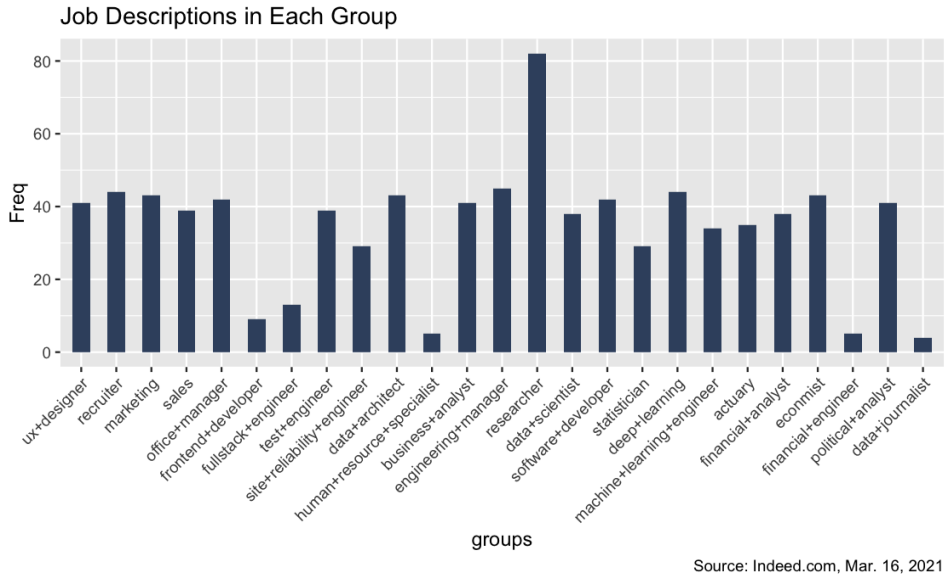
Kate Marsh
Jiaxin Li
March 30, 2021

Project 2: Text-Mining

Introduction:
In this project, we aimed to create an algorithm that takes an input of a resume and outputs three corresponding job descriptions that correspond with the interests and experience on the resume. We had two goals: alignment with the resume and a range of options. Alignment will be discussed as a "goodness" metric, while range will be discussed as a "diversity" metric.

Data Description:
The data is from an API call to Indeed.com that Professor Wayne Lee did, requesting full-time jobs in New York State of 25 different categories (researcher was called twice for some reason). These categories ranged from UX designer to data journalist. Below is a distribution of the job types given by Indeed.com. Something to note is that "researcher" was two separate categories in the API call, but we have combined it into one for the histogram below and our analysis at large, hence the spike in the graph below in the "researcher" column.

Note: Since Kate's resume mainly features jobs on political campaigns and climate change research, it will be difficult to find things from this dataset that align with her keywords. However, Jiaxin has lots of experience associated with statistics and programming and thus will likely have better luck having a good fit from this dataset.



The data cleaning that we did started with isolating the job descriptions into a large list and taking out the stopwords from the qdap dictionary of the top 200 words in the English language.

We also removed punctuation marks and numbers. From this, we created a document-feature matrix which was then made into a TF-IDF or term-frequency inverse document frequency matrix to show how important a word is to the full body of job descriptions. Something that might be significant is that the data includes some concatenated words that should be two words, such as “locationclinical” or “requiredlicense”. Since words like this exist in that dataset, they might skew our results so that both words in these combinations will appear as less frequent in our job descriptions.

To validate the TF-IDF, we looked at the top 15 tokens in each document to make sure that they were reasonable words without (see Figure 1). Another way we validated the TF-IDF was to run Principal Component Analysis on it to see what the loadings were (these came out to be job, salary, hires, ‘timenumber’ and ‘yearjob’). Since the loadings are supposed to be the most important components, this shows that the TF-IDF prioritizes the expected words like “job” and “salary” that are important to each job description.

Algorithms:
We stacked the tf-idx matrix of job description and resume together to form an overall tf-idf matrix. Each row of the matrix represents a document (either job description or resume), and then we calculated cosine similarity between the job descriptions and resume. Users can set the number of job_descriptions they would like to get and the file path of their resume. The algorithms will return job descriptions with the top n similarity scores that have a diversity score of 0.

Our diversity score is a scale from 0 to 2 where 0 is no overlap in cluster or job type, 1 is an overlap in either of those, and 2 for an overlap in both cluster and job type with a prior job description. The job descriptions are ranked from highest to least similar so the algorithm can make sure to output only job descriptions that have a 0 diversity score so that they all have a different cluster and job type. We created the clustering with a hierarchical cluster of type “ward.D2” and chose 17 clusters from the height of 2 after looking at the dendrogram (Figure 7). From there, the clustering was validating by running the model a few times to make sure the clusters did not change, plotting a histogram to make sure the clusters were significantly large (Figure 8), and summarizing the top words in each cluster to make sure they were groups that made sense to a human. For example, the top words in cluster 1 are “design, experience, user, ux, product”.

Our good metric is cosine similarity of the job description and the resume. It is then normalized to a score between 0 and 1 so you can tell how "good" of a recommendation it is, regardless of the raw cosine similarity.

Validating the Model:
For model validation, we first passed in Jiaxin’s resume, and the recommended job descriptions are machine learning engineer, data scientist and deep learning. For Kate’s resume, the recommended job descriptions are related to positions as researcher, economist, political analyst

and statistician. We also passed in a resume of insurance sales to test our algorithm. The returned job descriptions with top 3 similarity scores are all related to sales position (Figure 2).

As seen in Figures 4-6, the proportion of the diversity score were similar for all three resumes, with only a small portion having the 0 diversity score and most of the descriptions having a diversity score of 2, meaning that they are not the first description with both their cluster or resume. This shows that the diversity metric is working as planned, choosing only the small group of descriptions with a high similarity score, unique cluster and unique job type.

We also tested our algorithm by adding 'good' job descriptions (which was found in Jiaxin's resume). Then we used Jiaxin's resume as input, and the appended 'good' job description was recommended by our algorithm as the top choice.

Conclusion:
Our algorithm takes in a resume and amount of recommendations and outputs the top number of job descriptions with highest similarity score and a ranked diversity score. For the example resume (insurance sales resume), the algorithm correctly found they were interested in sales but then gave them some options in other job types and clusters in case they were looking for something different but still relevant to their resume.

Appendix:

Figure 1. First Six Rows of Keyword Lists. Each row represents a different job description. Each column ranks the top tokens in the job descriptions.

	X1 <chr>	X2 <chr>	X3 <chr>	X4 <chr>	X5 <chr>	X6 <chr>
1	creative	strong	experience	years	prestige	ui
2	complex	product	experience	experiences	create	management
3	princeton	online	ux	ui	prototypes	able
4	experience	remote	level	experiences	passionate	developers
5	design	experience	designer	experiences	newspapers	hearsts
6	experience	ui	working	cover	delivering	best

Figure 2. Top 10 Job Descriptions vs Cosine Similarity.

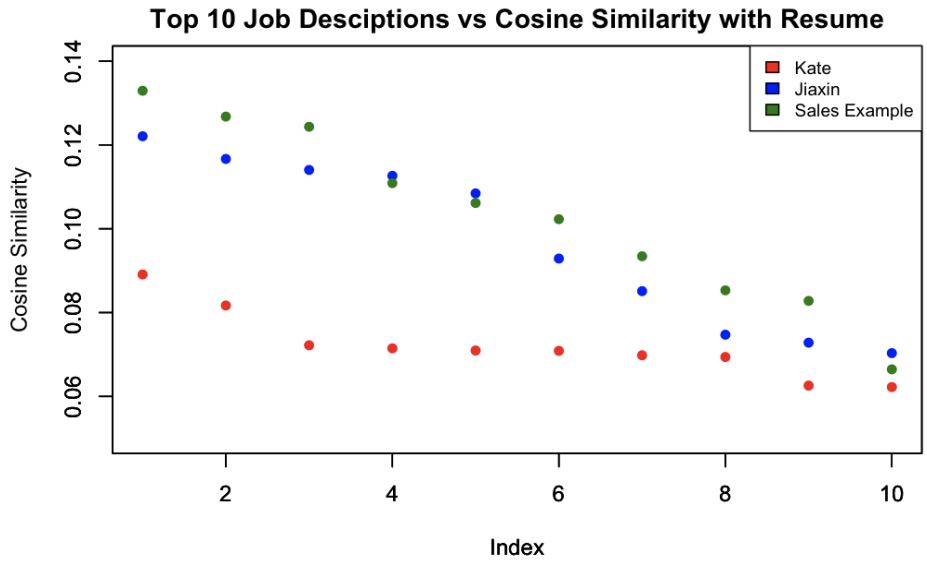


Figure 4-6. Proportions of the Diversity Metric After Running Algorithm on Each Resume.

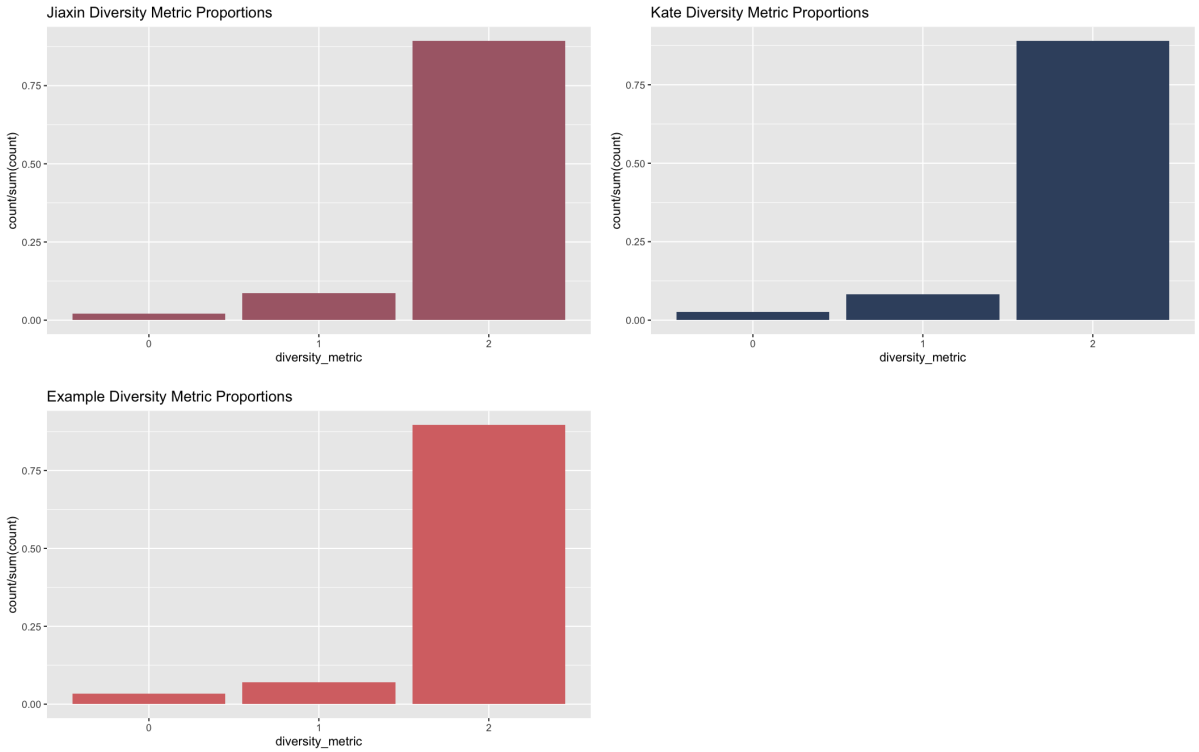


Figure 7. Dendrogram of Hierarchical Clustering of TF-IDF for Job Descriptions. Red Boxes are around the cluster groups from height = 2.

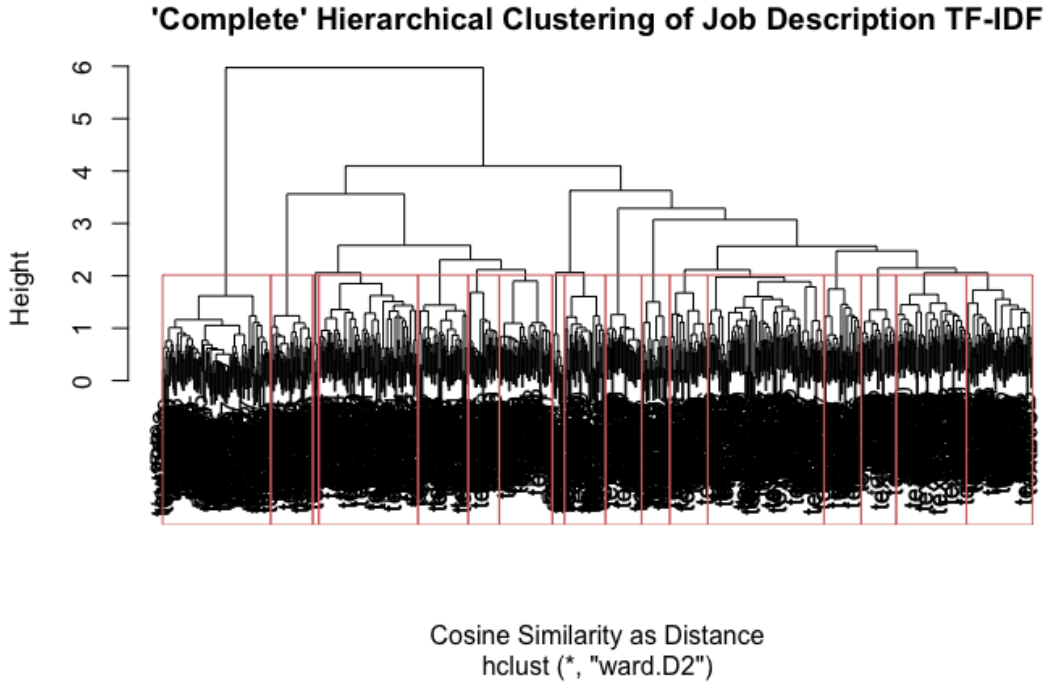


Figure 8. Histogram of the Amount of Job Descriptions in Each Cluster.

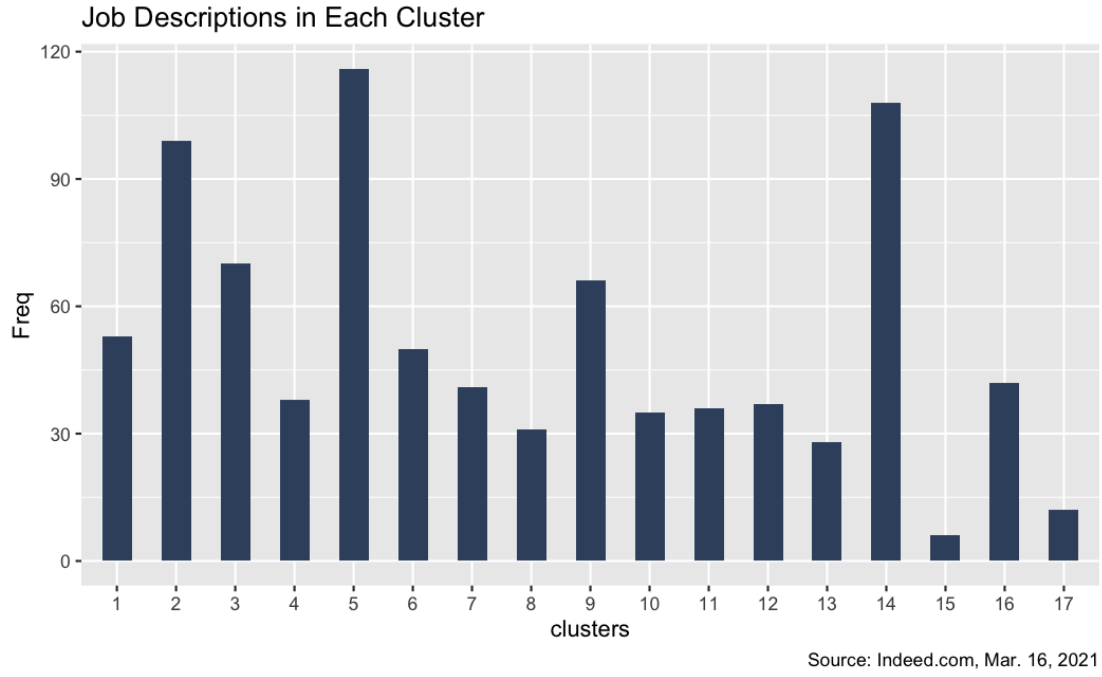


Figure 9. Top Words in Each Cluster.

cluster	size	top_words
1	53	design, experience, user, ux, product
2	99	technical, business, clinical, employment, status
3	70	job, full, years, per, hires
4	38	financial, investment, finance, management, analysis
5	116	actuarial, business, pricing, project, insurance
6	50	team, product, olo, engineering, spotify
7	41	research, ibm, information, insights, user
8	31	software, infrastructure, experience, cloud, systems
9	66	office, customer, equipment, service, maintain
10	35	hiring, recruiting, recruiter, recruitment, talent
11	36	sales, customer, training, per, looking
12	37	marketing, media, social, content, hbo
13	28	public, city, program, health, service
14	108	data, business, science, analytics, learning
15	6	mount, sinai, medicine, system, health
16	42	learning, machine, ml, deep, systems
17	12	research, investment, industry, equity, level

Figure 10. Example Output for Sales-Oriented Resume Found Online.

	similarity	recommended_position	cluster_name	good_metric	diversity_metric
160	0.132908311934115	sales	11	1	0
415	0.126747141035862	engineering+manager	3	0.940583730443776	0
270	0.124320320402496	test+engineer	1	0.917180284686814	0
171	0.110869944904669	office+manager	9	0.787469366192755	0
103	0.106139157970092	marketing	6	0.74184724072996	0