Workforce Analytics Group Coursework

The Python code for this homework is submitted with filename ***WorkforceGrp7HW.ipynb***

1a. Main Competitors

To be a competitor, the individual has to first be a suitable candidate for the job. To assess suitability, the candidate's cover letter is matched against the job description and their similarity is checked. Next, having filtered out the unsuitable candidates, we compare our group (Group 7) candidate's cover letter similarity with the remaining suitable candidates; the candidate with the highest similarity with us is chosen as the main competitor. The detailed text processing steps and similarity calculations are described below.

Building list of words

The following text processing steps are performed before calculating the similarities of cover letters:

1. Nouns and Verbs Extraction
    - For each cover letter and the job description, we begin by tagging each word (using NLTK's POS tagging) with their appropriate part of speech. Next, we extract the nouns and verbs and build the list of verbs and list of nouns separately.
    - We keep only the nouns and verbs because these words could represent the skills and knowledges a candidate has. These are important information to assess a candidate's suitability and similarity with other candidates.
2. Lemmatization
    - These lists (of nouns and verbs) are then lemmatized. This is to prevent treating similar words like "presentation" and "presentations" differently. Two words (with the same lemma) but written in different form (e.g. different tenses), should still be considered a match.
3. Stopwords Filtering
    - First, we remove the common stopwords (found in WordNet database) from the list of words extracted from the Cover Letters and Job Description.
    - Using the list of words from the Cover Letters, we build two global word lists (one global list for nouns, another for verbs). The words that appear most frequently (top 25 percentile) in each of the global lists are classified as stop words. The reason behind this "frequent stopwords filtering" is that these overused words (which appear frequently in most candidates' Cover Letters) would not differentiate a candidate from the rest. Hence it would not be informative in assessing job suitability or skillset similarity.
    - These frequent stopwords are then removed from the Cover Letters wordlists. The result is a list of unique verbs and a list of unique nouns for each group that appear in the respective cover letter, excluding the frequent stop words.
    - Note that frequent stop words are not removed from the job description. This prevents removing important key words in the job description that might be important to the job (e.g. "Python" should not be removed).

Assessing Similarity

Similarity is based on the number of words that appear in both groups' list of words. The word matching can be either verb or noun. The similarity score (of two texts) is calculated as follows.

$$Similarity = \frac{Number\ of\ unique\ words\ (verbs\ \&\ nouns)\ appearing\ in\ both\ texts}{Number\ of\ unique\ words\ (verbs\ \&\ nouns)\ appearing\ in\ either\ text}$$

Results – Job Suitability

To assess each candidate's suitability for the job, we first calculate the similarity between each group's cover letter and the job description. Given the seven similarity scores, we filter out the bottom four groups, keeping only the top three candidates who are most suitable for the job (highlighted in bold in Table 1). The seven similarity scores are shown below.

Group	1	**2**	3	**4**	**5**	6	7	8
Similarity	0.0380	**0.0438**	0.0347	**0.0414**	**0.0564**	0.0162	0.0267	0.0287

Table 1 Job Suitability (i.e. Similarity with Job Description) Score of Each Candidate – Top 3 Candidates Highlighted

Results – Candidates Most Similar to Group 7

After narrowing down to the top three candidates, we will now need their similarity scores (with our group – Group 7) to determine the strongest competitor. The one who is most similar to Group 7 will be our main competitor.

To derive the skillset similarities between groups, we calculate the Cover Letter similarities between each pair of groups. The similarities are calculated and represented in the graph below, where node represent group (the top three candidates are highlighted in yellow) and edge represent the similarity strength between two nodes (the darker the edge, the higher the similarity score).

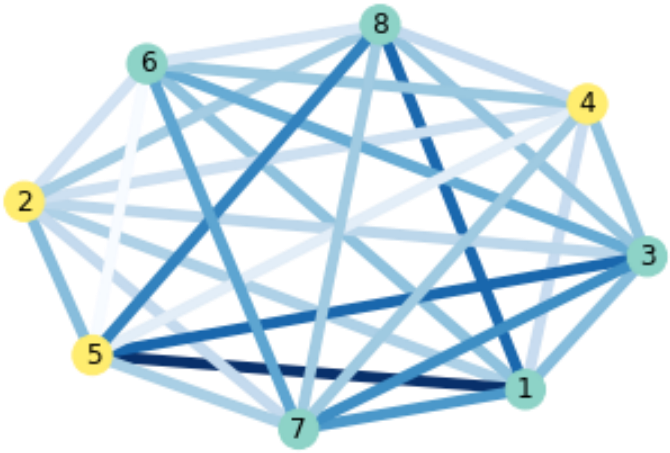

Figure 1 Cover Letter Similarity Network

From Table 1, groups 2, 4 and 5 are the top candidates. We compare their Cover Letter similarities to our group (Group 7). The similarity score are tabulated below:

Group	Similarity with Group 7
2	0.0281
4	**0.0378**
5	0.0352

Table 2 Top Candidates' Similarity Scores with Group 7

We find that amongst the three groups, group 4 is most similar to our group, and is thus our strongest competitor.

1b. Limitations of Techniques (Approx. 100 words)

Limitations include the lack of spell checker, as well as the inability to recognize certain synonyms. WordNet has a database of synonyms but it is unable to recognize synonyms in this context (e.g. unable to recognize "programmer" and "developer" are synonyms in software context). Also, the technique matches word but not phrase (e.g. "Machine Learning" should be matched as a phrase, instead of matching "Machine" and "Learning" individually).

In addition, the similarity score is calculated based on Cover Letter word matching. This is entirely based on what the interviewees wrote on Cover Letter rather than actual evidence of their abilities. Further, some important keywords might be filtered as stopwords (if they appear frequently in all Cover Letters) and won't be factored in for the job suitability assessment.

Lastly, keyword matching is not good in creating a complete picture of the candidate. It does not accurately assess other professional experience (e.g. marketing skills honed during internships, additional big data programming language such as Spark) since these words do not appear in the original job description.

2. Best Complement and Methodology

The first subsection below briefly summarises the techniques (in deriving the candidate best complement our group) in approximately 100 words. The following subsections describe the methodology in more details.

Brief Description of Techniques (Approx. 100 words)

To complement our candidate's skills, an individual would need to have least skillset similarity with our candidate. In addition, to ensure that the individual makes an effective team with our candidate, we would prefer moderate amount of team diversity. We use ethnicity as a proxy for inherent diversity, and highest education level as a proxy for acquired diversity. The steps are:

1. Extract Education Level and Last Names from the Cover Letters
    - The last name will be used to identify ethnicity of the candidate. Exact last name matching will first be performed. If no exact match is found, metaphone matching will be used.
2. Calculate diversity score based on ethnicity (inherent diversity) and education level (acquired diversity).
3. Narrow down the candidates with moderate amount of team diversity when teaming with our group – Group 7.
4. Among these candidates, the one will lowest skillset similarity with Group 7 will be the complement. Two different ways of assessing skillset similarities have been attempted and they yield consistent results, where Group 5 best complements our candidate (described in 'Results' section).

Detailed Description of Methodology

Diversity Variables

Ethnicity (Proxy for Inherent Diversity)

To extract ethnicity, we first extract the last name of each candidate. The last name is always the last word in the list, as candidates sign off with their names. For each cover letter, the last word in the list is used for exact last name matching in order to identify the ethnicity. If there is no exact match in the ethnicity, we match using metaphones of the last name. If there is more than one metaphone match, we randomly assign from the possible ethnicities. If there are still no matches (for metaphone matching), then the ethnicity of is labelled as "Others".

Education Level (Proxy for Acquired Diversity)

All candidates in the pool have either a PhD or a Master's degree. To identify a candidate's highest qualification, we search the entire Cover Letter for the keywords. If "PhD" is mentioned in the cover letter, we assume that to be their highest qualification. Otherwise, if some variations of a Master's degree like "MSc", "Master's" or "Master" is mentioned instead, then the highest qualification is assigned as "Masters".

Combined Diversity Scores

Next, we would like to combine the two proxies into a single diversity score when an individual teams with our Group 7 candidate to form a 2-person team. For this purpose, we use the following formula.

$$Diversity = 1 - \frac{Number\ of\ common\ characteristics}{Number\ of\ characteristics\ groups\ considered\ (i.e.\ ethnicitiy\ \&\ education)}$$

Since we have only 2 characteristics groups (ethnicity & education), the possible values will be:

- 0, when both ethnicity and education level are the same
- 0.5, when one of the two characteristics is the same
- 1, when both characteristics are different

Recall that for optimal team performance, we would like to have moderate amount of diversity so that the team members can complement each other while avoiding potential communication issues. Hence, we will narrow down to individuals who have a diversity score of 0.5 when teaming with Group 7 candidate. The diversity scores are tabulated below and the groups which can potentially complement our group are highlighted in bold.

Group	Ethnicity	Highest Education	Diversity with Group 7
1	**Others**	**Masters**	**0.5**
2	Others	PhD	1
3	Chinese	Masters	0
4	English	PhD	1
5	**Others**	**Masters**	**0.5**
6	**Indian**	**Masters**	**0.5**
7	Chinese	Masters	-
8	**Others**	**Masters**	**0.5**

Table 3 Each Group's Diversity Score when Teaming with Group 7

Skillset Dissimilarity

Now we have a pool of candidates who can potentially complement our group. Next, we would need to identify the candidate who has different skillsets from our group's candidate. This allows both individuals to complement each other's skill gaps.

To estimate this skillset dissimilarity, we adopt two methods:

1. Use the Cover Letter similarity scores from Part 1 to find out the candidate who is most dissimilar to our own candidate.
2. Since the Digital Interviewer job advertisement is pertaining to a Sales Engineer role, the ideal candidate should possess both Sales and Software Engineer skills. Hence two typical job descriptions from the web (Workable, 2017) are used (one for sales job and the other one for a software engineer job).
    - In order to assess a candidate's suitability to either role, we check the similarity of his/her cover letter with the respective job description. We obtain the similarity score for both sales and software engineer job descriptions. An individual is thought to possess a certain skills if his similarity score (of the respective category) is above average.
    - Ideally, if our group's candidate possesses sales skills, an ideal complement would be someone with software engineering skills.

Results

Method 1 – Similarity Score in Part 1 (Cover Letter Word Matching)

From Table 3, group 1, 5, 6, and 8 are the candidates who can potentially complement our group. Thus, we revisit their Cover Letter similarity scores with Group 7, as derived in Part 1.

Group	Similarity with Group 7
1	0.0561
5	**0.0352**
6	0.0515
8	0.0372

Table 4 Potential Candidates' Similarity Scores with Group 7

As observed from Table 4, group 5 is most dissimilar to our group and is hence the best complement. We now use the other method and find out if it yields consistent result.

Method 2 – Similarity Score with Sales and Software Engineer Job Descriptions

Using the two job descriptions from web, we calculate each group's cover letter similarity with these two JDs. As a result, we have two scores for each group: sales JD similarity and software engineer JD similarity.

The two scores for each group are then plotted on a 2-D graph, in Figure 2. The x-axis represents the sales JD similarity (proxy for sales-related skills) and y-axis represents software engineer JD similarity (proxy for software engineer-related skills). The two dashed lines indicate the average similarity values with these two JDs.

As seen in the graph, our group (group 7) candidate has above average sales skills but below average software engineering skills. Thus, a suitable complement would be candidates 1 or 5 (who have above average software engineering skills and moderate diversity score when teaming with Group 7). Out of these two candidates, candidate 5 possesses better software engineering skills and hence would complement our candidate (who has good sales-related skills).

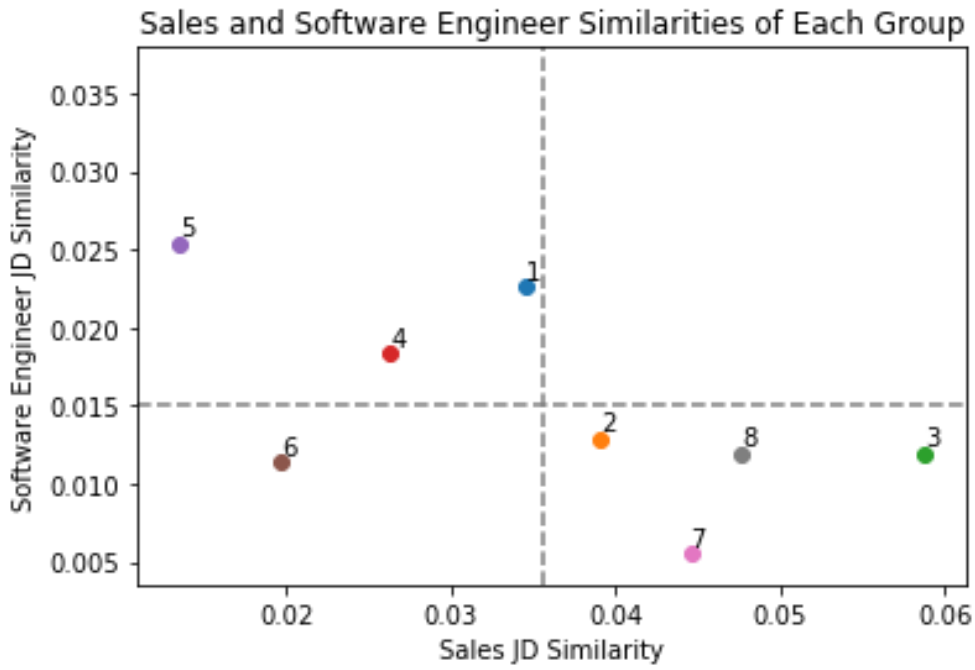

Figure 2 Sales and Software Engineer JD Similarities for Each Group

Since both methods produce consistent results, we conclude that group 5 is our best complement (moderate diversity score and most dissimilar skillsets).

Conclusion

As described in Part 1b, the text processing techniques used in this assignment have their limitations. Consequently, the job suitability assessment and similarity measurement may be flawed since word matching does not give a complete picture of the candidate's full experience and skills. While the analytical methods discussed here may be useful in reducing the large candidate pool to a few potentially good candidates, a human interviewer will still be required to give a holistic assessment of the candidates' experience, capabilities, and suitability to the job advertised.

References

Workable. (2017) *Sales Representative job description.* Available from: https://resources.workable.com/sales-representative-job-description [Accessed 1st June 2017]

Workable. (2017) *Software Developer job description.* Available from: https://resources.workable.com/software-developer-job-description [Accessed 1st June 2017]