Project 4

Can we predict salary and job title of a position?

Angie Sheng

26/08/2019

1. Raw Data

Seek.com.au

- Data Scientist
- Data Analyst
- Data Engineer
- BI Analyst

Last scraped on 21/08/2019

2. Data Cleaning

Columns =['Salary Range', 'Link', 'Job Title', 'Job Teaser', 'Advertiser', 'Classification', 'Location', 'Strong Words', 'Job Description', 'Category']

Salary Range =['0-70k','70k-120k','over 120k']

A Glimpse at the Data Set

	Salary Range	Link	Job Title	Job Teaser	Advertiser	Classification	Location	Strong Words	Job Description	Category
1071	120000-999999	https://www.seek.com.au/job/39576650	Lead Business Intelligence Analyst	Make your mark working with the industry leade...	McMillan Shakespeare	Information & Communication Technology	Melbourne	To succeed as a Lead Business Intelligence Ana...	As a result of growth within the business we a...	BI Analyst
1072	120000-999999	https://www.seek.com.au/job/39597045	Data Analyst - Business Intelligence / Data Wa...	Data Analyst - Business Intelligence / Data Wa...	Infinity Pro	Information & Communication Technology	Toowoomba & Darling Downs	Your Benefits: your CV will need to reflect on...	Your Benefits: Immediate Start; Great Rates Po...	BI Analyst
1073	120000-999999	https://www.seek.com.au/job/39579332	BI / Data Warehouse Analyst Programmer	Join this Government organisation leading the ...	Eden Ritchie Recruitment	Information & Communication Technology	Brisbane	Business Intelligence/Data Warehouse Analyst P...	CBD Location Initial 6 month contract 95–1...	BI Analyst
1074	120000-999999	https://www.seek.com.au/job/39588897	Business Intelligence DW Analyst Programmer	Great contract for a large government departme	Finite IT Recruitment Solutions	Information & Communication Technology	Brisbane	The following work will be involved for the po...	Our client is a large government department	BI Analyst

Question 1 What factors decided Salary

```
#Get TDIDF scores.. convert all to lowercase

vect = TfidfVectorizer(stop_words='english',
                        lowercase=True, preprocessor= None,
                        analyzer= 'word', token_pattern='(?u)\\b\\w\\w+\\b',
                        ngram_range=(1, 5), max_df=10.0, min_df=1,
                        max_features=180,
                        use_idf=True,
                        smooth_idf=True,
                        tokenizer= tokenizer.tokenize,
                        sublinear_tf=False)

vect.fit(jobs['Job Description'].apply(str))

vect.get_feature_names()

#I will use the document term matrix(DTM) below as a predictor attribute for both question one and two.  I decided to set
#number of features to 180

dtm = pd.DataFrame(vect.transform(jobs['Job Description'].apply(str)).todense(),
                        columns=vect.get_feature_names())
```

Get TD-IDF Score

Convert it to Document Term Matrix(DTM)

Question 1

```
from sklearn.ensemble import RandomForestClassifier


#X contains dummy variables for every title, classification,location of job plus the DTM with 180 features.
#The total predictor is over 700 variables.

X = pd.concat([X1_title, X2_classification, X3_location, dtm] ,axis=1, join_axes=[X1_title.index])

#The dependent variable is binary. 1 for high salary and 0 for low salary jobs.
y = jobs['salary_bin'].astype(float)

#test train split. I follow this format for splitting the data set for most
#of the analysis below other than the logistic regression model where I train the model of a five
#fold cross val again with all the same predictors
X_train, X_test, y_train, y_test = train_test_split(X, y,random_state=1)



rf = RandomForestCla
rf.fit(X_train, y_tr

rf.predict(X_train)
rf.score(X_test,y_te

print(rf.score(X_tes
```

```
from sklearn.tree import DecisionTreeClassifier



dt = DecisionTreeClassifier()
dt.fit(X_train, y_train)
importances = pd.DataFrame(zip(dt.feature_importances_,rf.feature_importances_),
                           index=X.columns, columns=['dt_importance','rf_importance']).sort_values('rf_importance',ascending=False)
dt.predict(X_test)
dt.score(X_test, y_test,sample_weight=None)

print(dt.score(X_test, y_test,sample_weight=None))
importances.head(50)
```

DTM

Dummy Variables:

Job Title (Data Scientist, Data Analyst…)

Classification(IT, HealthCare…)

Location(Sydney, Melbourne…)

Classifier

Random Forest Classifier
Decision Tree Classifier

Question 1

Accuracy: around 50%-60%

Title*

	dt_importance	rf_importance
Data Analyst	0.637364	0.417808
Data Scientist	0.352460	0.394401
BI Analyst	0.010176	0.094856
Data Engineer	0.000000	0.092935

* May due to unbalanced data set

Location*

	dt_importance	rf_importance
Darwin	0.336214	0.202780
Sunshine Coast	0.127352	0.087738
Melbourne	0.119355	0.081670
Brisbane	0.002257	0.067273
Wollongong, Illawarra & South Coast	0.057746	0.063555
Perth	0.000000	0.061882
ACT	0.010764	0.061597
Northern QLD	0.116302	0.060016
Newcastle, Maitland & Hunter	0.025553	0.059247
Gold Coast	0.024337	0.056116
South West Coast VIC	0.056135	0.053710
Sydney	0.054717	0.047117
Adelaide	0.018264	0.044749

Classification

	dt_importance	rf_importance
Information & Communication Technology	0.244618	0.145968
Accounting	0.055760	0.081041
Banking & Financial Services	0.090842	0.076374
Marketing & Communications	0.053327	0.074044
Government & Defence	0.009781	0.060986
Engineering	0.081781	0.056384
Science & Technology	0.073080	0.055894
Sport & Recreation	0.066496	0.055860
Administration & Office Support	0.068294	0.053224
Healthcare & Medical	0.032597	0.052492
Insurance & Superannuation	0.054138	0.048833
CEO & General Management	0.079186	0.045329
Mining, Resources & Energy	0.000000	0.044624
Education & Training	0.033478	0.039851
Consulting & Strategy	0.035274	0.031399

Question 1

Job Description (DTM)

	dt_importance	rf_importance
experience	0.015822	0.013853
team	0.016889	0.012314
data	0.011268	0.012136
skills	0.000000	0.011013
business	0.000000	0.010892
contract	0.030723	0.010539
role	0.014452	0.010458
work	0.001101	0.010454
analysis	0.019407	0.009925
senior	0.025231	0.009732
working	0.022916	0.009536
support	0.000000	0.009516
data science	0.039160	0.009172
apply	0.000000	0.009118
key	0.006607	0.008734

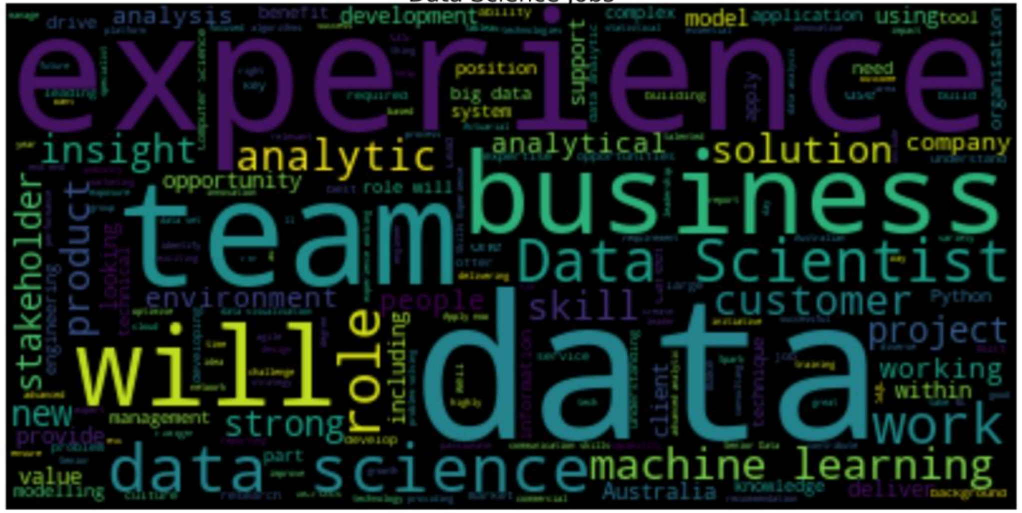
Experience

Team Work

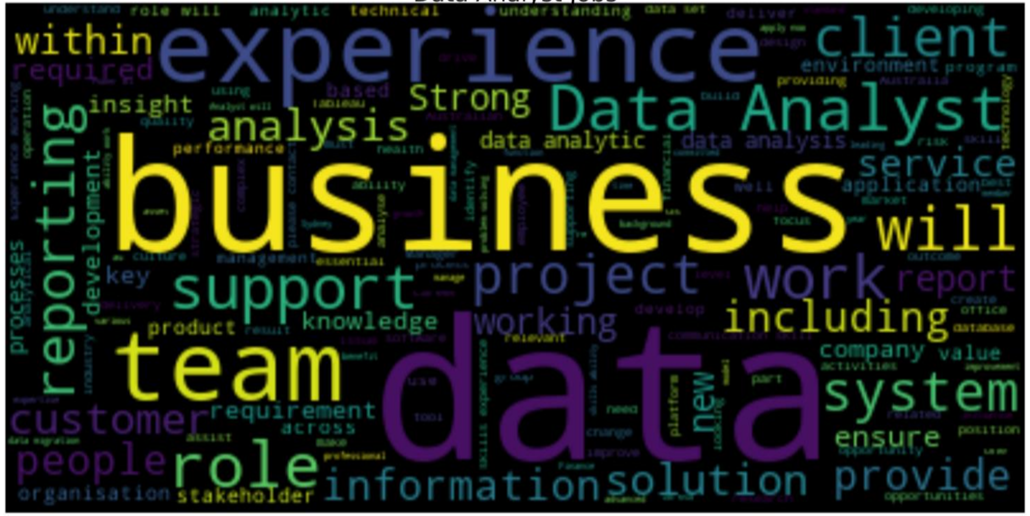
Data manipulation

Question 2 what distinguish different job classifications? ----- 1) Word Cloud

Data Science Jobs

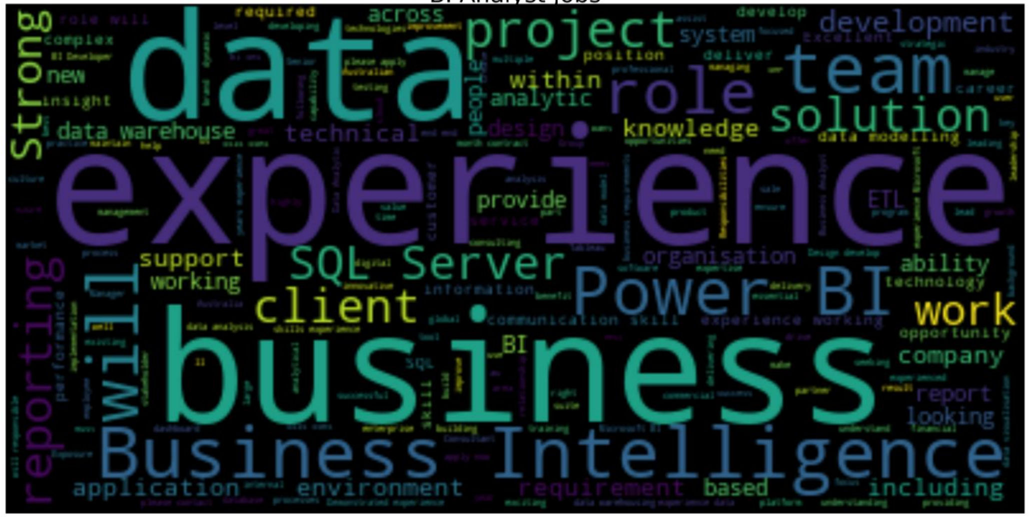


Data Analyst Jobs



Data Engineer Jobs

BI Analyst Jobs



Question 2 what distinguish different job classifications? ------ 2) Word2Vec

```
# sg - skip gram |  window = size of the window | size = vector dimension
size = 400 #It could be smaller but I would tend higher with this model than the size of the TFIDF features
window_size = 6 # sentences weren't too long
epochs = 50
min_count = 5
workers = 4

# train word2vec model using gensim
model = Word2Vec(corpus, alpha=0.01,sg=1,window=window_size,size=size,\
                              min_count=min_count,workers=workers,iter=epochs,batch_words=1,negative=25,seed=100)

model.build_vocab(sentences=corpus,update=True)


model.train(sentences=corpus,epochs=50,total_examples=model.corpus_count)

model.save('w2v_bftest')

model = Word2Vec.load('w2v_bftest')

w2v = dict(zip(model.wv.index2word, model.wv.syn0))
```

Question 2 what distinguish different job classifications? ------ 2) Word2Vec

```
1  model.wv.most_similar(positive=['data' ,'scientist'],
```

```
[('cleansing,', 0.46503371000289917),
 ('-Experience', 0.44138303399086),
 ('discovery', 0.43523725867271423),
 ('collection,', 0.43272876739501953),
 ('warehouses,', 0.431367218494153),
 ('economics', 0.4286949634552002),
 ('patterns,', 0.427321195602417),
 ('Assemble', 0.42301872372626726),
 ('junior', 0.4221442341804504 4),
 ('turning', 0.42023640871047974)]
```

```
1  model.wv.most_similar(positive=['data' ,'analyst'],
```

```
[('-Experience', 0.504447877407074),
 ('data-related', 0.4661821722984314),
 ('Owner', 0.4444394409656524 7),
 ('developers,', 0.4438074827194214),
 ('convergence', 0.4404323101043701),
 ('staging,', 0.43646693229675293),
 ('defining', 0.4364025294780731),
 ('BA', 0.4351705312728882),
 ('cleansing,', 0.4298211336135864),
 ('feasible', 0.42217501997947693)]
```

Question 2 what distinguish different job classifications? ------ 2) Word2Vec

```
1 model.wv.most_similar(positive=['data' ,'engineer'],
```

```
[('ingesting', 0.4779267013072967 5),
 ('lakes', 0.476145535707473 75),
 ('warehouses,', 0.47126805782318115),
 ('optimizing', 0.4665679931640625),
 ('Assemble', 0.45403429865837097),
 ('script', 0.4496055841445923),
 ('-Experience', 0.44864872097969055),
 ('Seeking', 0.44813841581344604),
 ('lake', 0.4401501715183258),
 ('cloud-based', 0.43680235743522644)]
```

```
1 model.wv.most_similar(positive=['BI' ,'intelligence']
```

```
[('Power', 0.5266002416610718),
 ('suite,', 0.4839787483215332),
 ('BI)', 0.4639700651168823),
 ('Pivot,', 0.429243803024292),
 ('MSBI', 0.4163907766342163),
 ('(E', 0.4100627601146698),
 ('business', 0.40011683106422424),
 ('Server', 0.3976868391036987 3),
 ('warehouses,', 0.3961943984031677),
 ('SSRS,', 0.3934440910816192 6)]
```

Question 2 what distinguish different job classifications? ---- 3) Random Forest & Decision Tree

```
#below are the various Random forests models the word embedding models were also used
#against just one dummy which was the data science job title dummy.

#Decision trees and random forests for the Data Scientist dummy

jobs['title_ds'] = np.where(jobs['Category'].str.contains("Data Scientist"), 1, 0).astype(float)
jobs['title_da'] = np.where(jobs['Category'].str.contains("Data Analyst"), 1, 0).astype(float)
jobs['title_de'] = np.where(jobs['Categor
jobs['title_bi'] = np.where(jobs['Categor
```

```
rf = RandomForestClassifier(n_estimators=50, max_features= 7, n_jobs=-1,class_weight='balanced', oob_score=True)
rf.fit(X_train, y_train)

rf.predict(X_train)
rf.score(X_test,y_test, sample_weight=None)

print(rf.score(X_test,y_test, sample_weight=None))

importances = rf.feature_importances_

from sklearn.tree import DecisionTreeClassifier


dt = DecisionTreeClassifier(class_weight='balanced')
dt.fit(X_train, y_train)
importances = pd.DataFrame(zip(dt.feature_importances_,rf.feature_importances_,),
                           index=dtm.columns, columns=['dt_importance','rf_importance']).sort_values('rf_importance',ascending=False)
dt.predict(X_test)
dt.score(X_test, y_test,sample_weight=None)

print(dt.score(X_test, y_test,sample_weight=None))
importances.head(20)
```

TFIDF Score

Document Term Matrix(DTM)

Question 2 what distinguish different job classifications? ---- 3) Random Forest & Decision Tree

Data Scientist

	dt_importance	rf_importance
science	0.179325	0.070131
python	0.062679	0.064421
learning	0.000000	0.061016
machine learning	0.372433	0.043068
data science	0.000000	0.036859
machine	0.000000	0.036568
bi	0.000000	0.035654
analyst	0.000000	0.025415
etl	0.016208	0.015871
engineer	0.000000	0.013474
power	0.000000	0.012072
design	0.000000	0.011979
data analyst	0.000000	0.010054
australia	0.059587	0.009463
data engineer	0.066143	0.009345
processes	0.000000	0.008772
ssis	0.000000	0.008729

Data Scientist and Data Analyst has different requirement towards python

Data Analyst

	dt_importance	rf_importance
analyst	0.000000	0.081386
data analyst	0.434660	0.063734
analysis	0.088193	0.035460
bi	0.042385	0.031847
big data	0.035591	0.021012
data analysis	0.000000	0.018805
data engineer	0.004935	0.015863
engineer	0.000000	0.015292
developer	0.000000	0.014247
business intelligence	0.000000	0.013938
intelligence	0.000000	0.012256
ssis	0.000000	0.012126
machine learning	0.000000	0.011995
experience	0.000000	0.011769
power bi	0.000000	0.011075
agile	0.000000	0.010862
python	0.000000	0.010858

Question 2 what distinguish different job classifications? ---- 3) Random Forest & Decision Tree

	dt_importance	rf_importance
engineer	0.000000	0.103651
data engineer	0.576597	0.073004
big data	0.000000	0.030943
big	0.000000	0.027481
technologies	0.133238	0.027452
cloud	0.017461	0.025944
analyst	0.000000	0.024649
data	0.042703	0.024591
data analyst	0.000000	0.019686
analysis	0.000000	0.019554
platform	0.060169	0.018117
reporting	0.036610	0.018090
business	0.004467	0.018036
azure	0.000000	0.016703
aws	0.000000	0.016375
bi	0.000000	0.014046
experience	0.000000	0.013243

Data Engineer

BI Analyst

Requirements on different skill sets

	dt_importance	rf_importance
bi	0.548854	0.091455
data analyst	0.070546	0.042312
data	0.040310	0.041490
ssis	0.000000	0.030524
business intelligence	0.104812	0.029546
developer	0.000000	0.029065
power bi	0.000000	0.027582
python	0.000000	0.021714
power	0.000000	0.021245
intelligence	0.000000	0.021127
sql server	0.000000	0.018889
data engineer	0.052440	0.014328
science	0.015184	0.014024
engineer	0.000000	0.013474
reports	0.000000	0.012936
engineering	0.000000	0.012639

Question 2 what distinguish junior and senior jobs?

Hard skills

```
1  model.wv.most_similar(('junior'),topn=15)
```

```
[('developers,', 0.47188881039619446),
 ('Mentor', 0.4698949456214905),
 ('analysts', 0.4618791341781616),
 ('scientist', 0.4532897174358368),
 ('engineers,', 0.444827139377594),
 ('mentored', 0.4279401600360870 4),
 ('graduate', 0.4171069860458374),
 ('6-month', 0.4057192206382751 5),
 ('intelligent', 0.4032573997974396),
 ('feasible', 0.4026352167129516 6),
 ('full-stack', 0.40107619762420654),
 ('fill', 0.394310861825943),
 ('analysts,', 0.39385557174682617),
 ('hunt', 0.3905479311943054),
 ('Hydrogen', 0.3891531229019165)]
```

Soft skills

```
1  model.wv.most_similar(('senior'),topn=15)
```

```
[('executives', 0.4336004257202148 4),
 ('makers', 0.3778500556945801),
 ('stakeholders;', 0.3667801320552826),
 ('mid', 0.3590776026248932),
 ('advisor', 0.3565120398998260 5),
 ('executive', 0.3523682355880737 3),
 ('That', 0.3428587913513183 6),
 ('influence', 0.341877818107605),
 ('presents', 0.3396634459495544 4),
 ('advisors', 0.3356225490570068 4),
 ('non', 0.3338938653469085 7),
 ('engineer,', 0.329532146453857 4),
 ('Kubernetes', 0.3277855813503265 4),
 ('Executive', 0.322480499744415 3),
 ('analysts,', 0.3202930688858032)]
```

Insights

1. Experience is the most important part that employee will look at

2. If you would like to be promoted into a senior role, besides python, SQL Tableau…Soft skills like stakeholder management, power of influencing others, team working… will be a must