Resume matching to a job

Uddhipan Thakur

June 25, 2019

The job of a resume matching platform used by a small business would be to create a ranked list of resumes for a particular job. Typically, the company will post a job on a platform and within the specified time, applicants will upload their resumes- usually in a pdf format.

Our job is to process both the job posting data and the applicant's resumes before we come out with a similarity measure. Once we have the similarity measure for a particular job J and the list of resumes $R_{1,2..N}$ we can create a sorted list. The top resumes in the sorted list can then be given priority in the hiring process.

1 Similarity measures

Each job usually contains a bunch of requirement features. Features will be skillset, education, location, salary expectation, seniority, etc. One way to match a job to a resume would be to extract the features from a job as keywords an then try to match those keywords with the words of the resumes. Another approach will be to directly consider the document similarity between the job posting and the resume. This approach already takes into account keyword matching, as more matching keywords will give a higher value.

There are several methods to measure document similarity. One easier method to do so will be with cosine similarity.

To measure the cosine similarity between two documents, we have to convert the documents into vectors. We first calculate the total number of unique words in both the documents combined. Then we represent the documents on the word basis i.e the document vector has the length of the total vocabulary and the number of time each word appears in the document will be the positional value in the vector.

If the two document vectors are represented by $\vec{u}, \vec{u}$ then their cosine similarity is given by

$$Sim(U, V) = cos(\theta) = \frac{\vec{u}.\vec{v}}{||u||||v||}$$

We can calculate the cosine similarities between all the resumes and the job, and then sort the scores to get the better suited resumes.

2 Using the code

- The job advertisement is stored in the text file called *job.txt*
- All the candidate resumes in pdf's are stored in a folder called *Candidates*.
- The *pdf2txt.py* python code runs over the resumes in the folder and converts them into text files for processing. The text files are stored in a automatically created folder called *Candidates_txt*. This code should be run first.
- Then the *cv_job_similarity.py* is run. This creates a dictionary called score which is then sorted and printed. The output is a sorted dictionary giving the resume name and scores in descending order.

3 Comments

Both the job text and the resumes are automatically preprocessed. The documents are tokenized, stemmed and filtered for stopwords. This increase the accuracy of the score.

Word embedding approaches are in progress, so that the semantic meaning of the words can be captured better.