LinkedIn Profile Analysis for Career Planning Strategy

Team: Christy Chen, Xi Fan, Ran Meng, Oliver Wu, Julien Yu

Introduction and Motivation

Nowadays, it is time-consuming for a graduate to find a suitable job. According to Consumer Affairs, the average graduate needs 7.4 months to find a job [1]. Our team comes out with a potential explanation: there is an informational gap between job seeking and hiring. After we spend four years studying at university, we develop certain competitive skills. We think some of these skills are crucial for our dream job; however, companies might not necessarily want them. Worse, career fair and online platforms still need improvements in evaluating candidates and matching them with job positions. Through analyzing 3 million LinkedIn profiles, we are aspiring to address this recruitment problem and find important factors of representative jobs.

Data Preprocessing

The original dataset includes 3 million LinkedIn profiles, and we randomly choose 1 million. Each profile is a dictionary that includes 18 features such as *name, experience, education*, and *skills*. Because the data is scrapped from LinkedIn, it is very messy. We preprocess these profiles by the following 4 steps: extracting important features, selecting useful profiles, removing punctuations, and engineering new features. A clean dataframe is therefore obtained.

We first extract 4 keys: *experience, education, skills*, and *industry*, because they contain information that is most relevant with the latest occupation. Then, we remove profiles that are not in English. If a profile does not have all of these 3 key, *experience, education* or *skills*, we also remove them. As a result, we reduce the data from 1 million to 107,6000 rows.

Next, we remove punctuations, split the feature *experience* into two features: *occupation* and *company*. For the feature *education,* we split it into three: *degree, major* and *institution*. Additionally, we engineer a new feature, *years of work experience,* based on one's graduation year from college and the earliest date of work. Last but not least, we create a feature named *latest occupation* from *occupation*. The *latest occupation* is dependent variable for supervised learning, because our goal is to find important factors of jobs. The first five rows of dataframe are shown in the Appendix, as *Figure 1*.

Text Quantification: Word2Vec

After we built a clean dataframe with one text label (*latest occupation*), six text features (*skills*, *major*, *degree*, *institution*, *industry*, and *company*) and one numeric feature (*years of work experience*), we convert text into vectors by using the NLTK and Gensim library. For each text feature (e.g. *skills*), we first remove all punctuations and transvert all text to lowercase. We then use lemmatization, reducing the inflectional forms of words. Tokenization is performed on *skills*,

major, *degree* and *industry* but not on *institution* and *company*, as individual words within *institution* and *company* lose their actual meanings. (For instance, "berkeley" appears in both "uc berkeley" and "berkeley city college"). Next, we remove all stopwords in *skills*, *major*, *degree* and *industry* to avoid redundant information in our word2vec dictionaries.

After text processing, we then train our word2vec models with bigram. The benefit of bigram is that the combination of two tokens often provides more information than one token. (For instance, "international_business" contains meanings that neither "international" or "business" has). For each text feature, a unique word2vec model is trained with manually chosen *min_count* (threshold of sparse word removal) and *size* (dimensionality of word embeddings). The *min_count* for *skills*, *major*, *degree*, *institution*, *industry*, and *company* are chosen to be 100, 200, 200, 20, 50 and 20, respectively. Each model outputs a dictionary of vocabularies, and each word has a coordinate in the vector space. The resulting dictionary sizes for *skills*, *major*, *degree*, *institution*, *industry*, and *company* are 1949, 152, 38, 1456, 146 and 1354, respectively.

Feature/Label Clustering

For each text feature (e.g. *skills*), we have a dictionary of vocabularies that can be expressed in vector form. However, these vector features are uninterpretable due to high dimensionality; therefore, they are not appropriate to be predictors of supervised learning models. To address this problem, we use hierarchical clustering to assign words to an appropriate number of clusters and then use these clusters as predictors. (*major* has a dictionary size of 152 and we split it into 20 clusters. For example, a candidate has "business", "business_administration" and "computer_science" majors. If "business" and "business_administration" are in cluster 3 and "computer_science" is in cluster 1, the *major* is expressed by the following 20-dimensional vector: (1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0). An example of hierarchical clustering on *major* is shown in the Appendix, as *Figure 2*.

For each text feature, the number of clusters is manually chosen based on its dictionary size. We do not perform clustering techniques on *years of work experience* and *degree*, since *years of work experience* is a quantitative variable and *degree* only has a dictionary size of 38. Each word in *degree* (e.g. "phd", "bs", "ba", etc.) serves as a predictor. The numbers of clusters we set for *skills*, *major*, *institution*, *industry*, and *company* are 50, 20, 50, 10 and 50, respectively. An integer array of 219 features including *degree* (38-dimensional) and *years of work experience* (1-dimensional) is used for model training. There are 107.6 thousand observations.

As mentioned in Data Preprocessing section, we use *latest occupation* as our dependent variable for supervised learning. Quantification of dependent variable is more challenging than independent variables, because dependent variable should exist uniquely for each profile in

training and testing sets. Thus, we set the *min_count* of *latest occupation*'s word2vec model to be small (5). For any word that appears less than 5 times, we try to map it onto the ten most similar words using the ".most_similar" attribute of word2vec. Still, only 56 percent of the 107.6 thousand profiles can be labeled due to sparsity, though *latest occupation*'s word2vec dictionary has a reasonable length (1814). We label each of the 60,428 profiles with a unique integer from 0 to 49 and ignore the rest with sparse *latest occupation*. Each integer label represents a cluster of occupations. Our final dataset has 60,428 labeled profiles, which are used for model training, validation, and testing.

Supervised Learning: Feature Importance

Once we finished feature engineering with word2vec and clustering, we applied supervised learning models to find the important features for representative occupation clusters. For dependent variable *latest occupation*, each label is an integer ranging from 0 to 49. These 50 numbers represent a unique occupation class. For independent variables *skills* (50 columns), *major* (20 columns), *degree* (38 columns), *institution* (50 columns), *industry* (10 columns), *company* (50 columns) and *years of work experience* (1 column), the feature array has 219 columns.

There are 60,428 rows. Each row corresponds to a profile, and each column of the feature array represents the number of occurrences of a clustered feature (e.g. "business" and other related *major*). In other words, a cell $X_{i,j}$ represents the count of words in profile i that belongs to a word cluster j. For example, a cluster is ["hospital", "health", "care", "mental"], which is the 9th *industry* cluster and the 167th column of the feature array X. The feature array X consists of positive integers. If the 4949th individual has two master's degrees ("master" is the 4th *degree* element and the 74th column of X), then $X_{4949,\,74} = 2$.

Setting the train-test ratio to be 4:1, we now start supervised learning. The goal of supervised learning in our project is not to achieve high accuracies, but to gain insights about important features. In other words, we are less interested in predicting occupations for individuals but more interested in understanding the importance of each feature cluster for chosen occupations. One reason why we do not focus on high accuracies is that dependent variable has 50 classes. The baseline (training set majority) method has a typical accuracy of 4-5 percent, and our models outperform the baseline by only 2-3 percent. Indeed, we can make multiple predictions by ranking the multiclass logistic probabilities with the "predict_proba" function in Sklearn, but the data size and our computation power do not allow us to delve further into it. Details of the predictive models and accuracies are in the "Model Build-Up" section of our Jupyter notebook.

The models we explore include Decision Tree (CART), K-Nearest-Neighbors (KNN) and Random Forest (RF). To identify the important features for each occupation class, tree-based models such as CART and RF are particularly useful: the more total impurity (Gini index) decreases while splitting a feature, the more significant the feature is for that occupation class.

To see how features affect occupation, we run RF and CART on the training data with 50 distinct labels. We output a 219-dimensional vector that indicates feature importance, as illustrated in *Figure 3*. As suggested by the non-zero terms, we can observe that both RF and CART identify *institution* and *company* as insignificant and *years of work experience, skills, major*, *degree* and *industry* as relatively significant. In general, the most important features for occupations are *years of work experience*, management *skills*, technology *skills*, engineering *skills*, and BA *degree*.

Besides the 50-class classification, we also train models with binary labels ("1" if a profile falls into a chosen occupation cluster and "0" if otherwise). Take "business and finance" related *occupation* (cluster 21) as an example. We label observations in cluster 21 with the binary value "1" and all others the value "0". Then, we obtain a 219-dimensional vector that is the important features for cluster 21 only. As seen in *Figure 4*, the individual importance of nonzero terms in the binary classification model is much higher than those of the 50-class classification model. This supports our hypothesis that for a specific occupation, only a subset of features are related to this occupation, and they have greater importance than other features.

Impact

According to Vaitkus, the major problem of employment is not a skill shortage but a skill mismatch [2]. Therefore, our data-driven analysis will help students make better career planning strategy. Important features (e.g. *skills*, *major*, *degree*) will be useful for students to develop the attribute that employers care about. From a sociological perspective, resolving the "skill mismatch" problem could not only increase companies' profit margins but also enhance the efficiency of government spendings on education sector. The so-called "skill shortage" problem that bothers employers for decades would be alleviated, as "skill mismatch" is reduced.

Citation

[1] Mark Huffman. *Despite low unemployment, many college grads are out of work*. ConsumerAffairs. June 18, 2018. https://www.consumeraffairs.com/news/despite-low-unemployment-many-college-grads-are-out-of-work-061818.html

[2] Laime Vaitkus. *It's a Skills Mismatch, Not a Skills Shortage*. Bloomberg Law. October 17, 2016. https://www.bna.com/skills-mismatch-not-b57982078736/

Occupations #
https://github.com/dnordfors/archetypes; http://www.dlt.ri.gov/lmi/pdf/soc.pdf
Skills #
https://www.onetonline.org/find/descriptor/browse/Skills/
Majors #
https://en.wikipedia.org/wiki/List_of_academic_fields
Industries #
https://en.wikipedia.org/wiki/Outline_of_industry

Appendix

	skills	institution	degree	major	industry	occupation	company	year_of_work_experience
0	[DNA, Nanotechnology, Molecular Biology, Softw...	[Harvard University, Yale University]	[PHD, BS]	[Biophysics, Computer Science]	Research	[Assistant Professor, Technology Development F...	[UCSF, Wyss Institute for Biologically Inspire...	16
1	[Interactive Marketing, Content Strategy, Affi...	[University of Virginia]	[BA]	[History]	Internet	[Social Media Marketing Manager, Board of Dire...	[Coca-Cola, Atlanta Interactive Marketing Asso...	21
2	[Primavera, Revit MEP, AutoCAD, Engineering, H...	[University of Petroleum & Energy Studies, IIT...	[, , , ,]	[, , , ,]	Oil & Energy	[Manager (International Business Development &...	[VOITH Hydro Pvt Ltd., VOITH Hydro Pvt. Ltd., ...	22
3	[Talent Acquisition, Recruiting, Talent Manage...	[Universitatea „Transilvania" din Braşov, Univ...	[BA,]	[Psychology & Science of Education,]	Human Resources	[Recruitment consultant, Scientific Staffing C...	[CGI, Kelly Services, Carmeuse, Education Inst...	18
4	[Brand Management, Integrated Marketing, Telev...	[Instituto de Diseño de Caracas]	[]	[]	Broadcast Media	[Creative Services, Vice President, Creative S...	[Discovery Communications, Warner Channel, War...	28

Figure 1: Data frame after preprocessing

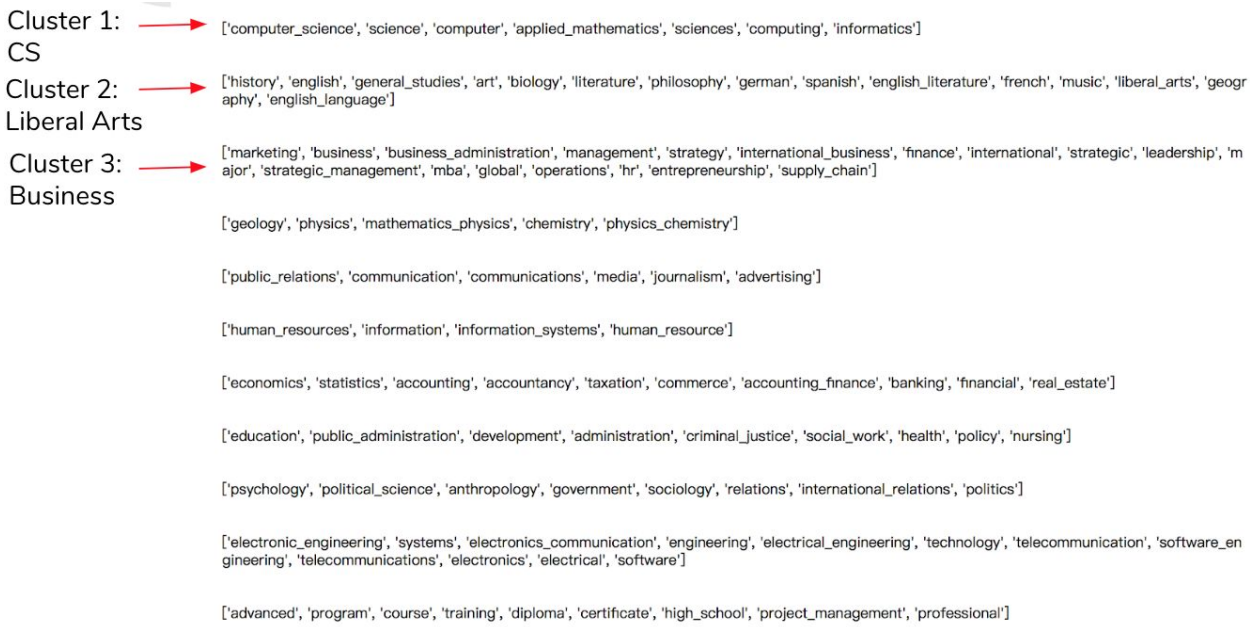

Figure 2: Example of hierarchical clustering (*major*)

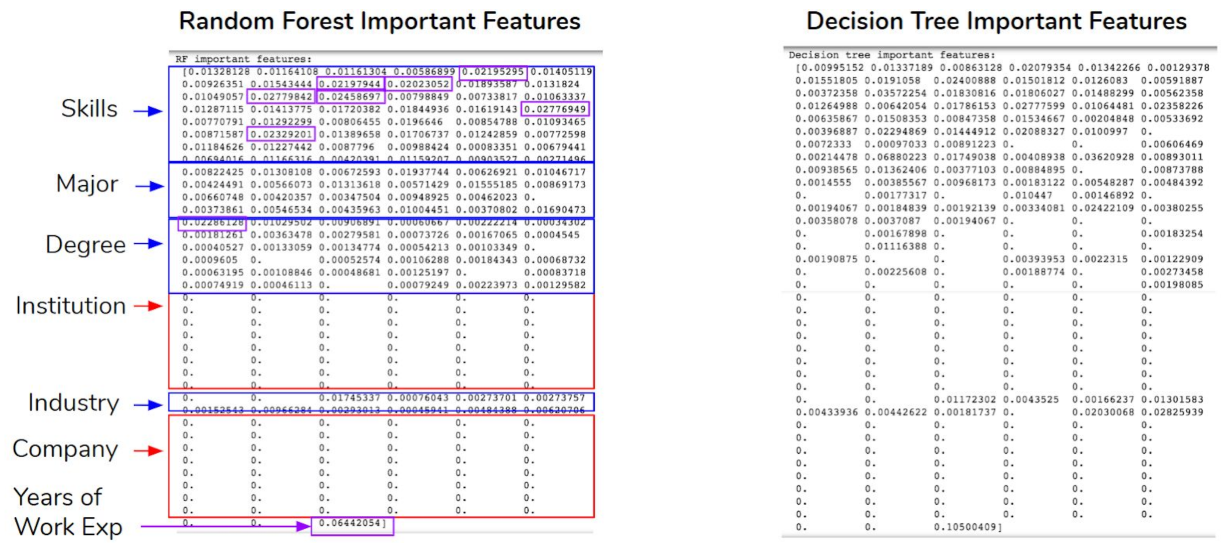

Figure 3: Weight matrix of 50-class classification

```
Decision tree test accuracy:  0.974820440848675


Decision tree confusion matrix:
 [[11808     2]
 [  303     0]]


Decision tree important features:
 [0.         0.04761248 0.         0.         0.         0.
 0.06770697 0.01584442 0.         0.07573484 0.06541598 0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.05102322 0.04134257 0.         0.08319757
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.06776805 0.03570936 0.
 0.         0.06777085 0.         0.         0.         0.
 0.         0.         0.         0.         0.05668216 0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.03360253
 0.         0.         0.         0.0693499  0.06777365 0.
 0.04038369 0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.06546927 0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
```

Figure 4: Binary classification model on *Business & Finance* (Cluster 21)