**Class:**.

**Date:** 11/21/2023

**Team 2:** Amjad Altuwayjiri, Muhannad Alwhebie, Nammin Woo

Project Proposal(updated)

**Object**

Our project aims to enable two-way communication by Integrated NLP model.

Firstly, it can help job seekers become more literate about occupations and the job market.

Secondly, it can help to match recruiters with the best candidates in their resume. This approach promises to refine the job-seeking and hiring process, by saving time for job seekers to acquire information and insight, and recruiters aligning job seekers with suitable opportunities more effectively.

**Database**

The [ONET® Database](https://www.onetcenter.org/database.html#all-files) is a comprehensive database that contains a rich set of variables describing work and worker characteristics, including skill requirements.

It sponsored by the U.S. Department of Labor, Employment and Training Administration (USDOL/ETA) and updated quarterly.

**Problem Description**

There are Classical search engines ([onet online](https://www.onetonline.org/), [my next move](https://www.mynextmove.org/)) sourcing from ONET DB. And, it is surely beneficial since they provide a wide range of information, which allows job seekers to explore different aspects of an occupation. However, it sometimes could be overwhelming if users have a specific question in mind because they might have to sort through a lot of data to find the solution.

A Chatbot, an NLP model based on ONET DB, could be a complementary method to resolve this problem. By using a chatbot, job seekers can save time by getting a direct answer to their questions. Furthermore, a feedback loop where job seekers can ask follow-up questions based on previous answers could be a powerful feature to lead to deeper insights. Also, the Chatbot might be a useful solution to resolve a challenge in effectively matching recruiters with the most appropriate candidates by accurately analyzing the skills, education, and experience outlined in their resumes and matching them with occupation information or job requirements.

**Design NLP Model**

**1. Reading ONET data**

**2-1. Main Chatbot**

**What is the goal of the chatbot?**

* Job seeker side (ex)

Q. which occupation is the hottest in terms of money in the data analysis field?

A. The hottest job at the moment is … and avg salary is …, needs skills.

* Recruiter side (ex)

Q. I want to perform quantitative screening for candidates’ resumes.

A. The best resume shows % matching score, because...

**WHAT SHOULD WE DEFINE Specifically to build a good chatbot model?**

**What kinds of transformer can we use, How to train and evaluate a model?**

- Pour the ONET data to Auto tokenizer, transformer ()

**2-2. Sub Models**

We will employ classical NLP methods such as vectorization techniques—namely, Bag of Words or TF-IDF and Rule based Matcher. Then apply cosine similarity to score and rank candidates against job descriptions. In addition to BERT to generate embeddings that capture the context around each skill or job requirement.

Ex. Matching models (or topic Modeling) for support Recruiter side.

- Recruiter input requirement for job position (skills, education, etc.)

- Matching model 1: aligning occupation codes (or topic clusters) based on ONET DB

- Recruiter chose appropriate occupations (generalize)

- Matching model 2: Input resumes and get the best matched resume: similarity (%) with explanation

**3. Steamlit (App implementation)**

Ex. Choose side (Job seeker or Recruiter)

Main Chatbot: Q and A section, Sub Matching Platform (Inquire – Input info – Answer)

**Reference**

[Rule Based Matching (SpaCy with Resume and Job Skill data)](https://deepnote.com/@abid/spaCy-Resume-Analysis-81ba1e4b-7fa8-45fe-ac7a-0b7bf3da7826?Job_Category=BUSINESS-DEVELOPMENT)

[Word2Vec model(offline)](https://www.analyticsvidhya.com/blog/2023/01/an-approach-to-extract-skills-from-resume-using-word2vec/)

[BERTopic model (+ GPT summary)](https://medium.com/gopenai/clustering-financial-news-using-bertopic-and-gpt-in-5-simple-steps-4e2f7a80c959)

**Chatbot (TBD)**

\*Supportive DBs

[Resume DB 1 (Kaggle)](https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset/data), [Resume DB 2 (huggingface)](https://huggingface.co/datasets/brackozi/Resume/viewer/default/train?f%5bCategory%5d%5bvalue%5d=%27Network%20Security%20Engineer%27), [Job skill DB](https://github.com/kingabzpro/jobzilla_ai/blob/main/jz_skill_patterns.jsonl)

***TBD***

***Packages used***

*Pandas, NumPy, Scikit-learn, NLTK*

***NLP tasks***

*Feature extraction, Similarity calculation, Ranking, Evaluation, Iteration and optimization, Contextual embedding generation, Fine-tuning of pre-trained models.*

***Performance evaluation***

*Manual Evaluation: manually review the recommendations for a sample of job descriptions to assess their relevance and quality.*

***Schedule***

***Day 1 - Day 2: Data Acquisition and Setup***

*- Obtain and load the resume dataset.*

*- Install necessary Python packages.*

***Day 3 - Day 4: Feature Extraction***

*- Apply TF-IDF to the resume data.*

*- Implement initial cosine similarity scoring.*

***Day 5 - Day 6: Evaluation Framework***

*- Set up a manual evaluation process.*

*- Define key performance metrics.*

***Day 7 - Day 8: System Refinement***

*- Refine feature vectors and similarity calculations.*

*- Gather initial feedback and adjust the model.*

***Day 9 - Day 10: Performance Tuning***

*- Analyze system performance and tune parameters.*

*- Implement recruiter feedback loop.*

***Day 11 - Day 12: Final Evaluation***

*- Conduct final evaluations with expanded job descriptions.*

*- Prepare the system for a pilot deployment.*

***Day 13: Reporting***

*- Document the methodology, system setup, and performance.*

*- Compile and finalize the project report.*

