Occupation prediction based on Linkedin Data

Christy Chen
Xi Fan
Ran Meng
Oliver Wu
Julien Yu

Motivation

Problem:

- The average graduate needs 7.4 months to find a suitable job

Potential Reason:

- Informational gap between job seeking and hiring

Our Approach:

- Analyze 3 million+ LinkedIn profiles to find important factors that various jobs require. (e.g. specific major and skills)

Example Raw Profile

(Dictionary of plain text)

Value of name is a dict

Value of skills is a list

Value of industry is a string


```
{'_id': 'in-00000001',
 'name': {'family_name': 'Mazalu MBA', 'given_name': 'Dr Catalin'},
 'locality': 'United States',
 'skills': ['Key Account Development',
  'Strategic Planning',
  'Market Planning',
  'Team Leadership',
  'Negotiation',
  'Forecasting',
  'Key Account Management'],
 'industry': 'Medical Devices',
 'summary': 'SALES MANAGEMENT / BUSINESS DEVELOPMENT / PROJECT MANAGEMENTDOMESTIC & INTERNATION
AL KEY ACCOUNT MANAGEMENTBusiness and Sales Executive with 20 years of accomplished career trac
k, reflecting extensive experience and dynamic record-breaking performance in the Medical Indus
try markets. Exceptional communicator, strong team player, flexible self-starter with consultat
ive sales style, strong negotiations skills, exceptional problem solving abilities, and accurat
e customer assessment aptitude. Manage and lead teams to success, drive new business through ke
y accounts management, establish partnerships, manage solid distributor relationship for increa
sed profitability and sales volumes. Very well organized, accurate and on-time administrative w
ork, with a track record that demonstrates self-motivation, creativity, sales team leadership,
initiative to achieve corporate, team and personal goals. Experience in the following markets:
Medical Devices, Medical Disposables, Capital Equipment, Pharmaceuticals.',
 'url': 'http://www.linkedin.com/in/00000001',
 'also_view': [{'url': 'http://www.linkedin.com/pub/krisa-drost/45/909/513',
   'id': 'pub-krisa-drost-45-909-513'},
  {'url': 'http://ro.linkedin.com/pub/florin-ut/18/b33/77b',
   'id': 'pub-florin-ut-18-b33-77b'},
```


How to clean?

How to quantify?

Approach

Data Preprocessing

Word2Vec

Clustering

Modeling

Preprocessing - JSON into Data Frame

- Extract features of interest: “experience”, “education”, “skills”, and “industry”
- Remove profiles with empty “experience”, “education” or “skills” (1M to 455.8K)
- Remove profiles with non-English “industry” or more than 25% of non-English words in “experience” (455.8K to 107.6K)
- Remove punctuations in “experience”, “education” and “industry”
- Split the raw feature “experience” into “occupation” (label) and “company”
- Split the raw feature “education” into “degree”, “major” and “institution”
- Manually standardize “degree”
- Create a new column “years of work experience”
- Load the 107.6K English profiles into a **data frame**

Preprocessing - JSON into Data Frame

Out[9]:

	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 Brașov, 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

Head of the cleaned data frame
Number of profiles: 107,632
(filtered out of 1,000,000 profiles)

Text to vectors - Word2Vec

- Q: Why **Word2Vec**?
- A: Our interim goal is to classify text features. Thus, we need to quantify them as vectors before the classification.

Text to vectors - Word2Vec

We use the nltk and **gensim** libraries to convert text into vectors.

For each feature column (i.e. skills), we perform the following:

- Remove punctuations again
- Transvert to lowercase
- Lemmatize (instead of stem)
- Tokenize (optional, not applicable to “institution” & “company”)
- Remove stopwords (optional, not applicable to “institution” & “company”)
- Use Phraser & Phrases functions in gensim to construct bigram
- Build Word2Vec model (i.e. convert each tokenized word into a vector)

Hierarchical Clustering

- Q: Why **Word2Vec**?
- A: Our interim goal is to classify text features. Thus, we need to quantify them as vectors before the classification.
- Q: Then, how to classify the vectorized text features?
- A: We perform **hierarchical clustering** on each vectorized text feature (i.e. skills, major, degree, institution, industry, and company), and then regard each cluster as a **sub-feature**.

Example: We split the “major” feature into 20 clusters. A candidate has majors “business” (belongs to cluster 3), “business_administration” (belongs to cluster 3) & “computer_science” (belongs to cluster 1). Then she has a “major” feature vector:

(1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

Example of Clustering: Major

Cluster 1: CS → ['computer_science', 'science', 'computer', 'applied_mathematics', 'sciences', 'computing', 'informatics']

Cluster 2: Liberal Arts → ['history', 'english', 'general_studies', 'art', 'biology', 'literature', 'philosophy', 'german', 'spanish', 'english_literature', 'french', 'music', 'liberal_arts', 'geography', 'english_language']

Cluster 3: Business → ['marketing', 'business', 'business_administration', 'management', 'strategy', 'international_business', 'finance', 'international', 'strategic', 'leadership', 'major', 'strategic_management', 'mba', 'global', 'operations', 'hr', 'entrepreneurship', 'supply_chain']

['geology', 'physics', 'mathematics_physics', 'chemistry', 'physics_chemistry']

['public_relations', 'communication', 'communications', 'media', 'journalism', 'advertising']

['human_resources', 'information', 'information_systems', 'human_resource']

['economics', 'statistics', 'accounting', 'accountancy', 'taxation', 'commerce', 'accounting_finance', 'banking', 'financial', 'real_estate']

['education', 'public_administration', 'development', 'administration', 'criminal_justice', 'social_work', 'health', 'policy', 'nursing']

['psychology', 'political_science', 'anthropology', 'government', 'sociology', 'relations', 'international_relations', 'politics']

['electronic_engineering', 'systems', 'electronics_communication', 'engineering', 'electrical_engineering', 'technology', 'telecommunication', 'software_engineering', 'telecommunications', 'electronics', 'electrical', 'software']

['advanced', 'program', 'course', 'training', 'diploma', 'certificate', 'high_school', 'project_management', 'professional']

Modeling & Prediction: (Latest) Occupation

```
Number of rows and columns for each array:
skills: 107632 ,  50
major: 107632 ,  20
degree: 107632 ,  38
institution: 107632 ,  50
industry: 107632 ,  10
company: 107632 ,  50
year_of_work_experience: 107632 ,  1
```

```
Sample feature row for one profile:
[[ 0.  0.  1.  0.  0.  0.  0.  0.  0.  1.  0.  1.  0.  0.  0.  0.  0.  0.
   0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1.  0.  0.  0.
   0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  2.  0.  0.  0.
   0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1.  1.
   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.  0.  0.  0.  0.  0.  0.  0.  0.  0.
   0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1.  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. 16.]]

Feature array dimensions:  (107632, 219)
```

- Labeling Y:
  Latest **Occupation**
  Hierarchical Clustering
  (0-49)
  Multiclass Classification
- Train-Test Split
  (4:1)

Built models:
- Baseline Model
- Decision Tree Classification
- Ridge Classification
- KNN Classification
- Random Forest Classification

Future work:
(need more computing power)
- SVM Classification
- One-vs-Rest Classification
- Multiclass Logistic Classification

- **Rank the Important Features for hiring and for specific job types**
- **Give Predictions for someone's occupation based on his/her past**

Model Accuracy

Model	Accuracy
Baseline (majority)	4.562%
CART (cross-validated max depth)	5.958%
RF (cross-validated n_estimators)	5.010%
KNN (cross-validated k)	6.481%

* Baseline predicts label with the most training occurrences.
* Test accuracy is the success rate of predicting 1 class out of 50 in 1 prediction.

Important Features for Hiring

We are more interested in evaluating the **most important features** for hiring.

Random Forest Important Features

Skills
Major
Degree
Institution
Industry
Company
Years of Work Exp

```
RF important features:
 [0.01328128 0.01164108 0.01161304 0.00586899 0.02195295 0.01405119
 0.00926351 0.01543444 0.02197944 0.02023052 0.01893587 0.0131824
 0.01049057 0.02779842 0.02458697 0.00798849 0.00733817 0.01063337
 0.01287115 0.01413775 0.01720382 0.01844936 0.01619143 0.02776949
 0.00770791 0.01292299 0.00806455 0.0196646  0.00854788 0.01093465
 0.00871587 0.02329201 0.01389658 0.01706737 0.01242859 0.00772598
 0.01184626 0.01227442 0.0087796  0.00988424 0.00083351 0.00679441
 0.00694016 0.01166316 0.00420391 0.01159207 0.00903527 0.00271496
 0.00822425 0.01308108 0.00672593 0.01937744 0.00626921 0.01046717
 0.00424491 0.00566073 0.01313618 0.00571429 0.01555185 0.00869173
 0.00660748 0.00420357 0.00347504 0.00948925 0.00462023 0.
 0.00373861 0.00546534 0.00435963 0.01004451 0.00370802 0.01690473
 0.02286128 0.01029502 0.00906891 0.00060667 0.00222214 0.00034302
 0.00181261 0.00363478 0.00279581 0.00073726 0.00167065 0.0004545
 0.00040527 0.00133059 0.00134774 0.00054213 0.00103349 0.
 0.0009605  0.         0.00052574 0.00106288 0.00184343 0.00068732
 0.00063195 0.00108846 0.00048681 0.00125197 0.         0.00083718
 0.00074919 0.00046113 0.         0.00079249 0.00223973 0.00129582
 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.         0.01745337 0.00076043 0.00273701 0.00273757
 0.00152543 0.00966284 0.00293013 0.00045941 0.00484388 0.00620706
 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.         0.06442054]
```

Decision Tree Important Features

```
Decision tree important features:
 [0.00995152 0.01337189 0.00863128 0.02079354 0.01342266 0.00129378
 0.01551805 0.0191058  0.02400888 0.01501812 0.0126083  0.00591887
 0.00372358 0.03572254 0.01830816 0.01806027 0.01488299 0.00562358
 0.01264988 0.00642054 0.01786153 0.02777599 0.01064481 0.02358226
 0.00635867 0.01508353 0.00847358 0.01534667 0.00204848 0.00533692
 0.00396887 0.02294869 0.01444912 0.02088327 0.0100997  0.
 0.0072333  0.00097033 0.00891223 0.         0.         0.00606469
 0.00214478 0.06880223 0.01749038 0.00408938 0.03620928 0.00893011
 0.00938565 0.01362406 0.00377103 0.00884895 0.         0.00873788
 0.0014555  0.00385567 0.00968173 0.00183122 0.00548287 0.00484392
 0.         0.00177317 0.         0.010447   0.00146892 0.
 0.00194067 0.00184839 0.00192139 0.00334081 0.02422109 0.00380255
 0.00358078 0.0037087  0.00194067 0.         0.         0.
 0.         0.00167898 0.         0.         0.         0.00183254
 0.         0.01116388 0.         0.         0.         0.
 0.00190875 0.         0.         0.00393953 0.0022315  0.00122909
 0.         0.00225608 0.         0.00188774 0.         0.00273458
 0.         0.         0.         0.         0.         0.00198085
 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.         0.01172302 0.0043525  0.00166237 0.01301583
 0.00433936 0.00442622 0.00181737 0.         0.02030068 0.02825939
 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.         0.10500409]
```

Important Features for a Specific Occupation Category

```
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.
```

Binary (finance & management)

```
Decision tree important features:
 [0.00995152 0.01337189 0.00863128 0.02079354 0.01342266 0.00129378
 0.01551805 0.0191058  0.02400888 0.01501812 0.0126083  0.00591887
 0.00372358 0.03572254 0.01830816 0.01806027 0.01488299 0.00562358
 0.01264988 0.00642054 0.01786153 0.02777599 0.01064481 0.02358226
 0.00635867 0.01508353 0.00847358 0.01534667 0.00204848 0.00533692
 0.00396887 0.02294869 0.01444912 0.02088327 0.0100997  0.
 0.0072333  0.00097033 0.00891223 0.         0.         0.00606469
 0.00214478 0.06880223 0.01749038 0.00408938 0.03620928 0.00893011
 0.00938565 0.01362406 0.00377103 0.00884895 0.         0.00873788
 0.0014555  0.00385567 0.00968173 0.00183122 0.00548287 0.00484392
 0.         0.00177317 0.         0.010447   0.00146892 0.
 0.00194067 0.00184839 0.00192139 0.00334081 0.02422109 0.00380255
 0.00358078 0.0037087  0.00194067 0.         0.         0.
 0.         0.00167898 0.         0.         0.         0.00183254
 0.         0.01116388 0.         0.         0.         0.
 0.00190875 0.         0.         0.00393953 0.0022315  0.00122909
 0.         0.00225608 0.         0.00188774 0.         0.00273458
 0.         0.         0.         0.         0.         0.00198085
 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.         0.01172302 0.0043525  0.00166237 0.01301583
 0.00433936 0.00442622 0.00181737 0.         0.02030068 0.02825939
 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.         0.10500409]
```

Multi- Class (50 classes)

Impact

- Help fresh grads find suitable jobs
- Improve matching between job candidates and jobs
- Add productivity to companies & provided help to students' decision makings

Appendix

Word2Vec & Clustering Statistics

Features	min_count	# of Word2Vec dict keys	# of clusters (k)
skill	100	1949	50
major	200	152	20
degree	200	38	38 (no clustering)
institution	20	1456	50
industry	50	146	10
company	20	1354	50

Label	min_count	# of Word2Vec dict keys	# of clusters (k)
(latest) occupation	2*	6391	50

* Note: If an occupation appears only once, it is “projected” into its most similar occupation and assigned the label.

Model Accuracy

Achieving high prediction accuracy is **not** the primary goal of our models.

- We currently assign labels into 50 different classes (and possibly more in the future). Thus, the "real" baseline prediction accuracy = 0.02
- A good prediction model returns the 5 (or 10) most likely occupation predictions (i.e. with "**predict_proba**" function in sklearn). Will do in the future.

Test accuracies (with 1 prediction, keep in mind there are 50 classes):

- Baseline (predict label with the most training occurrences): 4.562%
- Decision Tree (tuned max_depth): 5.958%
- Ridge (tuned alpha): **6.475%**
- Random Forest (tuned n-estimators): 5.010%
- K-Nearest Neighbors (tuned k): **6.481%**

Model Accuracy: Tuning Example

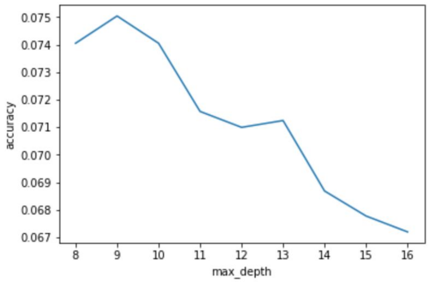
For the decision tree model, we tune the hyperparameter "max_depth".

In [43]:
```
from sklearn.tree import DecisionTreeClassifier
import matplotlib.pyplot as plt

# Decision tree classification (tune max_depth)
max_depth_list = [8,9,10,11,12,13,14,15,16]
dtree_accuracy_temp = []
for max_depth in max_depth_list:
    dtree_model_temp = DecisionTreeClassifier(max_depth = max_depth).fit(X_train, y_train)
    dtree_accuracy_temp.append(dtree_model_temp.score(X_test, y_test))

plt.plot(max_depth_list, dtree_accuracy_temp)
plt.ylabel('accuracy')
plt.xlabel('max_depth')
```

Out[43]:
```
Text(0.5, 0, 'max_depth')
```



Important Features for Hiring: Results

We are more interested in evaluating the **most important features** for hiring.

Significant Features (Overall)

- Years of work experience
- Skills
- Major
- Degree
- Industry

Insignificant Features

- Company
- Institution

Most Significant Features (Specified)

- Years of work experience
- Skill cluster: management, finance
- Skill cluster: technology
- Skill cluster: development
- Skill cluster: engineering
- Degree: BA

* Rank from Random Forest