What is Clustering?

- Makes group of similar things
- Help to analyze large data

What is the need for Clustering in our project?

- As This project is a two-way recommendation system so at the time of recommending resumes to a particular job we need to iterate each and every resume to find similarity between them and this become very time consuming so using clustering what we are doing we just make the cluster of various resumes and when we have to recommend resumes to a particular job we first find the similarity between each cluster representative and job and we can easily choose the cluster which has high similarity now we just need to compare only one cluster resumes with the particular job description
- So basically, clustering will reduce our model time needed for a run.

Why are we doing clustering for resumes, not for the job description?

- As we know job description has particular job title which is works as class and various features extracted from job description will works as tags
- And in resume we don't have any class name to make group of similar resumes that's why we need clustering on resumes not on job descriptions.

So, for clustering we are using k-mean clustering algorithms to cluster the large data of resumes.

Clean and preprocess the resumes data

- Extract features from resumes
- K-mean need input in numerical only so we need to convert it feature words into vectors
- Now cluster the large data of resumes using k-mean algorithm
- Also, K-mean need input value of k which denotes number of clusters we want to make So find this variable k we are using k-mean clustering algorithms testing method which is elbow method to choose best k for better results on k-mean algorithms

In this figure we can see that squared error is going decreasing with the increasing value of k and after an almost k=10 decreasing in the value of squared error very small so we can choose k = 10.

Here the figure of clusters of resumes when we are choosing k = 10 for k-mean algorithms

Now after clustering we don't need to find similarity between every resume with particular job description, we can just find similarity between cluster's representative and job description and whichever cluster give better similarity we can find top n resumes from that cluster.

What is the progress we made till now and what we will complete till last phase of evaluation?

We completed data extracting from resumes with preprocessing and cleaning of data
We completed job description dataset and resumes data set to train model batter way
We completed job description cleaning and pre processing
We completed classification of job description
We are almost there to complete clustering of resumes

We are going improve clustering of resumes first
Then we are going to implement similarity part where we need to show top n resumes for particular job description and also top n jobs for particular resume

That’s it about our progress on project

In last I want to thank you my all teammate and our guide for constantly support us.