User Guide

Steps to be followed

1. To resolve the problem of screening their resumes or not being shortlisted we came up with a resume skills recommender which solves this problem.
2. Requirement analysis is done by collecting a series of datasets of resume and analysis of those resumes based on the job description
3. The project is deployed in Google Colab which is an open source software with Colab notebooks that execute code on Google's cloud servers.
4. The data was downloaded from the online portal like LinkedIn and also from a collection of resumes. The data is in Excel format, with three column ID, Category, and Resume.
5. We have identified the scope of the data which is related and most important for our project based on view of jobs in broader range
6. After that we have scraped the resumes of the dataset to fit in the correct job archetype.
7. Since the data contains noise text processing is applied on the resumes being provided as input that would be cleansed to remove special or any junk characters that are there in the CVs/resumes.
8. The stop words such as and, the, was, etc. are frequently appeared in the text and not helpful for prediction process, hence they are removed.
9. Stemming is applied for decreasing word inflection to its root forms such as mapping a group of words to the same stem even though the stem itself is not a valid term in the language.
10. The next step after pre-processing is feature extraction. On pre-processed dataset, we have extracted the features using the Tf-Idf (term frequency, and inverse document frequency)
11. Then the Count Vectorizer concludes that resumes mentioning the “popular” words are more similar with each other as it puts more emphasis on these popular words.
12. Topic Modelling is done reducing the resumes into specific categories.
13. Topic modelling techniques are used for Dimension reduction like Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA). The objective of LSA is reducing dimension for classification. The idea is that words will occurs in similar pieces of text if they have similar meaning.

14. Then we apply the Similarity measure which is used to rank the documents by calculating the similarity between the document and the query. Commonly used measures are inner product, cosine coefficient, Jaccard coefficient etc.,
15. Calculating the similarity score gives the percentage or probability of getting shortlisted in that job description
16. Overall, the system performs pretty well with the current resources. As a part of our future work, we intend to improve the accuracy of our system by collecting more resumes from organizations and training our model for all the available roles.