes for a specific job requisition. This web application automates the process of fetching the profiles, screening them and ranking them. This system provides the facility to fetch the resumes from different job portals like workday, LinkedIn, indeed etc. For example, if the users or recruiters select workday, it provides various filters like region, country, requisition id to further drill down. The user can choose the job description file and can choose what key skills he is looking for in the candidates.

The output will be displayed in a tabular format showing the various details of the shortlisted candidates like email, phone number and the corresponding ranking. The recruiter has an option to click on the resume and view the resume he is interested in. He can download the resume as well. The table having the details also can be downloaded and exported to csv file.

The core engine of resume screening and ranking is developed using techniques like Natural Language Processing. The collected resumes are parsed and information is extracted using different techniques in NLP. The rank is calculated by assigning a weightage to each of the attributes in the resume like Experience, skillsets, Education etc.

Chapter 5: Project Methodology

The method proposed in this work offers an approach for screening the resumes based on a job description. This is a web application to help recruiters by screening the CVs, shortlisting candidates that best match the position, and filtering out those who don't.
Resumes are ranked based on the score in comparison with the job description.

The Proposed system consists of the following components:

1. Resume Collection
2. Resume Parser system
3. Candidate skill set database system
4. Ranking algorithm

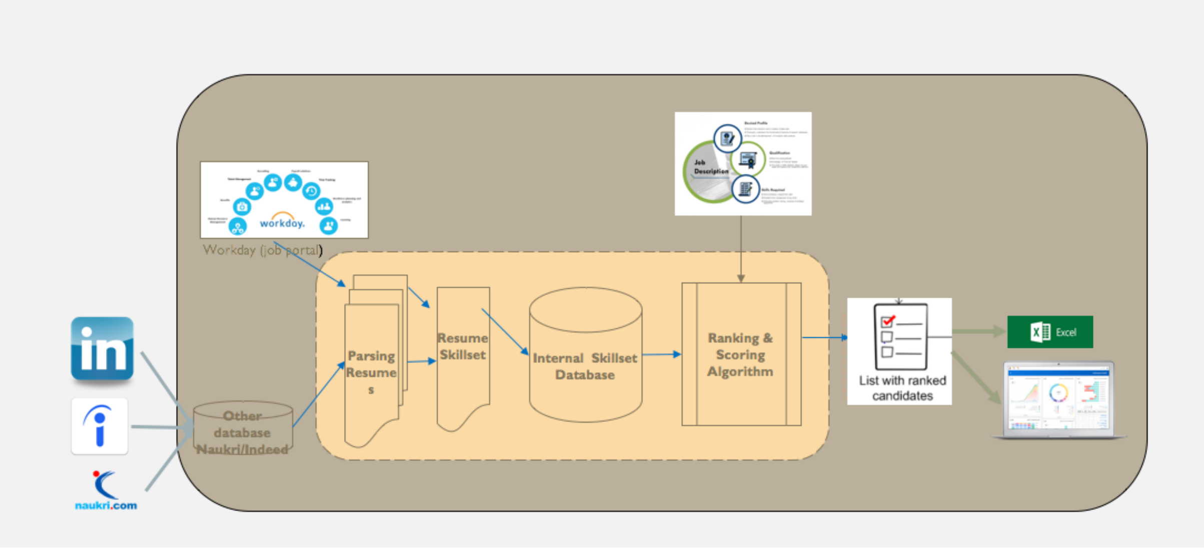

Workday (job portal)
workday.
Job Description
in
naukri.com
Other database Naukri/Indeed
Parsing Resumes
Resume Skillset
Internal Skillset Database
Ranking & Scoring Algorithm
List with ranked candidates
Excel

Fig-1

Resumes are collected from Workday using API(Application Programmable Interface). A python script is written to communicate with the Workday API and extract the resume data.
Data are collected from LinkedIn using Selenium(mimics human behavior).
Web application accepts job description as the input and based on the selection of the job portal, the resumes are collected by the back end code.

The next step is the resume parsing. The CVs will be of different formats like word .doc, .Docx, .pdf, etc. We convert all these CVs into text format which will be easy to analyze.

Resume Parser system

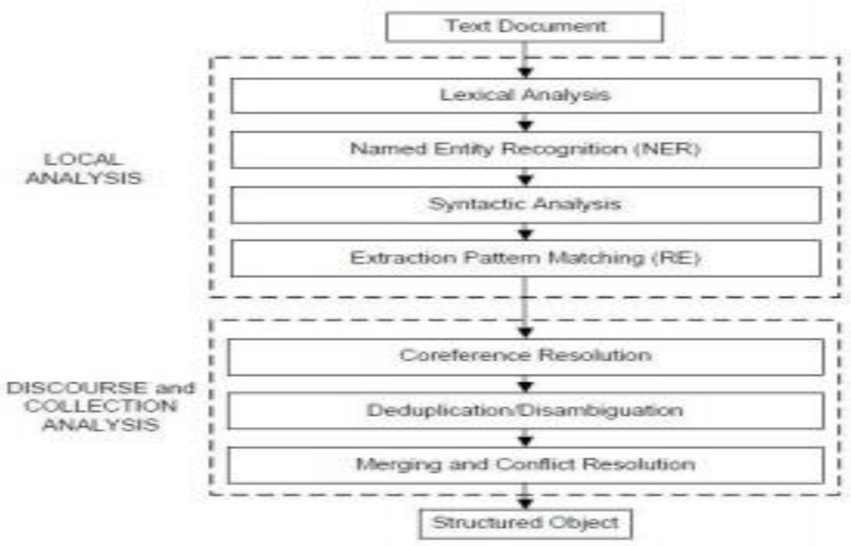

Text Document
LOCAL ANALYSIS
Lexical Analysis
Named Entity Recognition (NER)
Syntactic Analysis
Extraction Pattern Matching (RE)
DISCOURSE and COLLECTION ANALYSIS
Coreference Resolution
Deduplication/Disambiguation
Merging and Conflict Resolution
Structured Object

Fig 2

The resumes are ranked by assigning weightage to different attributes in the resume like skills, Experience, Education, Projects worked, etc, and compared with the job description.
Scores are assigned based on the match between resume data components and Job description data components.

Chapter 6: Business Understanding

Automation of Resume Screening and ranking improves the efficiency of the recruitment by automating the tedious task of parsinghuge amount huge amount of resumes for a single job post.It also helps to shortlist the suitable candidates that can be taken to the next step in the HR TA process.

The benefits of automated resume screening system

1. Increase profitability and performance.
2. Attract better candidates for the job profile.
3. Time and Cost of recruitment are saved.
4. metricslike time to fill and time to hire are improved.

(Using Effective Recruitment to Gain Competitive Advantage, n.d.)

Chapter 7: Data Understanding

Data comprises of unstructured resumes which are in the form of .doc or .docx or pdf format.
Resumes are collected from various job portals like Workday, LinkedIn, Naukri, etc.
Job Description which is in the format of .doc or .docx or .pdf.
Unstructured resumes are converted into structured text format.

Have used manually annotated resumes from recruiters as training data set.

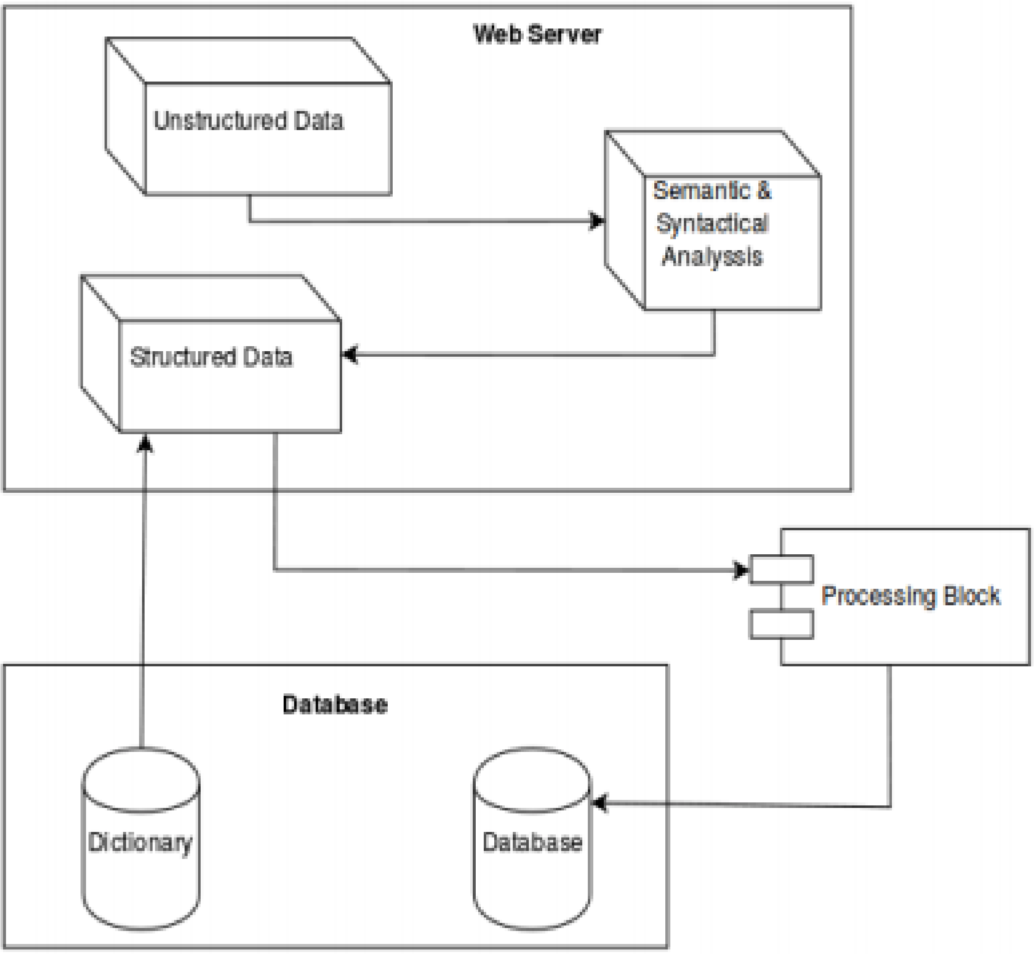


Fig-3

Chapter 8: Data Preparation

Resumes are collected from Workday using (application Programmable Interface).A python script is written to communicate with the Workday API and extract the resume data.
Data are collected from LinkedIn using Selenium (mimics the human behavior)
The resumes which are in different format (.doc or .docx or .pdf) are converted in to common .txt format using Python script.

The Various attributes from the resumes like

1. Experience
2. Technical Skills
3. Non-technical Skills
4. Email
5. Phone Number
6. Education

are extracted from the resume documents and stored in text format.

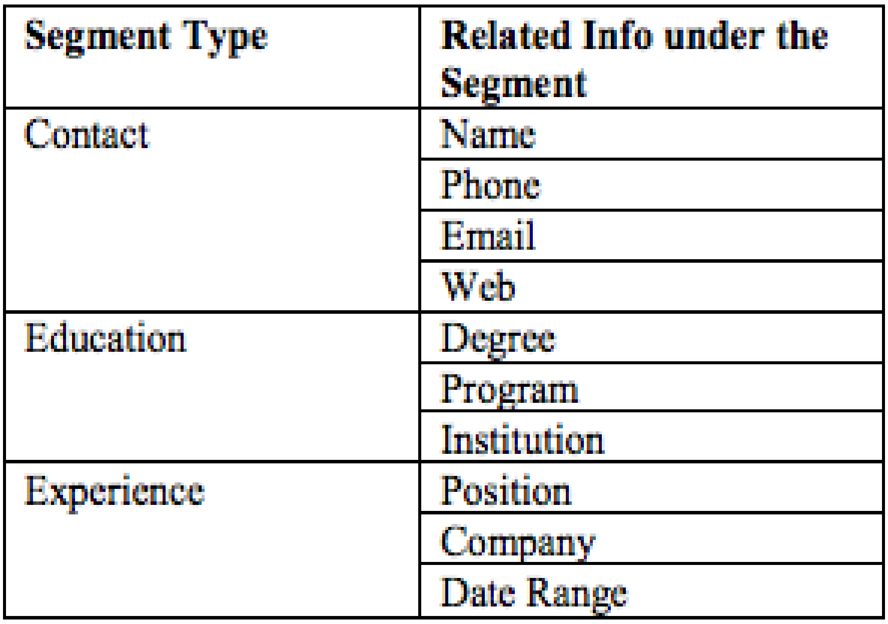
Chapter 9: Data Modeling

In this project, I have extracted the information which consists of four phases like Text Segmentation, NER, Named Entity Clustering, and Text Normalization. (Sonar & Bankar, 2012)

1. **Text Segmentation**

Before Text Segmentation, the resume undergoes the process of tokenization, stemming and lemmatization, stop words removal etc.

Each section and heading in a resume contains a segment of relevant information below it. This feature is used to segment the resumes in to different parts and separate them. Hence the CV will be segregated into segments called Experience, Education, Certifications, technical skills, non-technical skills, and personal information as shown in Figure 1. (Sonar & Bankar, 2012)



Segment Type	Related Info under the Segment
Contact	Name
	Phone
	Email
	Web
Education	Degree
	Program
	Institution
Experience	Position
	Company
	Date Range

Fig-4

2. **Named Entity Recognition (NER)**

The Named Entity Recognizer takes the tokenized text documents as input. The Named Entity refers already defined categories like dates, GPA or scores in education, Location, current company and previous companies, etc. Contact details and chronological information are also considered as Named Entities. (Sonar & Bankar, 2012) The Resume documents mainly consist of these entities and sentences. Due to this reason, an important task is to find out these named entities. Are specially designed for all types of information. Different Information categories are depicted in Table 1. In this phase, each chunker is analysed separately as explained in Figure 2. Chunkers (Sonar & Bankar, 2012) normally use four types of details to find named entities: Clue words (Sonar & Bankar, 2012) are nothing but prepositions. For example, inside the experience section in the resume. The term that follows “worked at” signifies a corporation name.

Well known or Famous names: (Sonar & Bankar, 2012) These represent well know institutions, good companies, renounced academic degrees.

From prefixes and suffixes of the word: (Sonar & Bankar, 2012) This entity denotes the prefixes and suffixes like university of, College etc., for institutions and Corp., Associates, etc., for corporations. (Sonar & Bankar, 2012)

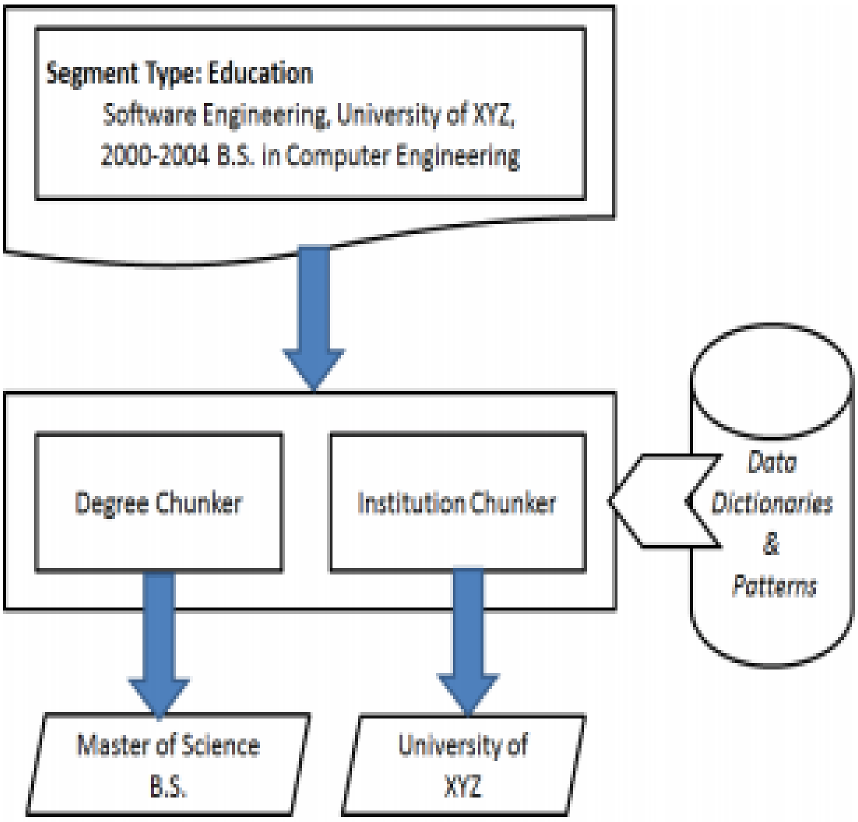

Segment Type: Education
Software Engineering, University of XYZ,
2000-2004 B.S. in Computer Engineering
Degree Chunker
Institution Chunker
Data Dictionaries & Patterns
Master of Science B.S.
University of XYZ

Fig-5

3. Named Entity Clustering

Clustering can be defined as “The process of separating objects into different groups in which the members of each group are quite in some way”. Hence cluster is defined as the objects that are having most similar items in each of them. Each cluster for example education , consists a section of relevant information. For example, Education cluster can show several segments of data about educational institutions which a person had gone to. Education section can contain College name, degree attained and the duration information. Table. 2 shows within the previous step we can have many separate entities. (Sonar & Bankar, 2012) The named entity that is shown within the block of data are supposed to be grouped for further processing and analysis. Table 1 depicts related information defined. Chunks or Named entities are grouped consistently as per their proximity and sort in a cluster. The algorithm depicted in Figure 3 tries to find an association between related entities.

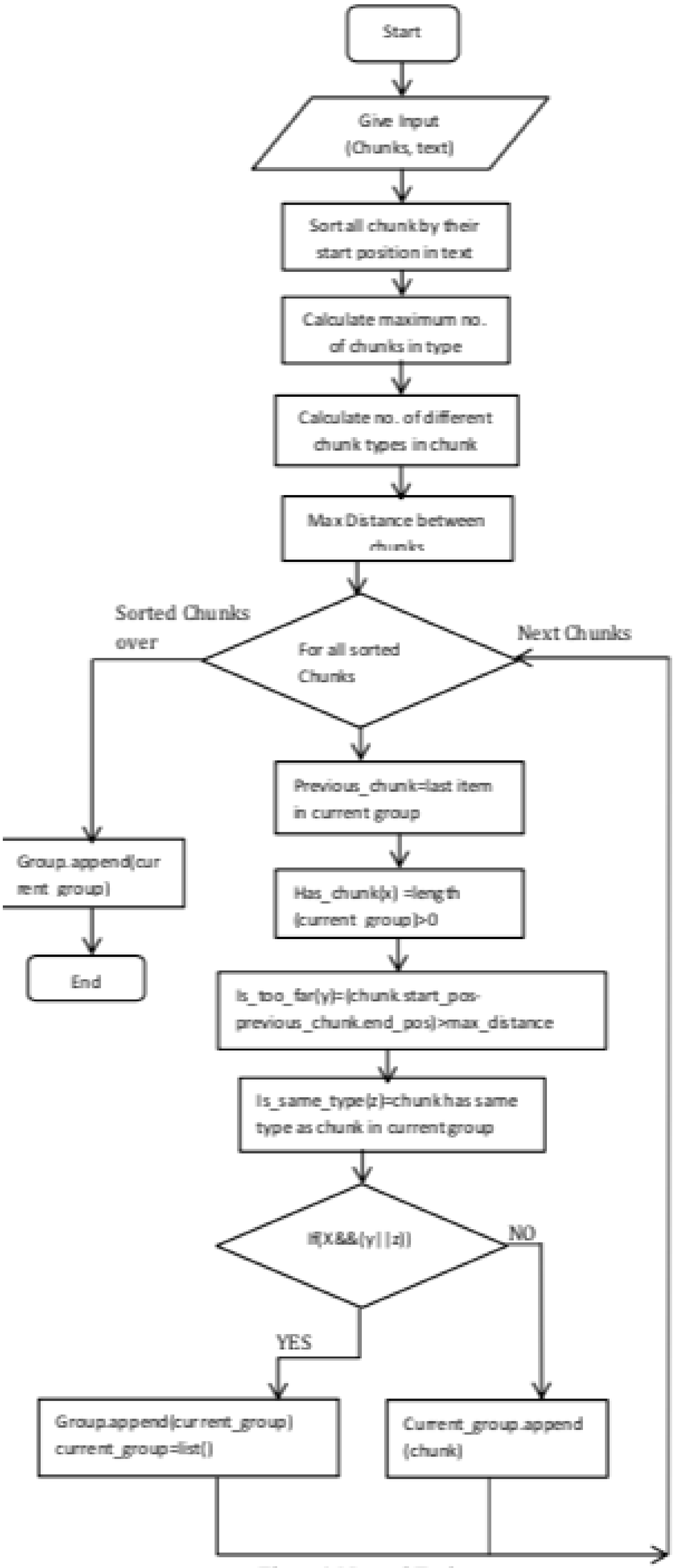
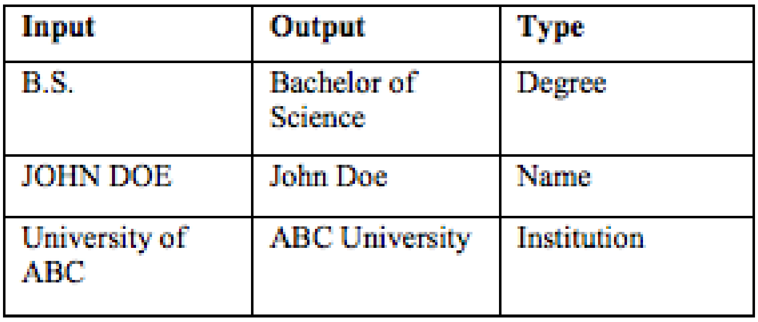


Figure3 Named Entity
Clustering Algorithm

Fig-6

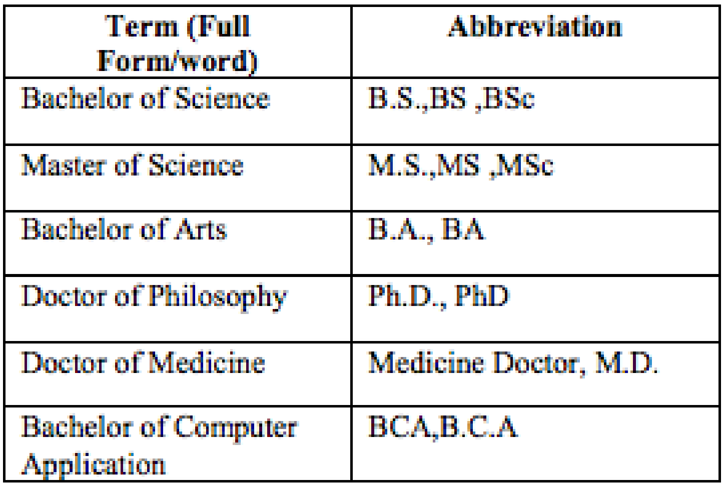
4. Text Normalisation

In this phase, numerous named entities are transformed to make all of them look consistent. Also, all the abbreviations are expanded using dictionaries.



Input	Output	Type
B.S.	Bachelor of Science	Degree
JOHN DOE	John Doe	Name
University of ABC	ABC University	Institution

Table No.1

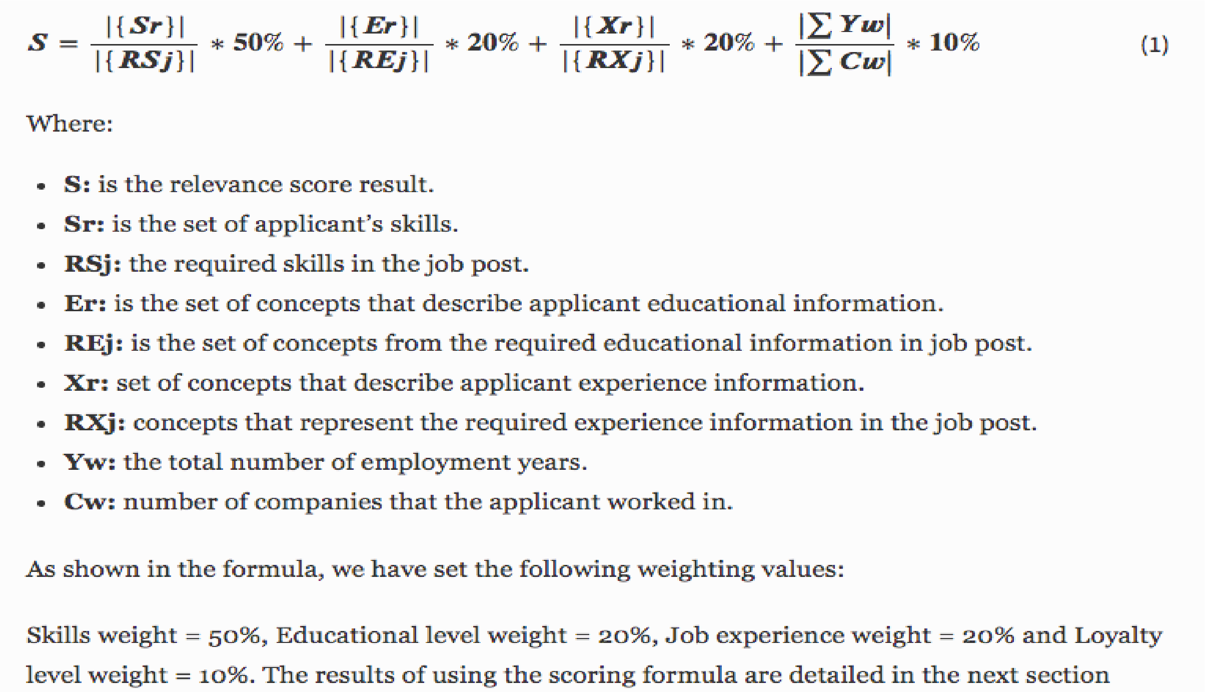


Term (Full Form/word)	Abbreviation
Bachelor of Science	B.S.,BS ,BSc
Master of Science	M.S.,MS ,MSc
Bachelor of Arts	B.A., BA
Doctor of Philosophy	Ph.D., PhD
Doctor of Medicine	Medicine Doctor, M.D.
Bachelor of Computer Application	BCA,B.C.A

Table No.2

Matching Resumes and Job description

In this project to calculate the scoring of the resumes, we are following the method proposed by the authors of (Faerber et al., 2003). To obtain the semantic attributes of cv and job requisition, we are using similar semantic resources and statistical concept-relatedness measures. A new weighting parameter like the loyalty parameter which denotes the degree of loyalty to the corporate that the candidate worked or currently working in, to increase the effectiveness of the matching process. Following scoring formulae is used to calculate the ranking of the resumes.



$$\boldsymbol{S} = \frac{|\{\boldsymbol{Sr}\}|}{|\{\boldsymbol{RSj}\}|} * \mathbf{50\%} + \frac{|\{\boldsymbol{Er}\}|}{|\{\boldsymbol{REj}\}|} * \mathbf{20\%} + \frac{|\{\boldsymbol{Xr}\}|}{|\{\boldsymbol{RXj}\}|} * \mathbf{20\%} + \frac{|\sum \boldsymbol{Yw}|}{|\sum \boldsymbol{Cw}|} * \mathbf{10\%} \tag{1}$$

Where:

- **S:** is the relevance score result.
- **Sr:** is the set of applicant's skills.
- **RSj:** the required skills in the job post.
- **Er:** is the set of concepts that describe applicant educational information.
- **REj:** is the set of concepts from the required educational information in job post.
- **Xr:** set of concepts that describe applicant experience information.
- **RXj:** concepts that represent the required experience information in the job post.
- **Yw:** the total number of employment years.
- **Cw:** number of companies that the applicant worked in.

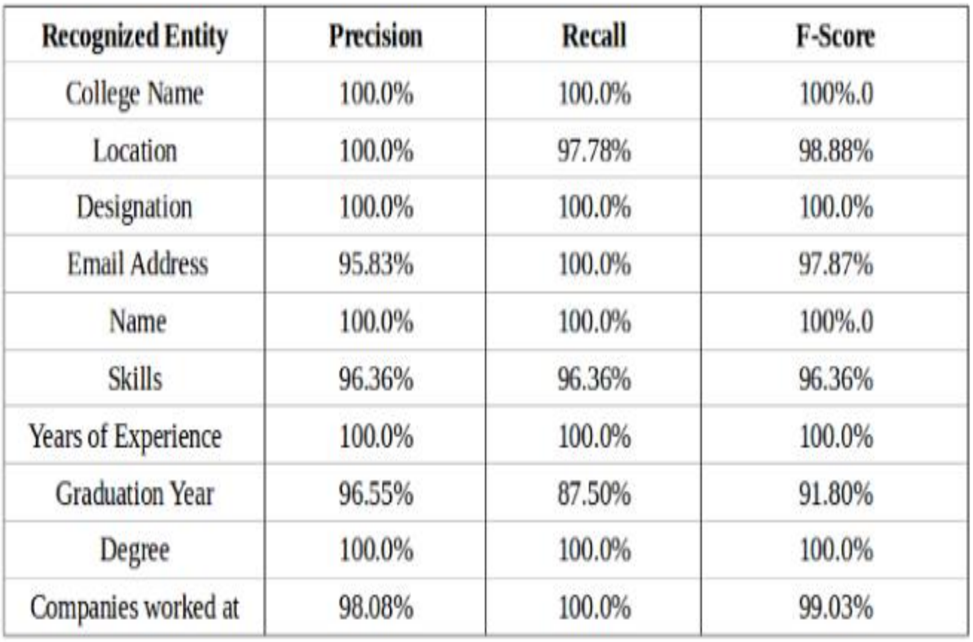
As shown in the formula, we have set the following weighting values:

Skills weight = 50%, Educational level weight = 20%, Job experience weight = 20% and Loyalty level weight = 10%. The results of using the scoring formula are detailed in the next section

Chapter 10: Data Evaluation

Evaluating the Resume Screening system

To evaluate our system, we used precision, recall, and measure. Precision is usually defined as the dividend of the number of related records extracted and the sum of number of related and not related records retrieved. For block segmentation, precision is calculated as the % of correct segments detected and extracted to the whole number of correct and wrong segments that were detected and extracted. For feature extraction, precision is the percentage of correctly extracted feature that matches to the ground truth data. The recall is calculated as the dividend of the no of related records extracted and the whole number of related records present in the database. In block segmentation, recall is the % of correct segments detected and retrieved to the sum number of correct segments present in the database. For feature extraction, recall is the % of correctly retrieved feature to total features present in the training data F1 score is-measured as the HM (Harmonic Mean) of above two metrics (precision and recall).



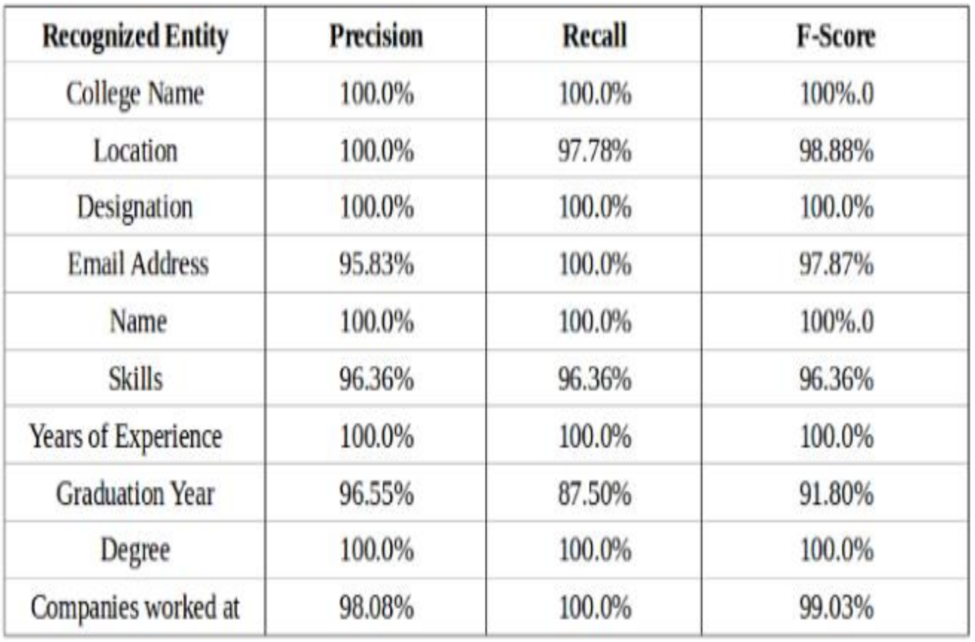
Recognized Entity	Precision	Recall	F-Score
College Name	100.0%	100.0%	100%.0
Location	100.0%	97.78%	98.88%
Designation	100.0%	100.0%	100.0%
Email Address	95.83%	100.0%	97.87%
Name	100.0%	100.0%	100%.0
Skills	96.36%	96.36%	96.36%
Years of Experience	100.0%	100.0%	100.0%
Graduation Year	96.55%	87.50%	91.80%
Degree	100.0%	100.0%	100.0%
Companies worked at	98.08%	100.0%	99.03%

Table No.3

Chapter 11: Analysis and Results

Evaluating the Resume Screening system

To evaluate our system, we used precision, recall, and measure. Precision is usually defined as the dividend of the number of related records extracted and the sum of number of related and not related records retrieved. For block segmentation, precision is calculated as the % of correct segments detected and extracted to the whole number of correct and wrong segments that were detected and extracted. For feature extraction, precision is the percentage of correctly extracted feature that matches to the ground truth data. The recall is calculated as the dividend of the no of related records extracted and the whole number of related records present in the database. In block segmentation, recall is the % of correct segments detected and retrieved to the sum number of correct segments present in the database. For feature extraction, recall is the % of correctly retrieved feature to total features present in the training data F1 score is-measured as the HM(harmonic mean) of above two metrics(precision and recall).



Recognized Entity	Precision	Recall	F-Score
College Name	100.0%	100.0%	100%.0
Location	100.0%	97.78%	98.88%
Designation	100.0%	100.0%	100.0%
Email Address	95.83%	100.0%	97.87%
Name	100.0%	100.0%	100%.0
Skills	96.36%	96.36%	96.36%
Years of Experience	100.0%	100.0%	100.0%
Graduation Year	96.55%	87.50%	91.80%
Degree	100.0%	100.0%	100.0%
Companies worked at	98.08%	100.0%	99.03%

Table No.4

Chapter 12: Conclusions and Recommendations for future work

In this project, we came with a web application that helps to shortlist the resumes of the applicants for a job posting. An important technical aspect that creates the application possible is that the pulling out various segments of data from resumes. (Singh et al., 2010) Since process necessities frequently specify necessities such as “a minimal of 3 years of Python Programmer” we have used records parsed from resumes to supply filters that permit recruiters to shortlist the applicants. This enables us to beat the restrictions inherent in natural keyword primarily based totally matching. In this project, the recruiters have experienced that this application aided in speeding up the hiring process. This is mainly accounted by two attributes. The first one is scoring the applicants to suit the job profile and use of the filters provided based on various data segments retrieved from the cv helps the recruiters to inspect far fewer resumes to shortlist a given number of candidates. Secondly, showing segments of the resume that match mostly to the profile and based on the content retrieved from the resume allows a recruiter to select or discard applicants quicker than manually scanning the entire resume. (Amin et al., 2019). As future work, Deep Learning can be to create a version for screening and rating the resumes.

Bibliography

Amin, S., Jayakar, N., Sunny, S., Babu, P., Kiruthika, M., & Gurjar, A. (2019). Web Application for Screening Resume. 2019 International Conference on Nascent Technologies in Engineering, ICNTE 2019 - Proceedings, Icnte, 1–7. https://doi.org/10.1109/ICNTE44896.2019.8945869

Faerber, F., Weitzel, T., Keim, T., & Färber, F. (2003). Association for Information Systems AIS Electronic Library (AISeL) An Automated Recommendation Approach to Selection in Personnel Recruitment AN AUTOMATED RECOMMENDATION APPROACH TO SELECTION IN PERSONNEL RECRUITMENT. http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1768&context=amcis2003%0Ahttp://aisel.aisnet.org/amcis2003/302

Laumer, S., & Eckhardt, A. (2009). Help to find the needle in a haystack - Integrating recommender systems in an IT supported staff recruitment system. SIGMIS CPR'09 - Proceedings of the 2009 ACM SIGMIS Computer Personnel Research Conference. https://doi.org/10.1145/1542130.1542133

Lee, I. (2007). An architecture for a next-generation holistic e-recruiting system. Communications of the ACM, 50(7), 81–85. https://doi.org/10.1145/1272516.1272518

Palshikar, G. K., Srivastava, R., Shah, M., & Pawar, S. (2018). Automatic shortlisting of candidates in recruitment. CEUR Workshop Proceedings, 2127(July), 5–11.

Singh, A., Catherine, R., Visweswariah, K., Chenthamarakshan, V., & Kambhatla, N. (2010). PROSPECT: A system for screening candidates for recruitment. International Conference on Information and Knowledge Management, Proceedings, October 2017, 659–668. https://doi.org/10.1145/1871437.1871523

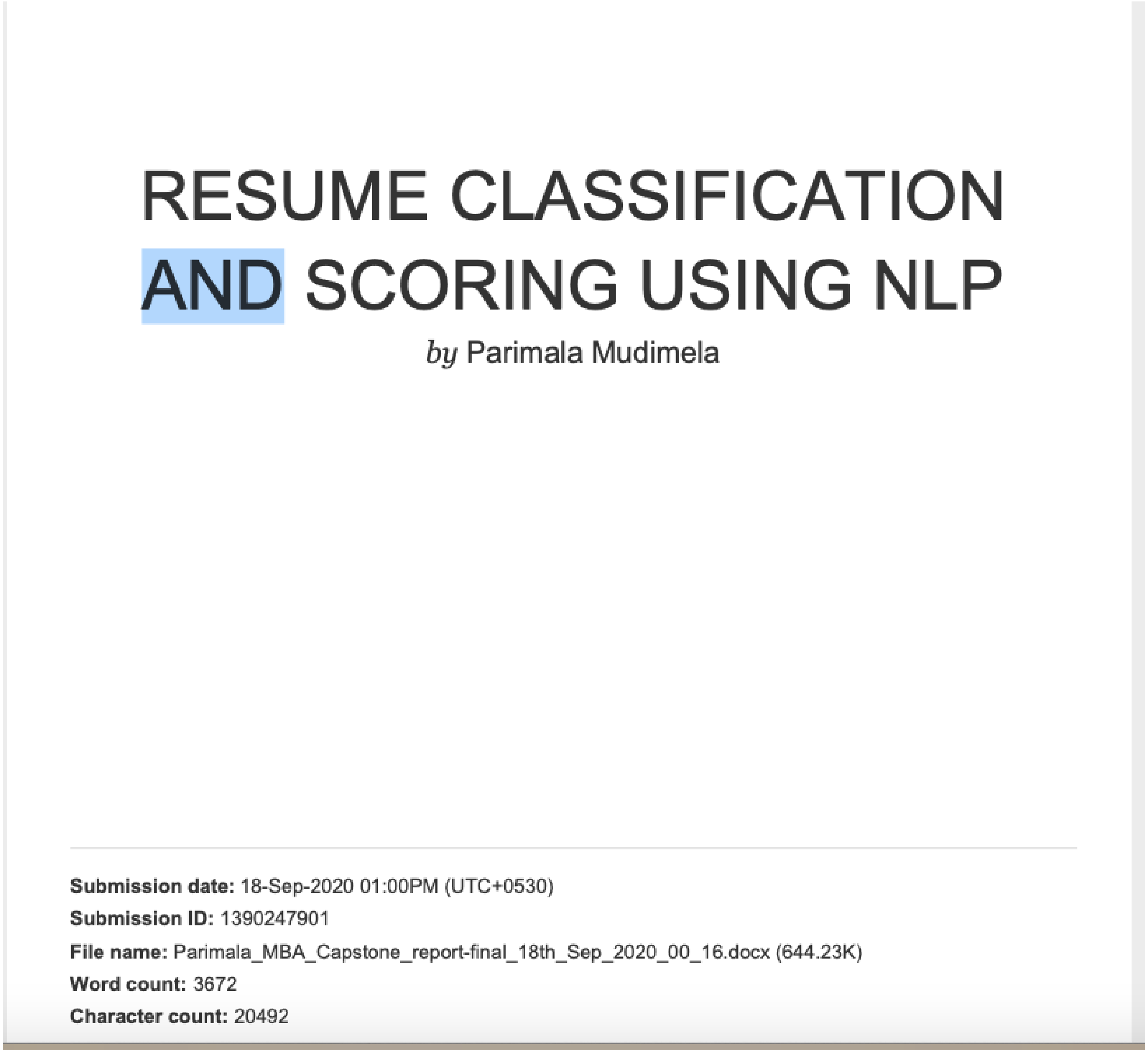
Sonar, S., & Bankar, B. (2012). Resume Parsing with Named Entity Clustering Algorithm. Paper, SVPM College of Engineering Baramati, Maharashtra, India. https://www.slideshare.net/swapnilmsonar/resume-parsing-with-named-entity-clustering-algorithm

Using effective recruitment to gain competitive advantage. (n.d.). https://greenbeanrpo.com/blog/using-effective-recruitment-gain-competitive-advantage/#:~:text=Shorter recruitment cycles make it,them tempted to look elsewhere.

Yi, X., Allan, J., & Croft, W. B. (2007). Matching resumes and jobs based on relevance models. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR’07, 809–810. https://doi.org/10.1145/1277741.1277920

Appendix

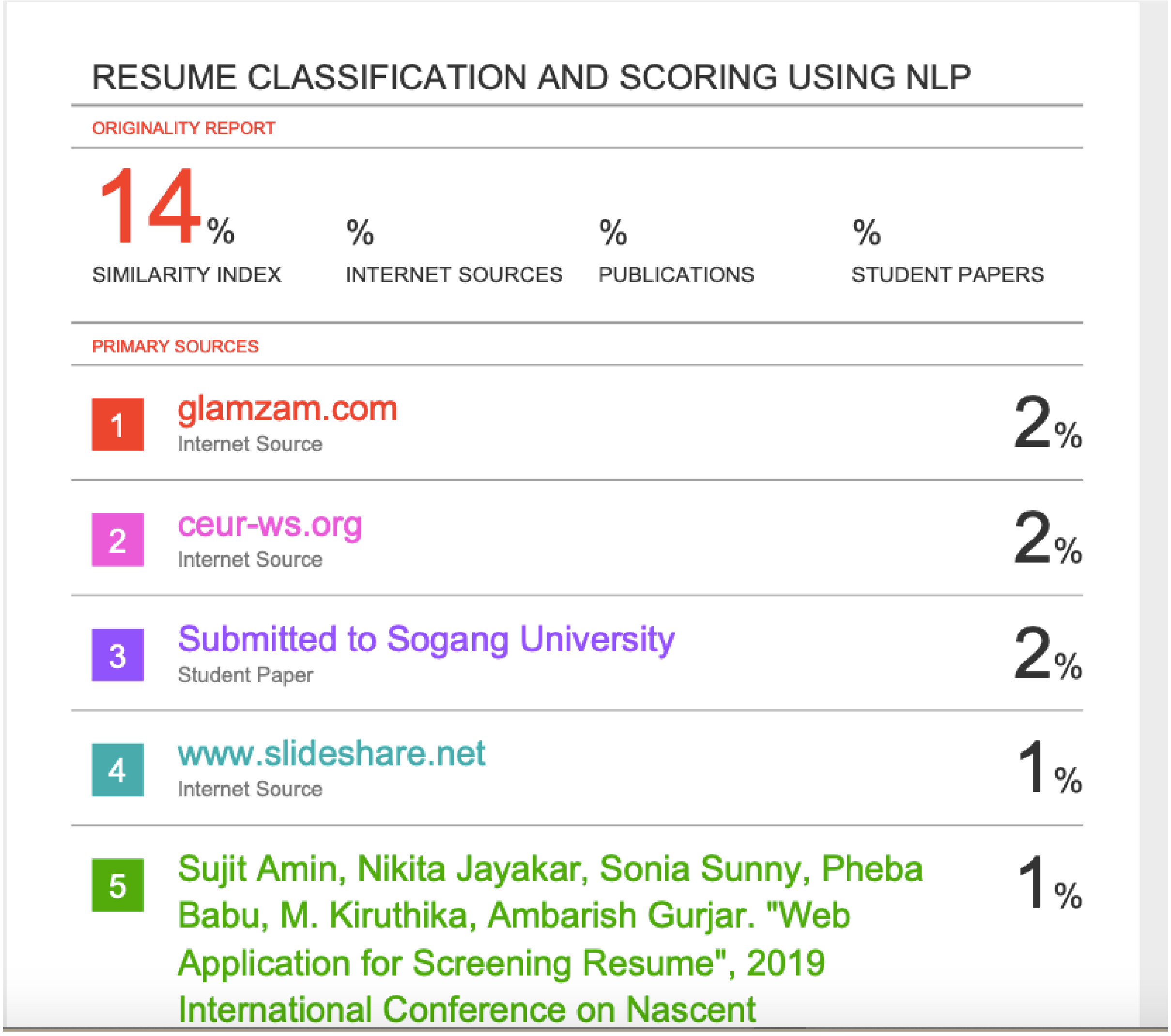
Plagiarism Report



RESUME CLASSIFICATION AND SCORING USING NLP

by Parimala Mudimela

Submission date: 18-Sep-2020 01:00PM (UTC+0530)
Submission ID: 1390247901
File name: Parimala_MBA_Capstone_report-final_18th_Sep_2020_00_16.docx (644.23K)
Word count: 3672
Character count: 20492



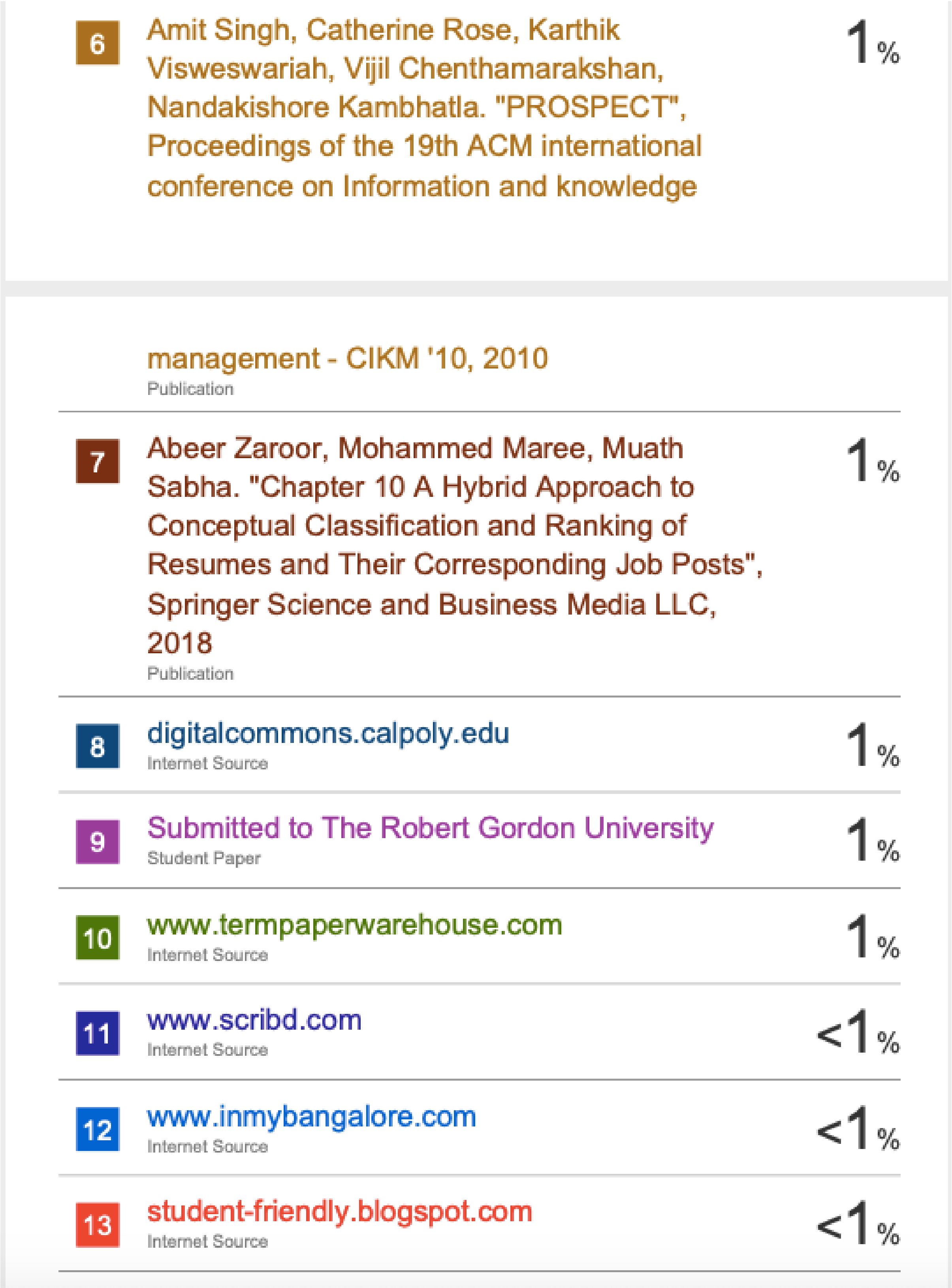
RESUME CLASSIFICATION AND SCORING USING NLP

ORIGINALITY REPORT

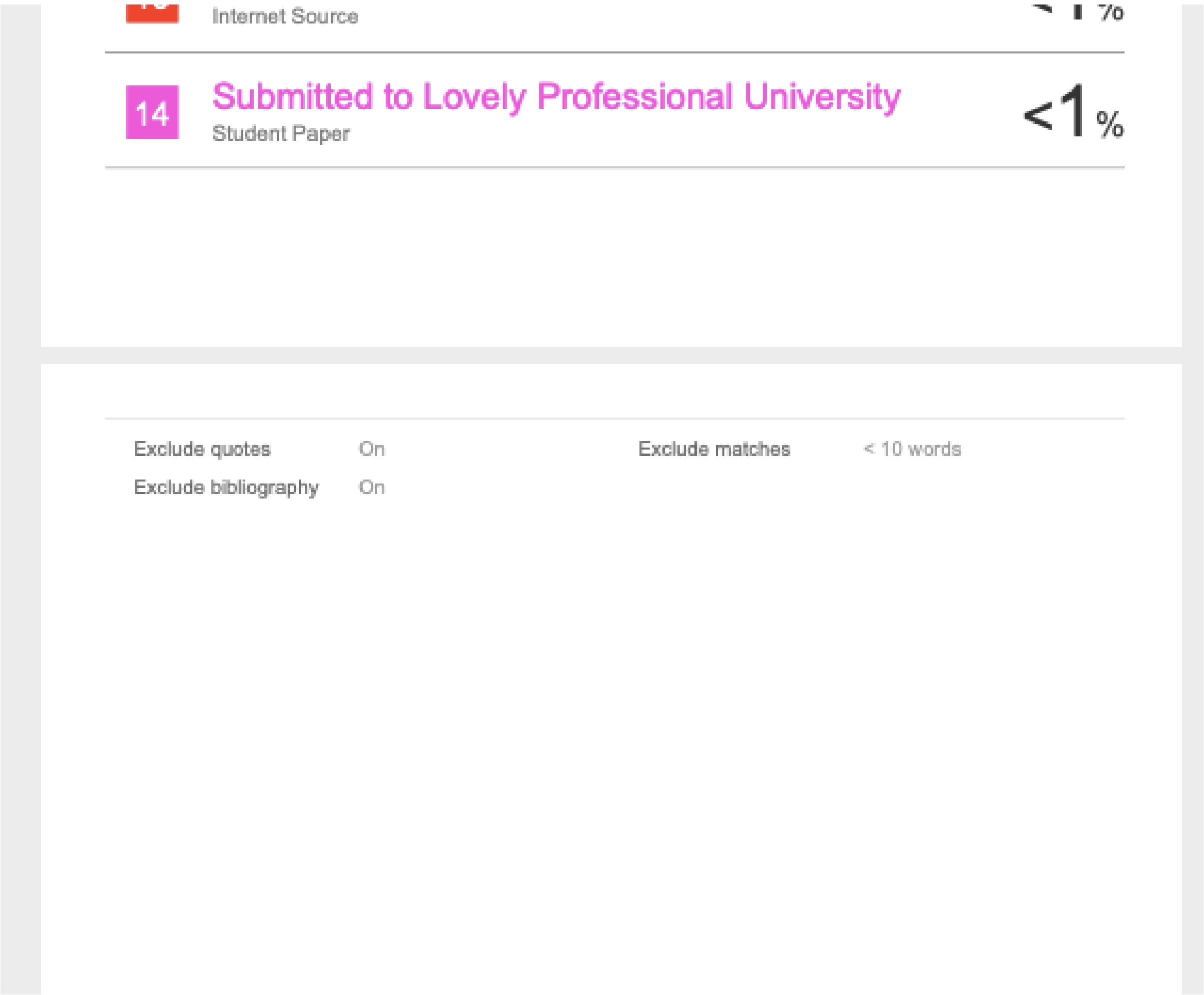
14%	%	%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	glamzam.com Internet Source	2%
2	ceur-ws.org Internet Source	2%
3	Submitted to Sogang University Student Paper	2%
4	www.slideshare.net Internet Source	1%
5	Sujit Amin, Nikita Jayakar, Sonia Sunny, Pheba Babu, M. Kiruthika, Ambarish Gurjar. "Web Application for Screening Resume", 2019 International Conference on Nascent	1%



6	Amit Singh, Catherine Rose, Karthik Visweswariah, Vijil Chenthamarakshan, Nandakishore Kambhatla. "PROSPECT", Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10, 2010 Publication	1 %
7	Abeer Zaroor, Mohammed Maree, Muath Sabha. "Chapter 10 A Hybrid Approach to Conceptual Classification and Ranking of Resumes and Their Corresponding Job Posts", Springer Science and Business Media LLC, 2018 Publication	1 %
8	digitalcommons.calpoly.edu Internet Source	1 %
9	Submitted to The Robert Gordon University Student Paper	1 %
10	www.termpaperwarehouse.com Internet Source	1 %
11	www.scribd.com Internet Source	<1 %
12	www.inmybangalore.com Internet Source	<1 %
13	student-friendly.blogspot.com Internet Source	<1 %



Internet Source

14	Submitted to Lovely Professional University Student Paper	<1%

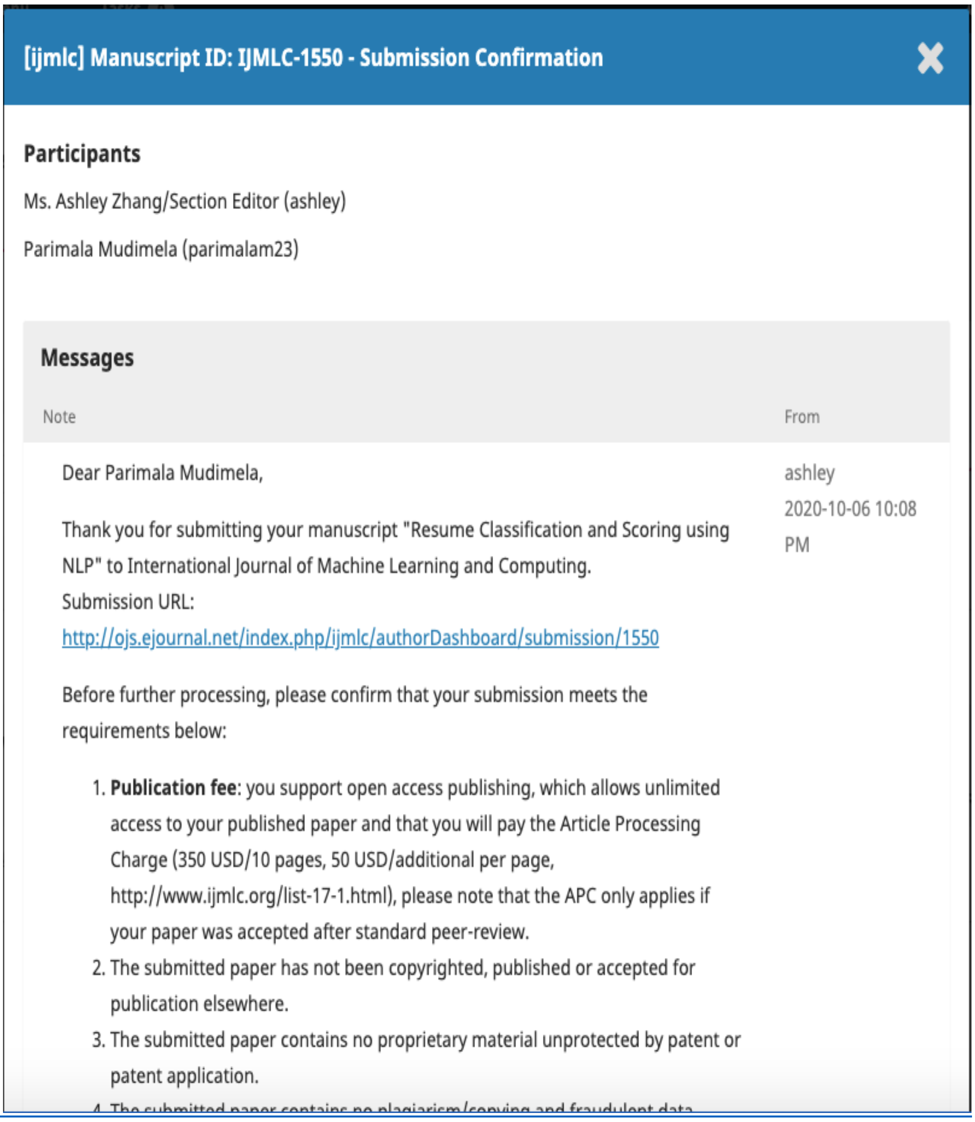
Exclude quotes On

Exclude bibliography On

Exclude matches < 10 words

Publications in a Journal/Conference Presented/White Paper

http://ojs.ejournal.net/index.php/ijmlc/authorDashboard/submission/1550



[ijmlc] Manuscript ID: IJMLC-1550 - Submission Confirmation

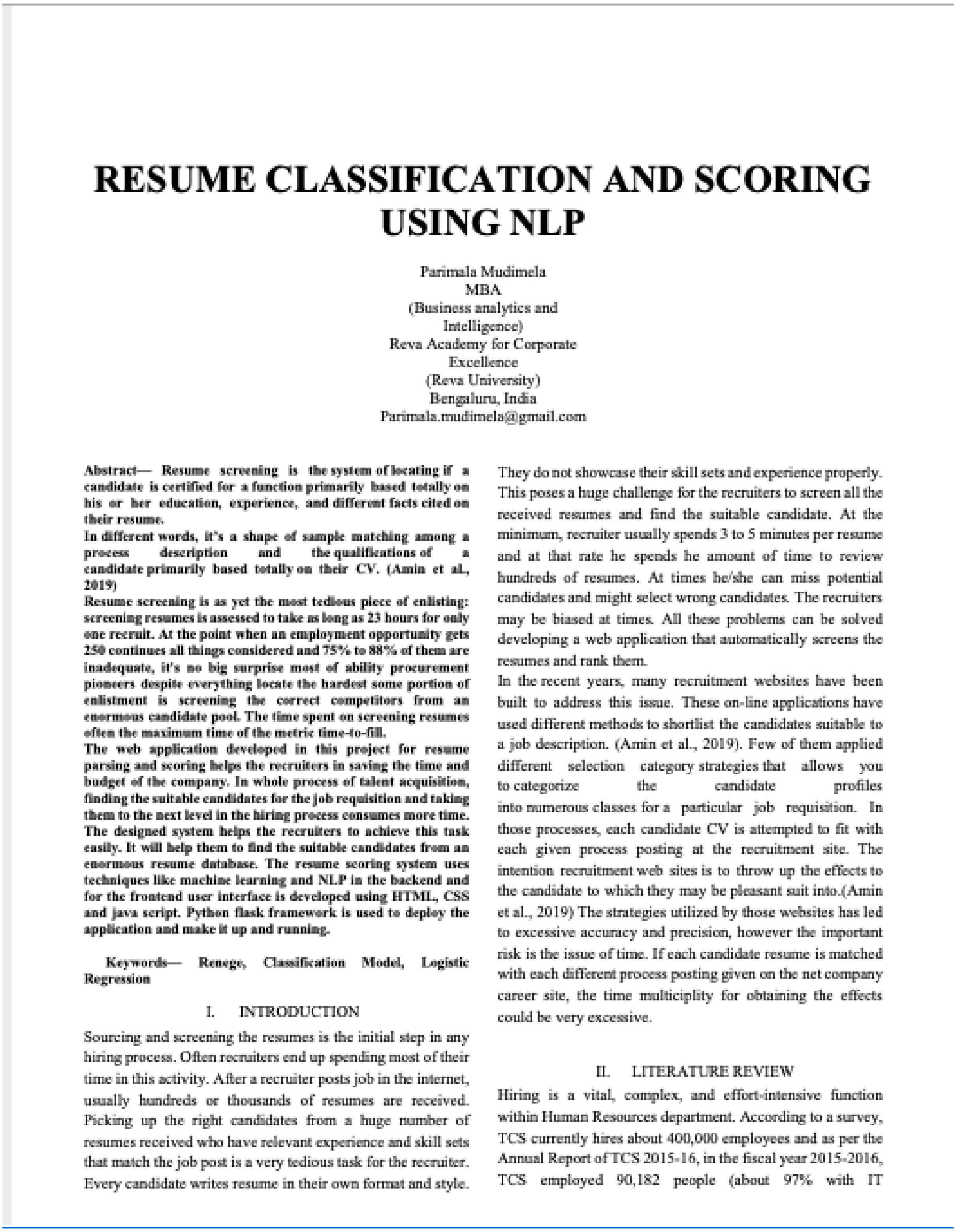
Participants

Ms. Ashley Zhang/Section Editor (ashley)

Parimala Mudimela (parimalam23)

Messages

Note	From
Dear Parimala Mudimela, Thank you for submitting your manuscript "Resume Classification and Scoring using NLP" to International Journal of Machine Learning and Computing. Submission URL: http://ojs.ejournal.net/index.php/ijmlc/authorDashboard/submission/1550 Before further processing, please confirm that your submission meets the requirements below: 1. **Publication fee**: you support open access publishing, which allows unlimited access to your published paper and that you will pay the Article Processing Charge (350 USD/10 pages, 50 USD/additional per page, http://www.ijmlc.org/list-17-1.html), please note that the APC only applies if your paper was accepted after standard peer-review. 2. The submitted paper has not been copyrighted, published or accepted for publication elsewhere. 3. The submitted paper contains no proprietary material unprotected by patent or patent application. 4. The submitted paper contains no plagiarism/copying and fraudulent data	ashley 2020-10-06 10:08 PM



RESUME CLASSIFICATION AND SCORING USING NLP

Parimala Mudimela
MBA
(Business analytics and
Intelligence)
Reva Academy for Corporate
Excellence
(Reva University)
Bengaluru, India
Parimala.mudimela@gmail.com

Abstract— Resume screening is the system of locating if a candidate is certified for a function primarily based totally on his or her education, experience, and different facts cited on their resume.

In different words, it's a shape of sample matching among a process description and the qualifications of a candidate primarily based totally on their CV. (Amin et al., 2019)

Resume screening is as yet the most tedious piece of enlisting: screening resumes is assessed to take as long as 23 hours for only one recruit. At the point when an employment opportunity gets 250 continues all things considered and 75% to 88% of them are inadequate, it's no big surprise most of ability procurement pioneers despite everything locate the hardest some portion of enlistment is screening the correct competitors from an enormous candidate pool. The time spent on screening resumes often the maximum time of the metric time-to-fill.

The web application developed in this project for resume parsing and scoring helps the recruiters in saving the time and budget of the company. In whole process of talent acquisition, finding the suitable candidates for the job requisition and taking them to the next level in the hiring process consumes more time. The designed system helps the recruiters to achieve this task easily. It will help them to find the suitable candidates from an enormous resume database. The resume scoring system uses techniques like machine learning and NLP in the backend and for the frontend user interface is developed using HTML, CSS and java script. Python flask framework is used to deploy the application and make it up and running.

Keywords— Renege, Classification Model, Logistic Regression

I. INTRODUCTION

Sourcing and screening the resumes is the initial step in any hiring process. Often recruiters end up spending most of their time in this activity. After a recruiter posts job in the internet, usually hundreds or thousands of resumes are received. Picking up the right candidates from a huge number of resumes received who have relevant experience and skill sets that match the job post is a very tedious task for the recruiter. Every candidate writes resume in their own format and style. They do not showcase their skill sets and experience properly. This poses a huge challenge for the recruiters to screen all the received resumes and find the suitable candidate. At the minimum, recruiter usually spends 3 to 5 minutes per resume and at that rate he spends he amount of time to review hundreds of resumes. At times he/she can miss potential candidates and might select wrong candidates. The recruiters may be biased at times. All these problems can be solved developing a web application that automatically screens the resumes and rank them.

In the recent years, many recruitment websites have been built to address this issue. These on-line applications have used different methods to shortlist the candidates suitable to a job description. (Amin et al., 2019). Few of them applied different selection category strategies that allows you to categorize the candidate profiles into numerous classes for a particular job requisition. In those processes, each candidate CV is attempted to fit with each given process posting at the recruitment site. The intention recruitment web sites is to throw up the effects to the candidate to which they may be pleasant suit into.(Amin et al., 2019) The strategies utilized by those websites has led to excessive accuracy and precision, however the important risk is the issue of time. If each candidate resume is matched with each different process posting given on the net company career site, the time multiciplity for obtaining the effects could be very excessive.

II. LITERATURE REVIEW

Hiring is a vital, complex, and effort-intensive function within Human Resources department. According to a survey, TCS currently hires about 400,000 employees and as per the Annual Report of TCS 2015-16, in the fiscal year 2015-2016, TCS employed 90,182 people (about 97% with IT background), 74,009 in India and 16,173 abroad. Taking as an assumption of 20% selection rate, on an average 5 people were interviewed. For shortlisting 5 persons per post, HR executives have to screen manually at least 10 resumes per requirement. Taking in to account an average effort of 350 seconds per profile, this interprets 70,000 man hours annually., spent solely for resume screening and shortlisting the prospective candidates for the interview. After going through so many a much tedious task, the manually evaluating the resumes is often error-prone, opaque, biased, does not facilitate comparisons, and opportunities for improving the quality of recruitment are difficult.(Palshikar et al., 2018)

Multiple trials were performed to automate several operations of talent acquisition. (Lee, 2007) For example, (Laumer & Eckhardt, 2009) recommends methods such as "collaborative filtering"(Singh et al., 2010) to shortlist the resumes suitable for a job a requirement. (Yi et al., 2007) defines a technique which uses "relevance models" to match to match the text content to match the resumes with the job requisition. After that, candidate profiles that match with the job postings are used to capture semantics that is not clearly cited in the job profile requirements. The approaches are based on assumption that the resumes are having manually labeled rankings.
In (Faerber et al., 2003) Collaborative Filtering and content-based similarity measures are combined for more accurate results in ranking. All the past research have been done on artificially generated data and did not use unstructured resume documents and job requisition document.(Singh et al., 2010) (Amin et al., 2019)

III. METHODS

The method proposed in this work offers an approach for screening the resumes based on a job description. This is a web application to help recruiters by screening the CVs, shortlisting candidates that best match the position, and filtering out those who don't.
Resumes are ranked based on the score in comparison with the job description.

The Proposed system consists of the following components:

1. Resume Collection
2. Resume Parser system
3. Candidate skill set database system
4. Ranking algorithm

Resumes are collected from Workday using API (Application Programmable Interface). A python script is written to communicate with the Workday API and extract the resume data.

Data are collected from LinkedIn using Selenium (mimics human behavior).
Web application accepts job description as the input and based on the selection of the job portal, the resumes are collected by the back-end code.

The next step is the resume parsing. The CVs will be of different formats like word.doc, Docx, .pdf, etc. We convert all these CVs into text format which will be easy to analyze.

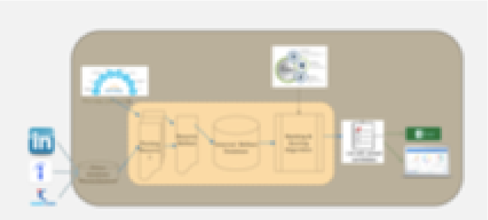


Fig 1

A. Resume Parser system

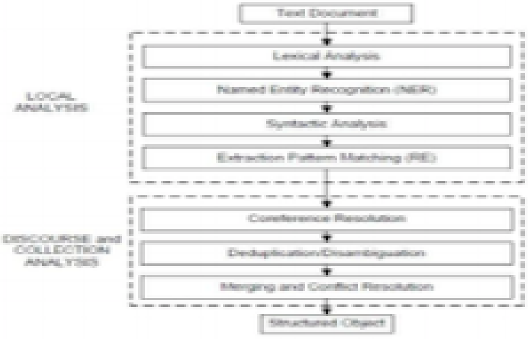
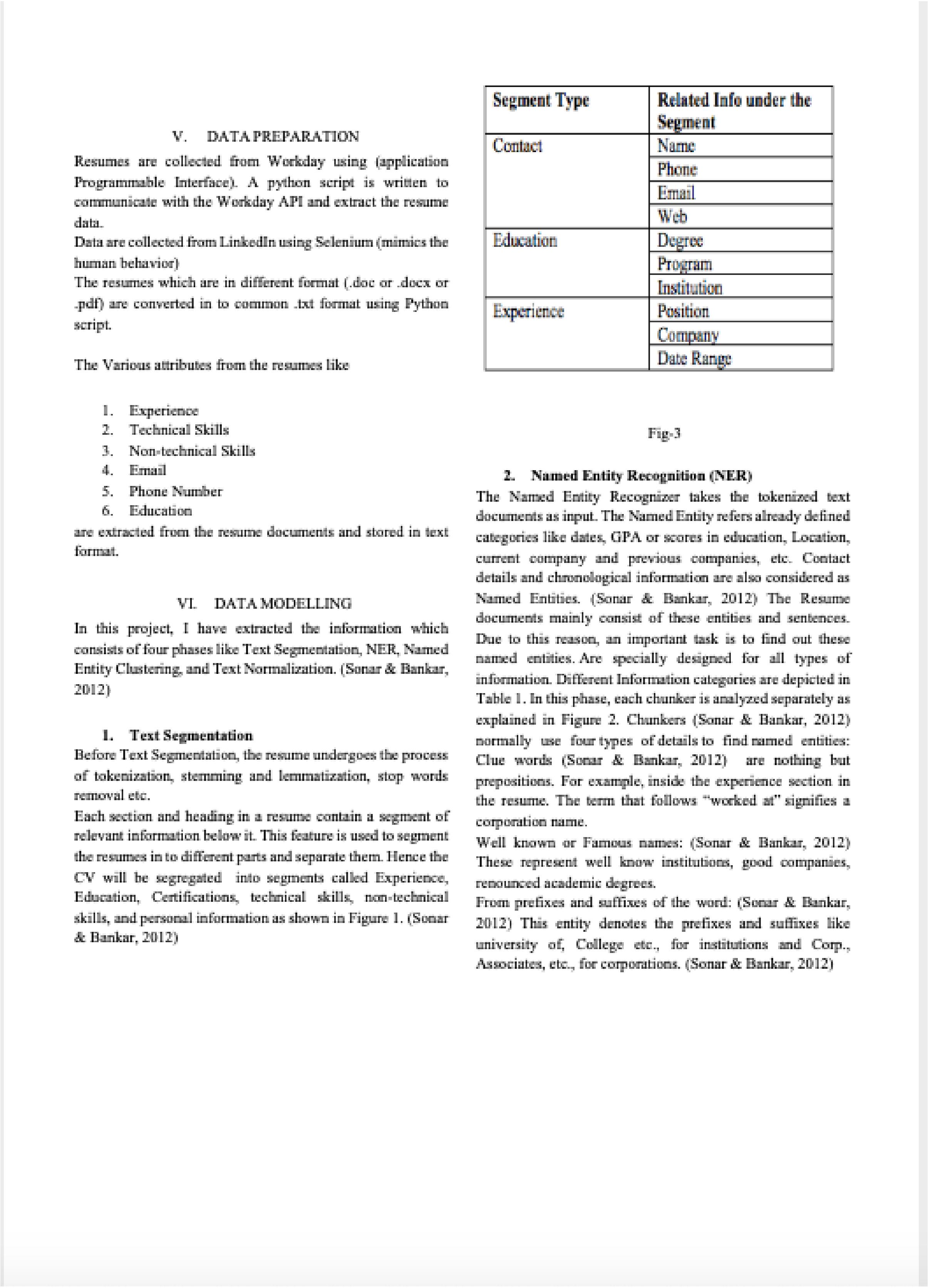

Text Document
LOCAL ANALYSIS
Lexical Analysis
Named Entity Recognition (NER)
Syntactic Analysis
Extraction Pattern Matching (RE)
DISCOURSE and COLLECTION ANALYSIS
Conference Resolution
Deduplication/Disambiguation
Merging and Conflict Resolution
Structured Object

Fig 2

The resumes are ranked by assigning weightage to different attributes in the resume like skills, Experience, Education, Projects worked, etc, and compared with the job description. Scores are assigned based on the match between resume data components and Job description data components.

IV. DATASET

Data comprises of unstructured resumes which are in the form of .doc or .docx or pdf format. Resumes are collected from various job portals like Workday, LinkedIn, Naukri, etc. Job Description which is in the format of .doc or .docx or .pdf. Unstructured resumes are converted into structured text format. Have used manually annotated resumes from recruiters as training data set.



V. DATA PREPARATION

Resumes are collected from Workday using (application Programmable Interface). A python script is written to communicate with the Workday API and extract the resume data.

Data are collected from LinkedIn using Selenium (mimics the human behavior)

The resumes which are in different format (.doc or .docx or .pdf) are converted in to common .txt format using Python script.

The Various attributes from the resumes like

1. Experience
2. Technical Skills
3. Non-technical Skills
4. Email
5. Phone Number
6. Education

are extracted from the resume documents and stored in text format.

VI. DATA MODELLING

In this project, I have extracted the information which consists of four phases like Text Segmentation, NER, Named Entity Clustering, and Text Normalization. (Sonar & Bankar, 2012)

1. Text Segmentation

Before Text Segmentation, the resume undergoes the process of tokenization, stemming and lemmatization, stop words removal etc.

Each section and heading in a resume contain a segment of relevant information below it. This feature is used to segment the resumes in to different parts and separate them. Hence the CV will be segregated into segments called Experience, Education, Certifications, technical skills, non-technical skills, and personal information as shown in Figure 1. (Sonar & Bankar, 2012)

Segment Type	Related Info under the Segment
Contact	Name
	Phone
	Email
	Web
Education	Degree
	Program
	Institution
Experience	Position
	Company
	Date Range

Fig-3

2. Named Entity Recognition (NER)

The Named Entity Recognizer takes the tokenized text documents as input. The Named Entity refers already defined categories like dates, GPA or scores in education, Location, current company and previous companies, etc. Contact details and chronological information are also considered as Named Entities. (Sonar & Bankar, 2012) The Resume documents mainly consist of these entities and sentences. Due to this reason, an important task is to find out these named entities. Are specially designed for all types of information. Different Information categories are depicted in Table 1. In this phase, each chunker is analyzed separately as explained in Figure 2. Chunkers (Sonar & Bankar, 2012) normally use four types of details to find named entities: Clue words (Sonar & Bankar, 2012) are nothing but prepositions. For example, inside the experience section in the resume. The term that follows "worked at" signifies a corporation name.

Well known or Famous names: (Sonar & Bankar, 2012) These represent well know institutions, good companies, renounced academic degrees.

From prefixes and suffixes of the word: (Sonar & Bankar, 2012) This entity denotes the prefixes and suffixes like university of, College etc., for institutions and Corp., Associates, etc., for corporations. (Sonar & Bankar, 2012)

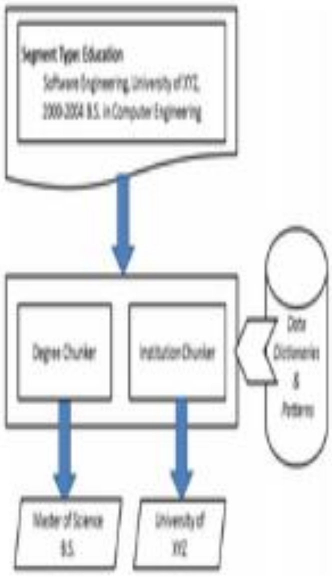


Fig 4

3. Named Entity Clustering

Clustering can be defined as "The process of separating objects into different groups in which the members of each group are quite in some way". Hence cluster is defined as the objects that are having most similar items in each of them. Each cluster for example education, consists a section of relevant information. For example, Education cluster can show several segments of data about educational institutions which a person had gone to. Education section can contain College name, degree attained and the duration information. Table. 2 shows within the previous step we can have many separate entities. (Sonar & Bankar, 2012) The named entity that is shown within the block of data are supposed to be grouped for further processing and analysis.

Table 1 depicts related information defined. Chunks or Named entities are grouped consistently as per their proximity and sort in a cluster. The algorithm depicted in Figure 3 tries to find an association between related entities.

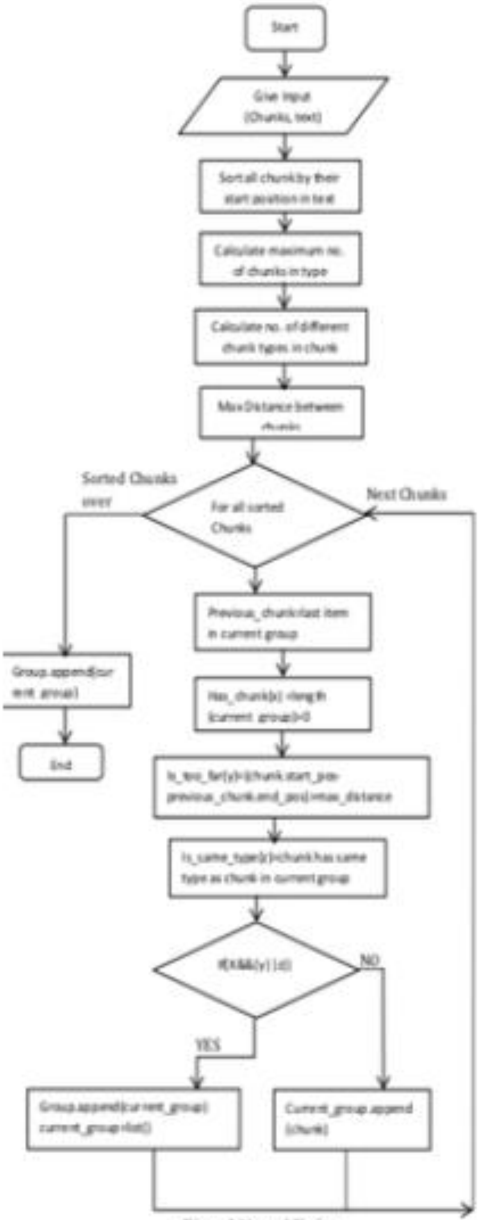
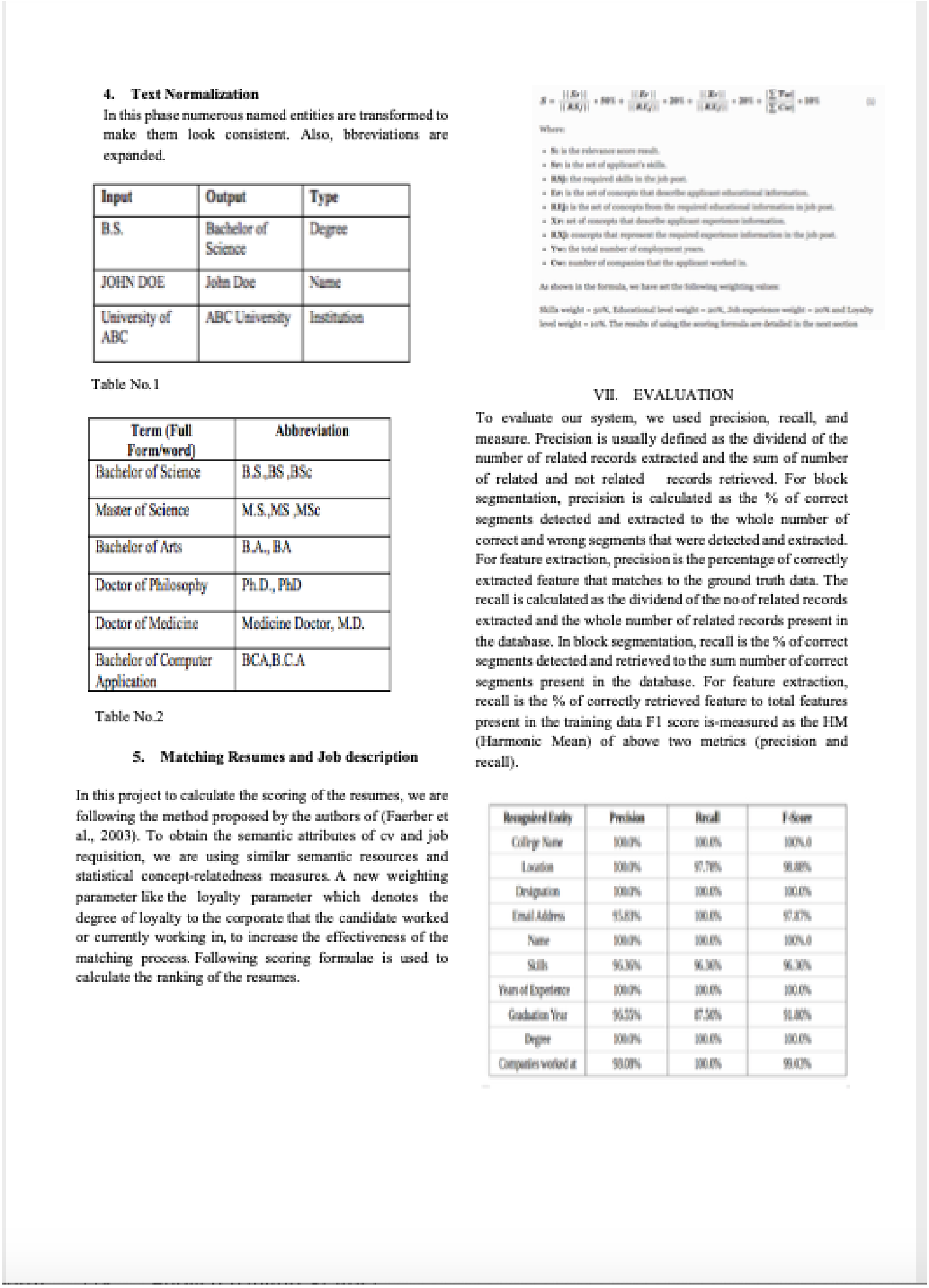

Start
Give Input
(Chunks, text)
Sort all chunk by their start position in text
Calculate maximum no. of chunks in type
Calculate no. of different chunk types in chunk
Max Distance between chunks
Sorted Chunks over
For all sorted Chunks
Next Chunks
Previous_chunk=last item in current group
Has_chunk(x) =length (current group)>0
Is_too_far(y)=(chunk.start_pos-previous_chunk.end_pos)>max_distance
Is_same_type(z)=chunk has same type as chunk in current group
If(X&&(y||z))
NO
YES
Group.append(current_group)
current_group=list()
Current_group.append (chunk)
Group.append(current_group)
End

Figure3 Named Entity Clustering Algorithm

Fig 5



4. Text Normalization

In this phase numerous named entities are transformed to make them look consistent. Also, bbreviations are expanded.

Input	Output	Type
B.S.	Bachelor of Science	Degree
JOHN DOE	John Doe	Name
University of ABC	ABC University	Institution

Table No.1

Term (Full Form/word)	Abbreviation
Bachelor of Science	B.S.,BS ,BSc
Master of Science	M.S.,MS ,MSc
Bachelor of Arts	B.A., BA
Doctor of Philosophy	Ph.D., PhD
Doctor of Medicine	Medicine Doctor, M.D.
Bachelor of Computer Application	BCA,B.C.A

Table No.2

5. Matching Resumes and Job description

In this project to calculate the scoring of the resumes, we are following the method proposed by the authors of (Faerber et al., 2003). To obtain the semantic attributes of cv and job requisition, we are using similar semantic resources and statistical concept-relatedness measures. A new weighting parameter like the loyalty parameter which denotes the degree of loyalty to the corporate that the candidate worked or currently working in, to increase the effectiveness of the matching process. Following scoring formulae is used to calculate the ranking of the resumes.

$$S = \frac{||Sr||}{||RSj||} * 50\% + \frac{||Er||}{||REj||} * 20\% + \frac{||Xr||}{||RXj||} * 20\% + \frac{\sum Twi}{\sum Cwi} * 10\% \quad (1)$$

Where:

- **S:** is the relevance score result.
- **Sr:** is the set of applicant's skills.
- **RSj:** the required skills in the job post.
- **Er:** is the set of concepts that describe applicant educational information.
- **REj:** is the set of concepts from the required educational information in job post.
- **Xr:** set of concepts that describe applicant experience information.
- **RXj:** concepts that represent the required experience information in the job post.
- **Tw:** the total number of employment years.
- **Cw:** number of companies that the applicant worked in.

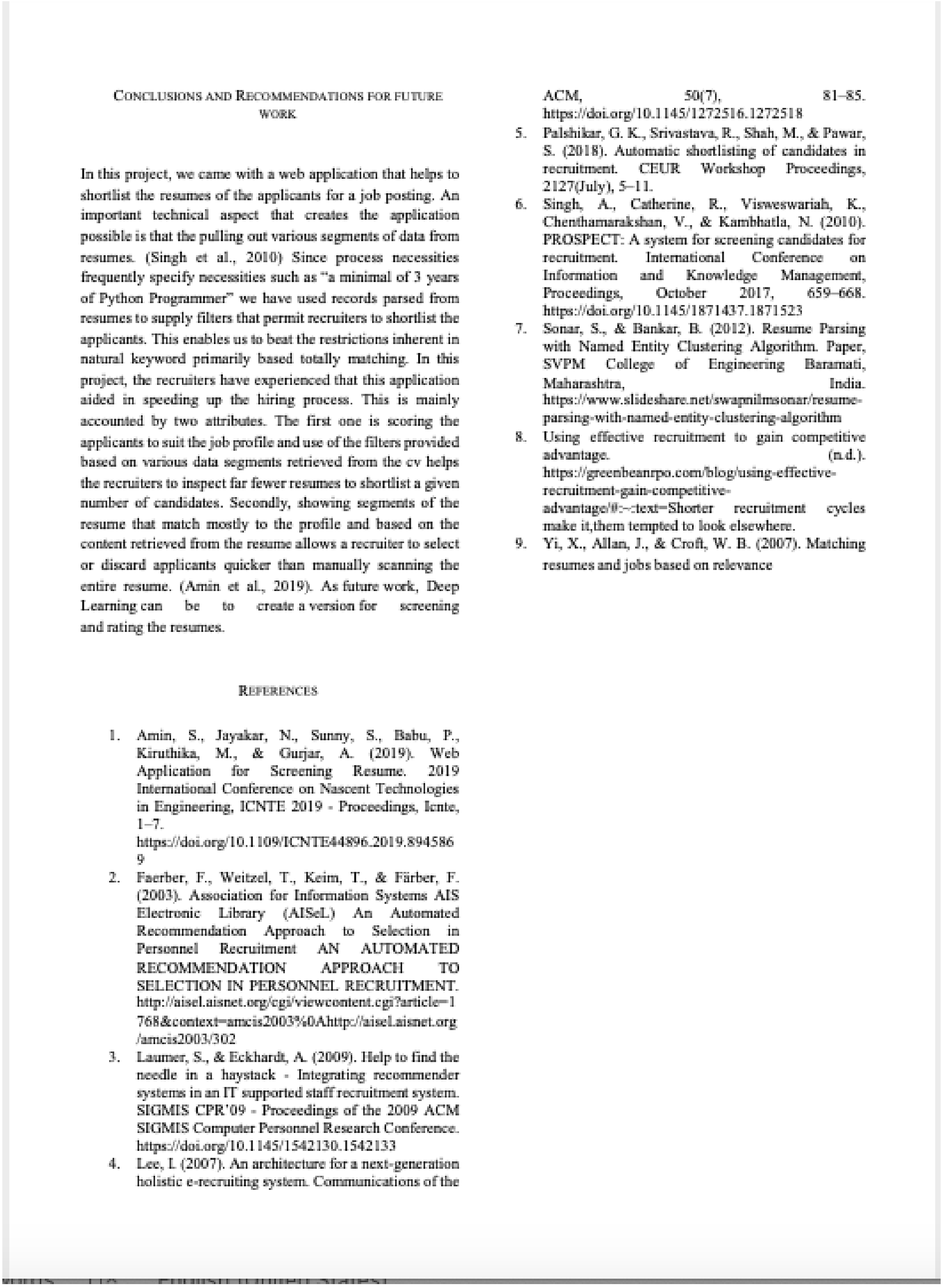
As shown in the formula, we have set the following weighting values:

Skills weight = 50%, Educational level weight = 20%, Job experience weight = 20% and Loyalty level weight = 10%. The results of using the scoring formula are detailed in the next section

VII. EVALUATION

To evaluate our system, we used precision, recall, and measure. Precision is usually defined as the dividend of the number of related records extracted and the sum of number of related and not related records retrieved. For block segmentation, precision is calculated as the % of correct segments detected and extracted to the whole number of correct and wrong segments that were detected and extracted. For feature extraction, precision is the percentage of correctly extracted feature that matches to the ground truth data. The recall is calculated as the dividend of the no of related records extracted and the whole number of related records present in the database. In block segmentation, recall is the % of correct segments detected and retrieved to the sum number of correct segments present in the database. For feature extraction, recall is the % of correctly retrieved feature to total features present in the training data F1 score is-measured as the HM (Harmonic Mean) of above two metrics (precision and recall).

Recognized Entity	Precision	Recall	F-Score
College Name	100.0%	100.0%	100%.0
Location	100.0%	97.78%	98.88%
Designation	100.0%	100.0%	100.0%
Email Address	95.83%	100.0%	97.87%
Name	100.0%	100.0%	100%.0
Skills	96.36%	96.30%	96.30%
Years of Experience	100.0%	100.0%	100.0%
Graduation Year	96.55%	87.50%	91.80%
Degree	100.0%	100.0%	100.0%
Companies worked at	98.08%	100.0%	99.03%



CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

In this project, we came with a web application that helps to shortlist the resumes of the applicants for a job posting. An important technical aspect that creates the application possible is that the pulling out various segments of data from resumes. (Singh et al., 2010) Since process necessities frequently specify necessities such as "a minimal of 3 years of Python Programmer" we have used records parsed from resumes to supply filters that permit recruiters to shortlist the applicants. This enables us to beat the restrictions inherent in natural keyword primarily based totally matching. In this project, the recruiters have experienced that this application aided in speeding up the hiring process. This is mainly accounted by two attributes. The first one is scoring the applicants to suit the job profile and use of the filters provided based on various data segments retrieved from the cv helps the recruiters to inspect far fewer resumes to shortlist a given number of candidates. Secondly, showing segments of the resume that match mostly to the profile and based on the content retrieved from the resume allows a recruiter to select or discard applicants quicker than manually scanning the entire resume. (Amin et al., 2019). As future work, Deep Learning can be to create a version for screening and rating the resumes.

REFERENCES

1. Amin, S., Jayakar, N., Sunny, S., Babu, P., Kiruthika, M., & Gurjar, A. (2019). Web Application for Screening Resume. 2019 International Conference on Nascent Technologies in Engineering, ICNTE 2019 - Proceedings, Icnte, 1–7. https://doi.org/10.1109/ICNTE44896.2019.8945869
2. Faerber, F., Weitzel, T., Keim, T., & Färber, F. (2003). Association for Information Systems AIS Electronic Library (AISeL) An Automated Recommendation Approach to Selection in Personnel Recruitment AN AUTOMATED RECOMMENDATION APPROACH TO SELECTION IN PERSONNEL RECRUITMENT. http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1768&context=amcis2003%0Ahttp://aisel.aisnet.org/amcis2003/302
3. Laumer, S., & Eckhardt, A. (2009). Help to find the needle in a haystack - Integrating recommender systems in an IT supported staff recruitment system. SIGMIS CPR'09 - Proceedings of the 2009 ACM SIGMIS Computer Personnel Research Conference. https://doi.org/10.1145/1542130.1542133
4. Lee, I. (2007). An architecture for a next-generation holistic e-recruiting system. Communications of the ACM, 50(7), 81–85. https://doi.org/10.1145/1272516.1272518
5. Palshikar, G. K., Srivastava, R., Shah, M., & Pawar, S. (2018). Automatic shortlisting of candidates in recruitment. CEUR Workshop Proceedings, 2127(July), 5–11.
6. Singh, A., Catherine, R., Visweswariah, K., Chenthamarakshan, V., & Kambhatla, N. (2010). PROSPECT: A system for screening candidates for recruitment. International Conference on Information and Knowledge Management, Proceedings, October 2017, 659–668. https://doi.org/10.1145/1871437.1871523
7. Sonar, S., & Bankar, B. (2012). Resume Parsing with Named Entity Clustering Algorithm. Paper, SVPM College of Engineering Baramati, Maharashtra, India. https://www.slideshare.net/swapnilmsonar/resume-parsing-with-named-entity-clustering-algorithm
8. Using effective recruitment to gain competitive advantage. (n.d.). https://greenbeanrpo.com/blog/using-effective-recruitment-gain-competitive-advantage/#:~:text=Shorter recruitment cycles make it,them tempted to look elsewhere.
9. Yi, X., Allan, J., & Croft, W. B. (2007). Matching resumes and jobs based on relevance