Leveraging Word Embedding for Automated Candidate Ranking in Talent Acquisition Processes

R.Jeyasumangala
MS22040930

A THESIS
SUBMITTED TO
SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

December 2024

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science in Information Technology.

____

Prof.Anuradha Karunasena

Approved for MSc. Research Project:

____

MSc. Programme Co-ordinator, SLIIT

Approved for MSc:

____

Head of Graduate Studies, FoC, SLIIT

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. Also, I hereby grant to Sri Lanka Institute of Information Technology the non-exclusive right to reproduce and distribute my dissertation in whole or part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Student Name – Jeyasumangala Rasanayagam

Registration Number – MS22040930

Signature: ………………………………… Date: 25.01.2025 …………………….

ABSTRACT

Leveraging Word Embedding for Automated Candidate Ranking in Talent Acquisition Processes

Jeyasumangala Rasanayagam

MSc. in Information Technology

Supervisor: Prof.Anuradha Karunasena

December 2024

Ranking the applicants who applied for a certain position in a company is mostly done manually. To ease this process, this system creates a ranking system by giving scores for each applicant based on the word embedding model trained using the past datasets. The job advertisements related to information technology fields or related to certain positions are collected and trained a model using the word embedding process. The system compares the resume of the applicant with this model and allocates a specific score for each applicant and orders them in the ascending order. Data crawling and scraping, text preprocessing and training the model are the main components in this research.

The goal of this research is to collect the data of job openings related to the information technology industry and collect the job seekers information through the web scraping and crawling and train a model to rank the applicants. The crawled data is used to prepare the corpus. Python scrapy is used to prepare the crawler script for this crawling mechanism. The crawled data is then undergone for the preprocessing. Finally, the preprocessed corpus is undergone for the word embedding. Word2Vec, Gensim are some algorithms used here to train a model. This model is used to compare the resumes of each applicant and get value from the model and finally it will output a total score for each resume and then the system finally ranks the applicants based on the scores they got in ascending order.

Key words: *Scraping, Crawling Mechanism, Text preprocessing, Stemming, Lemmatization, Word embedding, Machine Learning, Model, Gensim, Word2Vec*

ACKNOWLEDGEMENT

First and foremost, I express my gratitude to Almighty God for the blessings bestowed upon me throughout the duration of this research work. Under the research module during the second year of my Msc programme, I could have gained an invaluable experience and I enriched my skills in across various domains within the Information Technology. First of all, I extend my heartfelt appreciation to our supervisor Dr.Anuradha Karunasena for her exceptional guidance and unwavering encouragement throughout this journey. Even during challenging periods in the project completion, she provided invaluable advice, urging me to preserve and exert maximum effort towards the project's success. Their mentorship steered me in the right direction, illuminating the path towards achieving my objectives. Also, I am really grateful to them for giving their valuable time throughout this year. I also extend my sincere appreciation to all those contributed for the success of this project. I would really like to express my thanks to Sri Lanka Institute of Information Technology for providing us with the expertise to finish the mission. Nevertheless, I express my sincere thanks to my parents and family for believing in me and giving mental support and colleagues for their cooperation and continuous encouragement which helped me immensely in the final touch of this project.

Jeyasumangala Rasanayagam

Faculty of Graduate Studies and Research,

Sri Lanka Institute of Information Technology.
Sri Lanka.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

Abbreviation	**Description**
IT	Information Technology
UI	User Interface
ML	Machine Learning
URL	Uniform Resource Locator
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
RegEx	Regular Expression
POS	Part-of-Speech
PDF	Portable Document Format
CV	Curriculum Vitae