Automated Vacancy Recommender

Team Members:

Mahad Khalid Tarar (18K-0187)

Abdullah Raheel (18K-0170)

Muhammad Ammar bin Nasir (18K-1037)

Introduction

Talent acquisition is an important, complex, and time-consuming function within Human Resources (HR). The most challenging part is the lack of a standard structure and format for a resume which makes short listing of desired profiles for required roles very tedious and time-consuming.With a huge number of different job roles existing today along with the typically large number of applications received, short-listing poses a challenge for the human resource department. Which is only further worsened by the lack of diverse skill and domain knowledge within the HR department, required for effective screening.

Today the industry face three major challenges:

- Separating right candidates from the pack - This makes the whole hiring process slow and inefficient, costing resources to the companies.
- Making sense of candidate CVs - practically every resume in the market has a different structure and format. HR has to manually go through the CVs to find the right match to the job description. This is resource intensive and prone to error whereby a right candidate for the job might get missed in the process.
- Knowing that candidates can do the job before you hire them -The third and the major challenge is mapping the CV to the job description to understand if the candidate would be able to do the job for which he/she is being hired.

To overcome the mentioned issues in the resume short-listing process, we developed an automated Machine Learning based model. The model takes the features extracted from the candidate’s resume as input and finds their categories, further based on the required job description the categorised resume mapped and recommends the most suitable candidate’s profile to HR.

Methodology

The aim of this work is to find the right candidate's resume from the pool of resumes. To achieve this objective,we have developed a machine learning based solution. The ***Dataset*** to train the model was downloaded from ***Kaggle.***

Preprocessing:

In this process, the CVs being provided as input would be cleansed to remove special or any junk characters that are there in the CVs. In cleaning, all special characters, the numbers, and the single letter words are removed. We got the clean dataset after these steps having no special characters, multi-spaces or single letter word. The dataset is split into the tokens. Further, the preprocessing steps are applied on tokenized dataset such as ***stop word removal***, ***case-folding***, and ***lemmatization***.

Stop words removal: The stop words such as and, the, was, etc. frequently appear in the text and are not helpful for prediction process, hence they are removed.

Case-folding: The process of converting all the characters in a document into the same case, either all uppercase or lowercase, in order to speed up comparisons during the indexing process.

Lemmatization: lemmatization decreases the inflected phrases to ensure that the root word belongs to the language correctly**.**

Feature extraction:

On the preprocessed dataset, we have extracted the features using the Tf-Idf .
The machine learning based classification model or learning algorithms need a fixed size numerical vector as input to process it. ML based classifiers do not process the raw text having variable size in length. Therefore, the texts are converted to a required equal length of vector form during the preprocessing steps. Specifically, we have calculated tf-idf (term frequency, and inverse document frequency) for each term present in our dataset using the *scikit learn* library function:
sklearn.feature extraction.text.TfidfVectorizer to calculate a tf-idf vector.

Results

Two models have been built on the cleansed data: i) Classification - Based on the resume and category the model has been designed to categories the resume in the right category and ii) Recommendation -The model would give the list of most relevant resume based on the similarity between resume and jobs description.

Classification:

The classification was done using three different models and their accuracy score was recorded.

1. **Random Forest**: It is an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) of the individual trees.
2. **Logistic Regression**: It uses a logistic function to model a binary dependent variable.
3. **Linear Support Vector Classifier** (Linear SVM): A SVM is a supervised machine learning classifier which is defined by a separating hyperplane. In two-dimensional space, a hyperplane is a line which separates a plane into two separate planes, where each plane belongs to a Class.

The accuracy score of Linear Support Vector Classifier was higher compared to other models and we found this model to be reliable and best fit for our objective.

CV Recommendation Model:

The recommendation model is designed to take job description and CVs as input and provide the list of CVs which are closest to the provided job description.Considering this is the case of document similarity identification, we have gone with the Content-based recommender where Job Description provided by the employer is matched with the content of resumes in the space and the top n matching resumes are recommended to the recruiter. The model takes the cleansed resume data and job description and combines the two into a single data set, and then computes the cosine similarity between the job description and CVs.

Conclusion

Huge number of applications received by the organization for every job post. Finding the relevant candidate's application from the pool of resumes is a tedious task for any organization nowadays. The process of classifying the candidate's resume is manual, time consuming, and a waste of resources. To overcome this issue, we developed an automated machine learning based model which recommends suitable candidate's resumes to the HR based on given job description. The proposed model worked in two phases: fifirst, classify the resume into different categories. Second, recommends resumes based on the similarity index with the given job description.By involving the domain experts like HR professional would help to build a more accurate model, feedback of the HR professional helps to improve the model iteratively.

Github Link

https://github.com/PeinZero/Automated-Vacancy-Recommender

References

- ★ https://stackoverflow.com/
- ★ https://scikit-learn.org/stable/user_guide.html
- ★ https://datascience.stackexchange.com/
- ★ https://discuss.analyticsvidhya.com/