Seek Webscraping & NLP

Project 4 Submission- DSI Immersive – Pramod Paul

27/5/2019

Goal/Objective

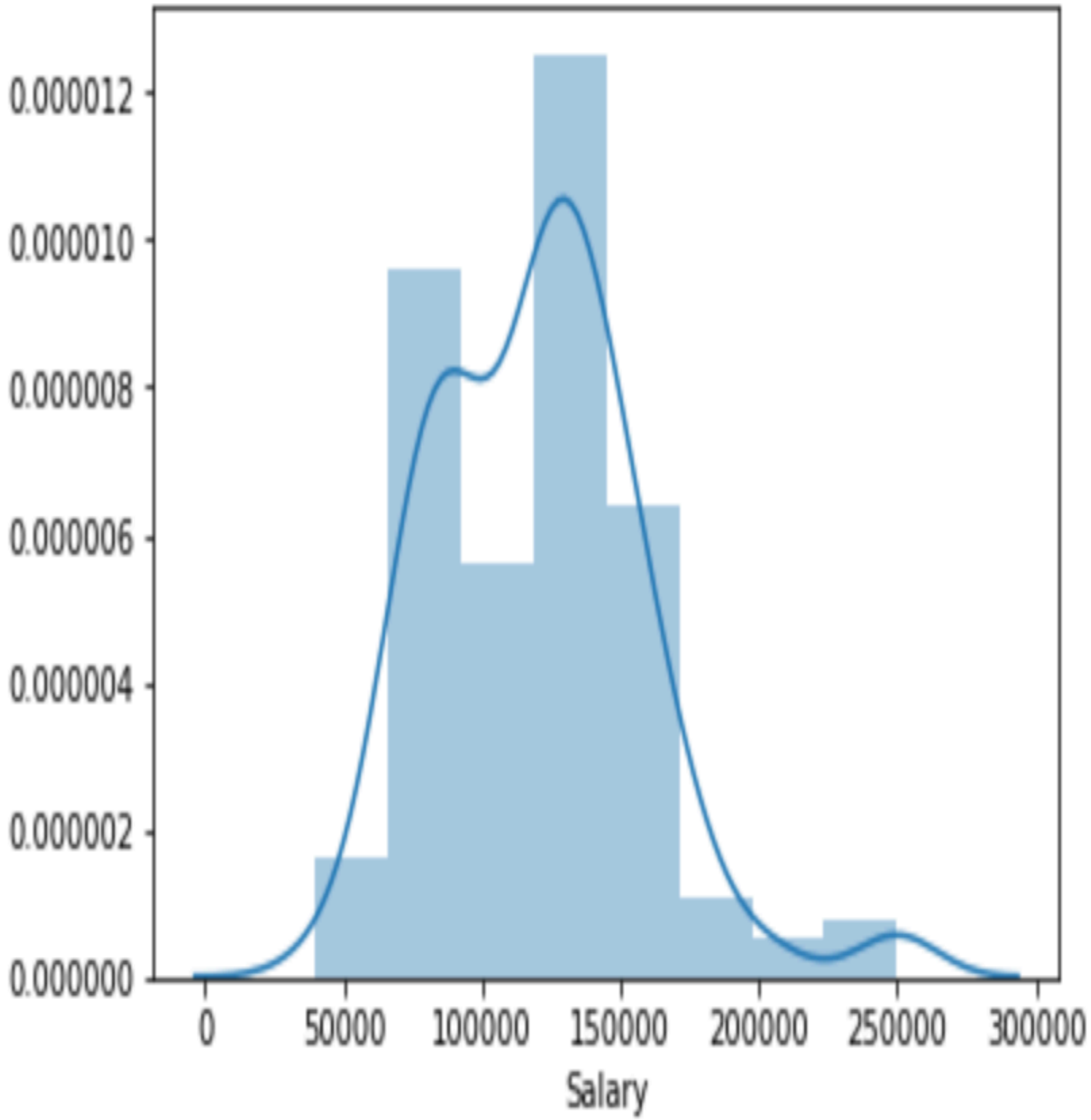
1)Create Data set:

Webscrape Data related job postings from Seek

2) Factors that impact Job salary

3) Factors that distinguish Job category

What factors could be impacting this wide range of salaries ?



1) Models: Tf-idf & elastic net

2) X = Job title, Y= Salary

3) Result is features showing the word impact and the co-efficients

Question I:

Factors that impact salary

Job title

(unfiltered text from the posting)

Feature names and their importance from Raw Job title string

Out[49]:

	word	elasticnet_coef
52	data	967.045750
29	business	966.454170
98	lead	802.100067
168	senior	716.314833
166	scientist	678.303401
56	development	470.400432
107	manager	417.289391
68	finance	295.017453
87	insights	262.293259
25	bi	253.465225
26	big	251.696488
189	technical	241.115718
143	privacy	223.680671
37	circa	218.557158
15	analyst	214.910118
24	bank	203.386437
133	partner	203.386435
2	250k	203.386423
157	reporting	175.507234
173	solutions	175.014943
148	protection	169.242296
62	engineer	159.362408

Feature names and their importance on salary from processed Job title string

Classified as Analyst, Scientist, Manager

Out[54]:

	word	elasticnet_coef
3	manager	602.923884
1	data	558.403912
5	scientist	352.314777
0	analyst	125.508429
2	info	-863.899158
4	no	-863.899168

Feature names and their importance from raw Job description from Seek

Out[59]:

	word	elasticnet_coef
2239	li	951.847852
525	business	687.003530
1012	data	532.021549
481	br	457.422712
3700	strong	387.642775
1550	financial	384.015475
4023	understanding	286.376259
973	csiro	283.699676
2326	machine	252.983892
3473	senior	251.124303
2213	learning	245.361759
3275	requirements	236.592037
2203	lead	234.907061
3729	success	230.334336
1369	enterprise	223.543700
3430	science	222.737059
2188	large	212.197990
2488	modelling	181.776713
2351	managing	178.951105
2789	partners	176.454826
626	change	170.080329

3646	stakeholder	160.850087
2489	models	159.507525
2651	on	158.758976
750	commercial	158.456679
437	bi	154.290781
3790	tableau	153.737203
3691	strategy	152.858855
3568	solutions	152.411463
518	building	151.945551
3823	technical	149.740819
556	capability	148.134738
3647	stakeholders	147.044865
2933	predictive	146.706178
46	across	146.036622
3331	revenue	143.740895
189	analytics	143.046993
3082	python	141.265497

Question 2:

Factors that distinguish each job category

1) Models: Tf-idf & others

2) X = Job Description

Y= Filtered Job Title

3) Result is features & accuracy

Accuracy:

Naïve Bayes**:** 0.535

Bernoulli Naive Bayes: 0.512

RandomForestClassifier: 0.535

SGD with Elastic-Net penalty: 0.512

LinearSVC: 0.535

Using word2vec on job description (CBOW)

```
In [236]: sim_words_analyst = model.wv.most_similar('analyst')
```

```
In [237]: sim_words_analyst
```

```
Out[237]: [('marketing', 0.9680184125900269),
 ('responsible', 0.954077422618866),
 ('experienced', 0.9521729946136475),
 ('salesforce', 0.9514082670211792),
 ('digital', 0.9474241733551025),
 ('lead', 0.9406147003173828),
 ('manager', 0.9394255876541138),
 ('cyber', 0.9337409138679504),
 ('team', 0.9316191673278809),
 ('based', 0.9229862689971924)]
```

```
In [238]: sim_words_scientist = model.wv.most_similar('scientist')
```

```
In [239]: sim_words_scientist
```

```
Out[239]: [('consultancy', 0.9959532022476196),
 ('accountant', 0.9951343536376953),
 ('currently', 0.9949975609779358),
 ('recruiting', 0.9945607781410217),
 ('manufacturing', 0.9943975210189819),
 ('registered', 0.9943151473999023),
 ('administrator', 0.9937311410903931),
 ('unique', 0.9935327768325806),
 ('contracts', 0.9927530288696289),
 ('growth', 0.9911949038505554)]
```

Using word2vec (CBOW)

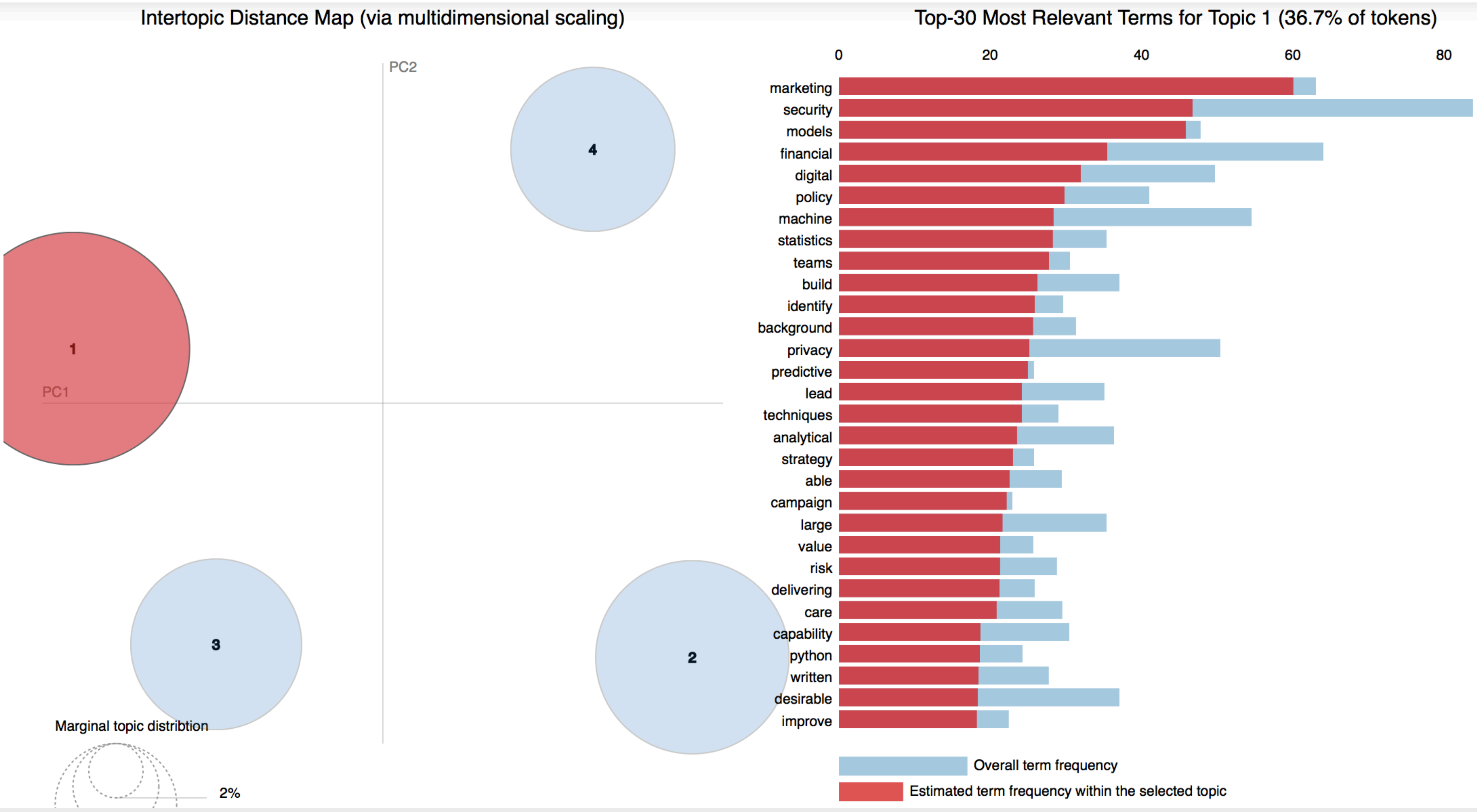
In [240]:
```
sim_words_manager = model.wv.most_similar('manager')
sim_words_manager
```

Out[240]:
```
[('digital', 0.9876937866210938),
 ('responsible', 0.9862974286079407),
 ('lead', 0.9824036359786987),
 ('marketing', 0.981643557548523),
 ('finance', 0.9794235229492188),
 ('partnering', 0.9793709516525269),
 ('journey', 0.9791109561920166),
 ('understand', 0.9790796041488647),
 ('team', 0.9787262678146362),
 ('deliver', 0.9778867959976196)]
```

Topic extraction using LDA

```
Fitting LDA models with tf features, n_samples=2000 and n_features=1000...
done in 0.832s.

Topics in LDA model:
Topic #0: li spatial ul marketing maritime reporting analysis management role questions project safety resume key wat
erway officer policy campaign processing information
Topic #1: strong li ul role experience payroll environment provided working team href providing reputation produce bu
siness benefits right market commitment salary
Topic #2: li xa br strong business ul security financial amp experience years contact href development enterprise man
aging skills key required management
Topic #3: li br strong ul analytics xa business skills insights experience advanced models collections role learning
work ability trends teams value
Topic #4: area bring environments like quantitative interpersonal developers sector undertaking non agile residential
operating required engineer software mailto studio undergraduate plans
Topic #5: li strong security ability apply ul zkuacf vifm technology description biology write evidence environment w
eb expert skills computational victorian understanding
Topic #6: xa governance profile questions lecturer applicant hoc customer responsible private melbourne unstructured
accounting check ca conferences bluefinresources submitting agencies currently
Topic #7: li strong xa br ul experience research apply href team csiro business work role target blank working scienc
e skills software
Topic #8: br li rmit dha ongoing service strong benefits customer prospective experience employee ul including positi
on property work portfolio performance sales
Topic #9: conduct leave security program li digital market science sound solve feeling expectations city related met
teaching extraction international driving led
```



Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
Marginal topic distribtion
2%
Top-30 Most Relevant Terms for Topic 1 (36.7% of tokens)
0
20
40
60
80
marketing
security
models
financial
digital
policy
machine
statistics
teams
build
identify
background
privacy
predictive
lead
techniques
analytical
strategy
able
campaign
large
value
risk
delivering
care
capability
python
written
desirable
improve
Overall term frequency
Estimated term frequency within the selected topic

Topic extraction using NMF

```
Fitting the NMF model (Frobenius norm) with tf-idf features, n_samples=2000 and n_features=1000...
done in 0.055s.

Topics in NMF model (Frobenius norm):
Topic #0: li strong ul business experience reporting skills role team financial management work contact requirements
amp client analyst company systems apply
Topic #1: strong monash university xa target blank https project research career href edu jobs apply faculty pageuppe
ople equity directions ai www
Topic #2: br strong business experience dha xa working hadoop scripting applications analytics rmit understanding clo
ud ongoing including service benefits performance team
Topic #3: xa research strong able consulting em world payroll demonstrate analysis related track oracle assisting tea
m professional help transformation field record
Topic #4: li analytics advanced learning machine models privacy predictive policy insights modelling value informatio
n collections statistical techniques python unstructured consent bluefinresources
Topic #5: em strong people regulations science worked diverse desirable policy applications position intelligence for
m job statistics demonstrating statement criteria role sets
Topic #6: csiro strong research software privacy australia au development blank target security www scientific balanc
e engineering br flexible phd technologies future
Topic #7: spatial maritime waterway safety questions officer project did challenges xa mapping responses letter usual
cover collection victoria ol resume available
Topic #8: marketing campaign campaigns li brand opportunities reporting analysis teams digital key performance identi
fy privacy global simple provide customer leading channels
Topic #9: care aged residential risk quality governance clinical eastern suburbs compliance standards facilities syst
ems place position processes nursing li staff management
```