Project Outline & Abstract

Project Title: Automated Keyword Extraction from Job Descriptions

Team Members:

1. Naveen Kumar Kundla - 016021941
2. Sivakrishna Yaganti - 017428853
3. Vinit Kanani - 016651323

Abstract

Applicant tracking systems (ATS) are commonly used by employers to automatically scan resumes for relevant keywords from job descriptions. However, many qualified candidates may be overlooked if their resumes lack explicit matches for keywords. This project aims to automate the extraction of key skills and qualifications from job descriptions using natural language processing. The extracted keywords can then be used to optimize resumes so they contain relevant terms that will enable candidates to successfully pass initial ATS screening. By improving keyword matching between applicant resumes and job descriptions, this system can help job-seekers better tailor their resumes while enabling employers to more accurately identify qualified applicants that may have been previously overlooked. Overall, automated keyword extraction will make the initial resume screening process more efficient for both applicants and employers.

Project Outline

1. Introduction
    a. Problem Statement
    b. Motivation
    c. Related Work
    d. Proposed Solution
2. Literature review
3. Methodology
    a. Data collection and dataset creation
        - Scraped 17000+ job descriptions from greenhouse.io and filtered 5000+ tech related jobs. Extract plain text from HTML pages through parsing and stripping HTML tags, scripts, styling, and other non-text content.
    b. Preprocessing techniques

- Performed text cleaning and normalization by removing special characters and html tags.
- Filtered out non-English job descriptions.

c. NER model selection, fine-tuning, and training.

1. Pre-trained model:
    a. Spacy Models
        - en_core_web_trf - Used the spacy pre-trained transformer model to extract keywords from job descriptions. The model is only able to recognize well-known entities such as Degree and Programming Language. The model is not able to recognize custom entities such as Skills and Qualifications.
2. Fine-tuned model:
    a. Fix annotatations for the job description.
    b. Train model on annotated data.
    c. Evaluate model performance

d. Keyword extraction using TF-IDF and Similarity metrics

- Extract keywords from job descriptions using TF-IDF and Similarity metrics

4. Results

    a. Compare performance of pre-trained and fine-tuned models

5. Conclusion

    a. Summary of results

    b. Future work

6. References

[1] Lison, J. Barnes, and A. Hubin, “skweak: Weak Supervision Made Easy for NLP,” arXiv (Cornell University), Jan. 2021, doi: https://doi.org/10.18653/v1/2021.acl-demo.40.

[2] “Application of Neural Network Keyword Extraction Methods for Student’s CV Compilation from Discipline Work Programs | IEEE Conference Publication | IEEE Xplore,” ieeexplore.ieee.org. https://ieeexplore.ieee.org/abstract/document/10159061 (accessed Nov. 22, 2023).

[3] “Training Pipelines & Models · spaCy Usage Documentation,” Training Pipelines & Models. https://spacy.io/usage/training

[4] “The Stanford Natural Language Processing Group,” nlp.stanford.edu. https://nlp.stanford.edu/software/CRF-NER.html

[5] M. Ghadge, “Building Your Own Custom Named Entity Recognition (NER) Model with spaCy V3: A Step-by-Step Guide,” Medium, Sep. 06, 2023. https://medium.com/@mjghadge9007/building-your-own-custom-named-entity-recognition-ner-model-with-spacy-v3-a-step-by-step-guide-15c7dcb1c416 (accessed Nov. 22, 2023).

Dataset

The data for this project will consist of 17000+ job descriptions scraped from greenhouse, a popular recruiting software platform.

Ethical considerations

Global Impact

The project would help job-seekers to better tailor their resumes resulting in more efficient job search. This would also benefit employers by enabling them to more accurately identify qualified applicants that may have been previously overlooked. But, the project can also be misused to bypass the initial screening process and get unqualified candidates through the door.

Economic Impact

This project would benefit both the job-seekers and employers by making the initial resume screening process more efficient. However, if the candidates misuse the project to bypass the initial screening process, it would result in higher costs for employers.

Environmental Impact

The project would not have any significant environmental impact. The training of models would require computing resources, but the impact would be negligible.

Societal Impact

The project can have a significant positive impact to the well qualified candidates. However, the project can also be misused to deceive the employers.