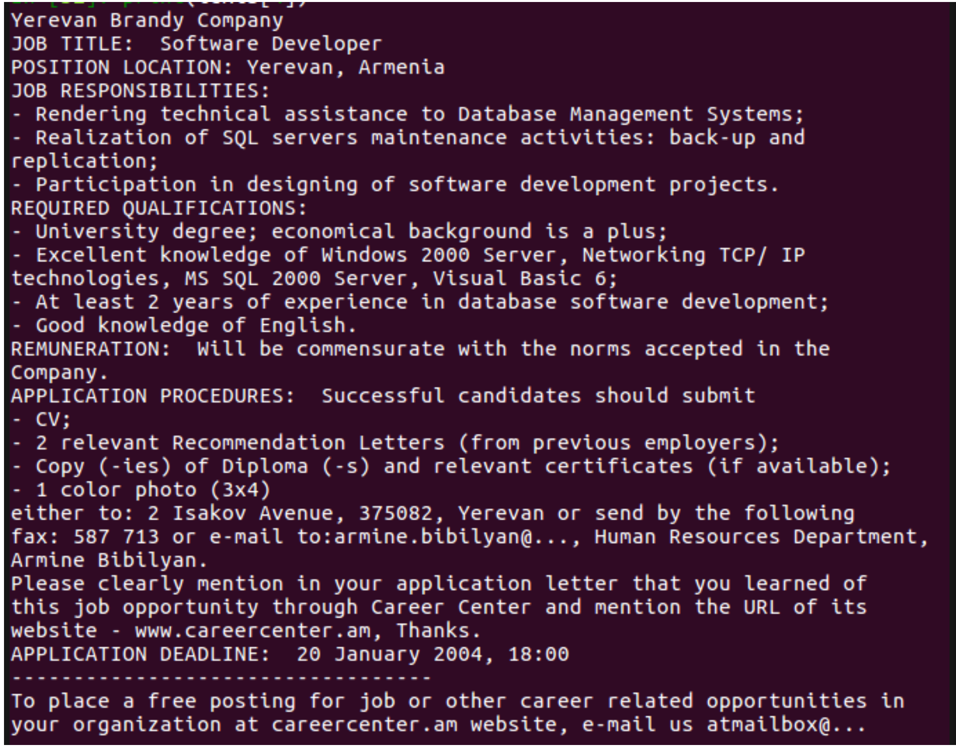
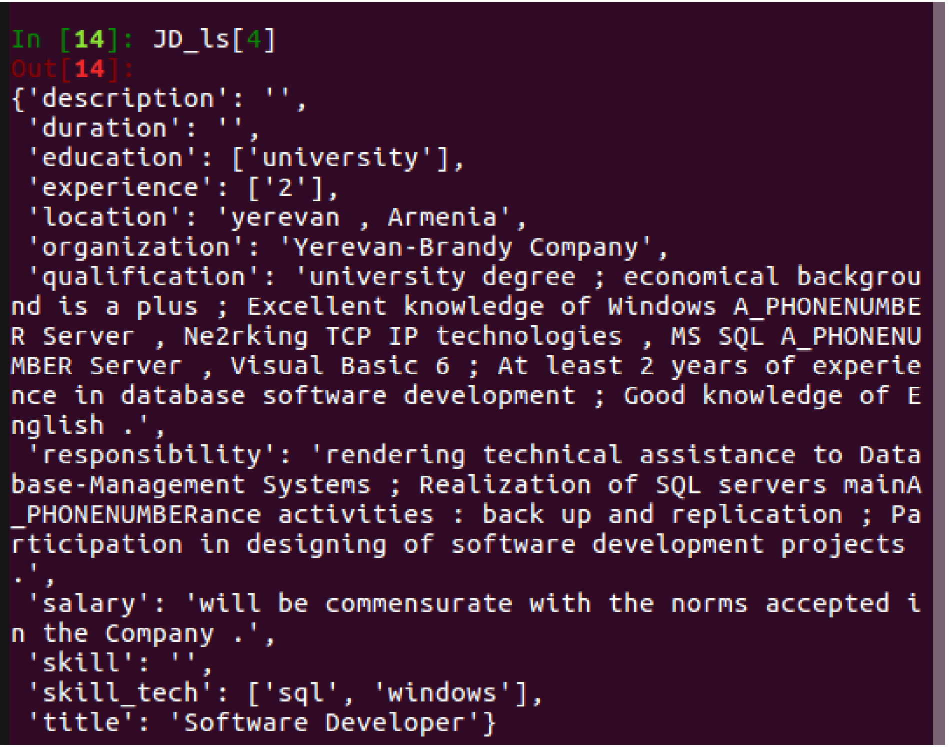
Match job description to job titles

Strayn Wang

Original data -> job description features



```
Yerevan Brandy Company
JOB TITLE:  Software Developer
POSITION LOCATION: Yerevan, Armenia
JOB RESPONSIBILITIES:
- Rendering technical assistance to Database Management Systems;
- Realization of SQL servers maintenance activities: back-up and
replication;
- Participation in designing of software development projects.
REQUIRED QUALIFICATIONS:
- University degree; economical background is a plus;
- Excellent knowledge of Windows 2000 Server, Networking TCP/ IP
technologies, MS SQL 2000 Server, Visual Basic 6;
- At least 2 years of experience in database software development;
- Good knowledge of English.
REMUNERATION:  Will be commensurate with the norms accepted in the
Company.
APPLICATION PROCEDURES:  Successful candidates should submit
- CV;
- 2 relevant Recommendation Letters (from previous employers);
- Copy (-ies) of Diploma (-s) and relevant certificates (if available);
- 1 color photo (3x4)
either to: 2 Isakov Avenue, 375082, Yerevan or send by the following
fax: 587 713 or e-mail to:armine.bibilyan@..., Human Resources Department,
Armine Bibilyan.
Please clearly mention in your application letter that you learned of
this job opportunity through Career Center and mention the URL of its
website - www.careercenter.am, Thanks.
APPLICATION DEADLINE:  20 January 2004, 18:00
----------------------------------
To place a free posting for job or other career related opportunities in
your organization at careercenter.am website, e-mail us atmailbox@...
```



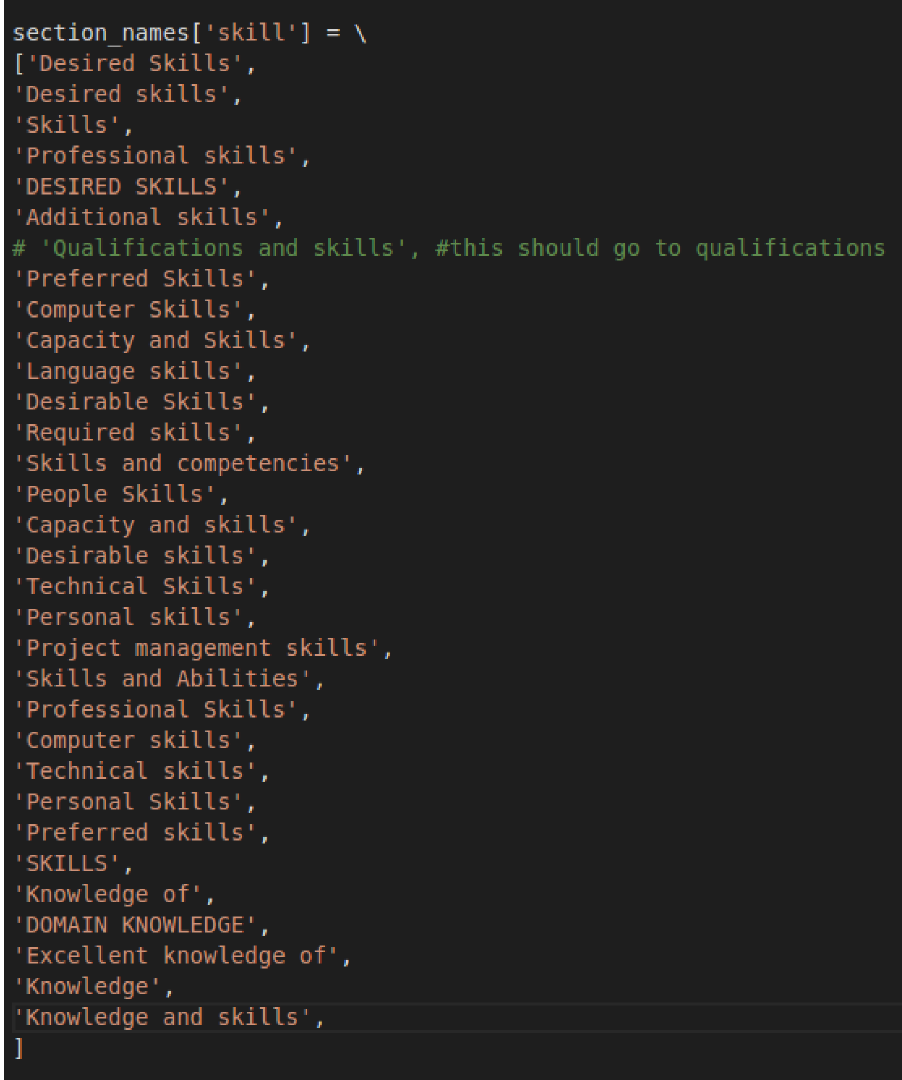
```
In [14]: JD_ls[4]
Out[14]:
{'description': '',
 'duration': '',
 'education': ['university'],
 'experience': ['2'],
 'location': 'yerevan , Armenia',
 'organization': 'Yerevan-Brandy Company',
 'qualification': 'university degree ; economical backgrou
nd is a plus ; Excellent knowledge of Windows A_PHONENUMBE
R Server , Ne2rking TCP IP technologies , MS SQL A_PHONENU
MBER Server , Visual Basic 6 ; At least 2 years of experie
nce in database software development ; Good knowledge of E
nglish .',
 'responsibility': 'rendering technical assistance to Data
base-Management Systems ; Realization of SQL servers mainA
_PHONENUMBERance activities : back up and replication ; Pa
rticipation in designing of software development projects
.',
 'salary': 'will be commensurate with the norms accepted i
n the Company .',
 'skill': '',
 'skill_tech': ['sql', 'windows'],
 'title': 'Software Developer'}
```

Get rid of noises

I split original data into sections, I get

<description, responsibility,qualification,education,experience,salary,skill,organization,title,duration,location>

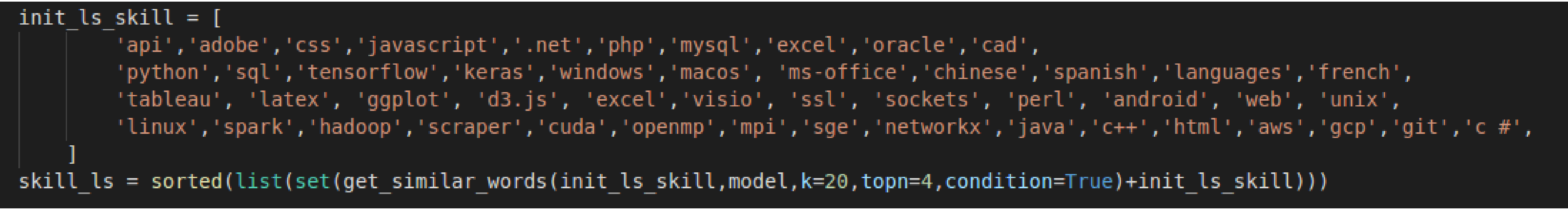
Use <responsibility, qualification> as predicting features



```
section_names['skill'] = \
['Desired Skills',
'Desired skills',
'Skills',
'Professional skills',
'DESIRED SKILLS',
'Additional skills',
# 'Qualifications and skills', #this should go to qualifications
'Preferred Skills',
'Computer Skills',
'Capacity and Skills',
'Language skills',
'Desirable Skills',
'Required skills',
'Skills and competencies',
'People Skills',
'Capacity and skills',
'Desirable skills',
'Technical Skills',
'Personal skills',
'Project management skills',
'Skills and Abilities',
'Professional Skills',
'Computer skills',
'Technical skills',
'Personal Skills',
'Preferred skills',
'SKILLS',
'Knowledge of',
'DOMAIN KNOWLEDGE',
'Excellent knowledge of',
'Knowledge',
'Knowledge and skills',
]
```

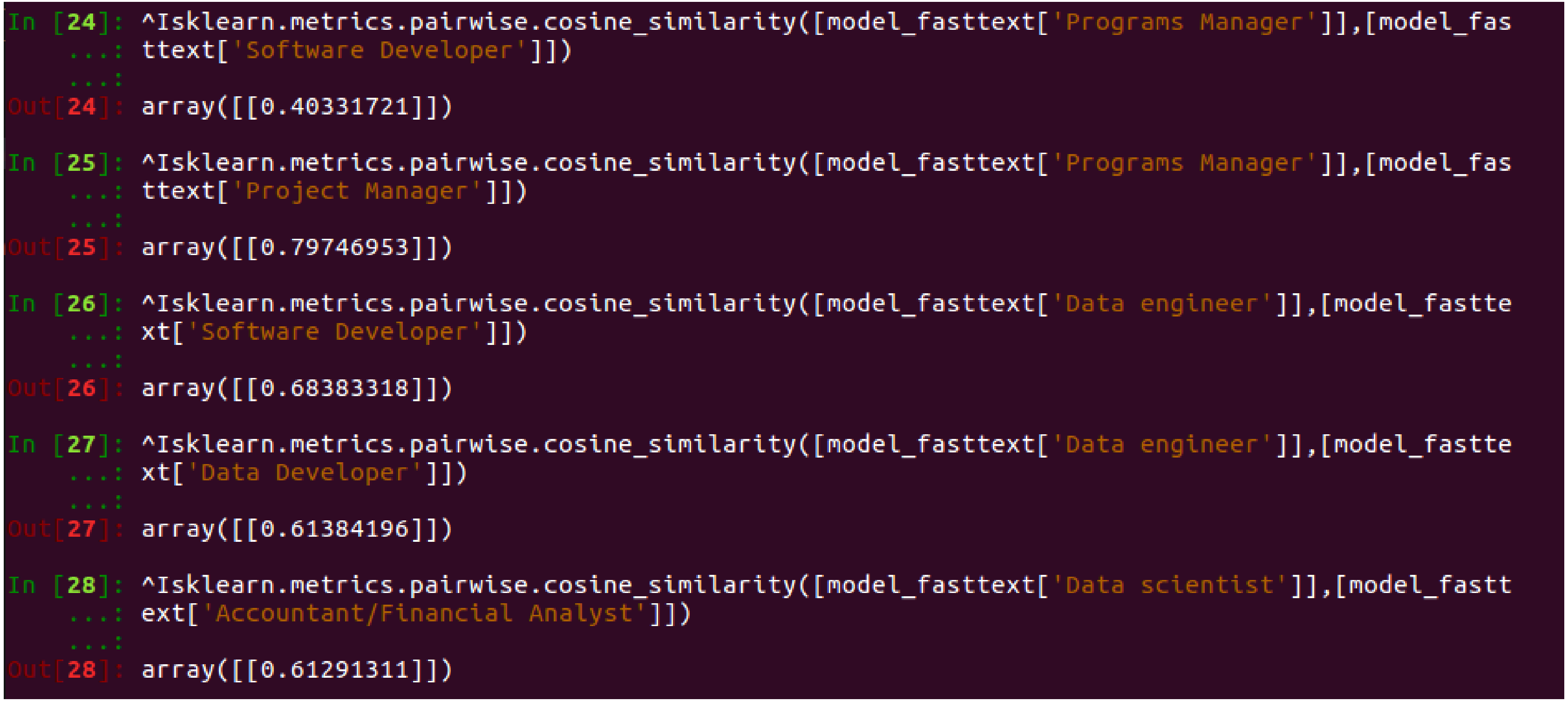
Skill extraction

1. Replace words: replace email/phone/urls etc.
2. Merge terms: concat consective captialized words to one word
3. Get nouns through part of speech tag
4. Get rid of stop words
5. Train Word2Vec to the word sets
6. Start from initialized skill and extract similar words from the model



```
init_ls_skill = [
        'api','adobe','css','javascript','.net','php','mysql','excel','oracle','cad',
        'python','sql','tensorflow','keras','windows','macos', 'ms-office','chinese','spanish','languages','french',
        'tableau', 'latex', 'ggplot', 'd3.js', 'excel','visio', 'ssl', 'sockets', 'perl', 'android', 'web', 'unix',
        'linux','spark','hadoop','scraper','cuda','openmp','mpi','sge','networkx','java','c++','html','aws','gcp','git','c #',
    ]
skill_ls = sorted(list(set(get_similar_words(init_ls_skill,model,k=20,topn=4,condition=True)+init_ls_skill)))
```

Modeling titles - fasttext (train on subwords)

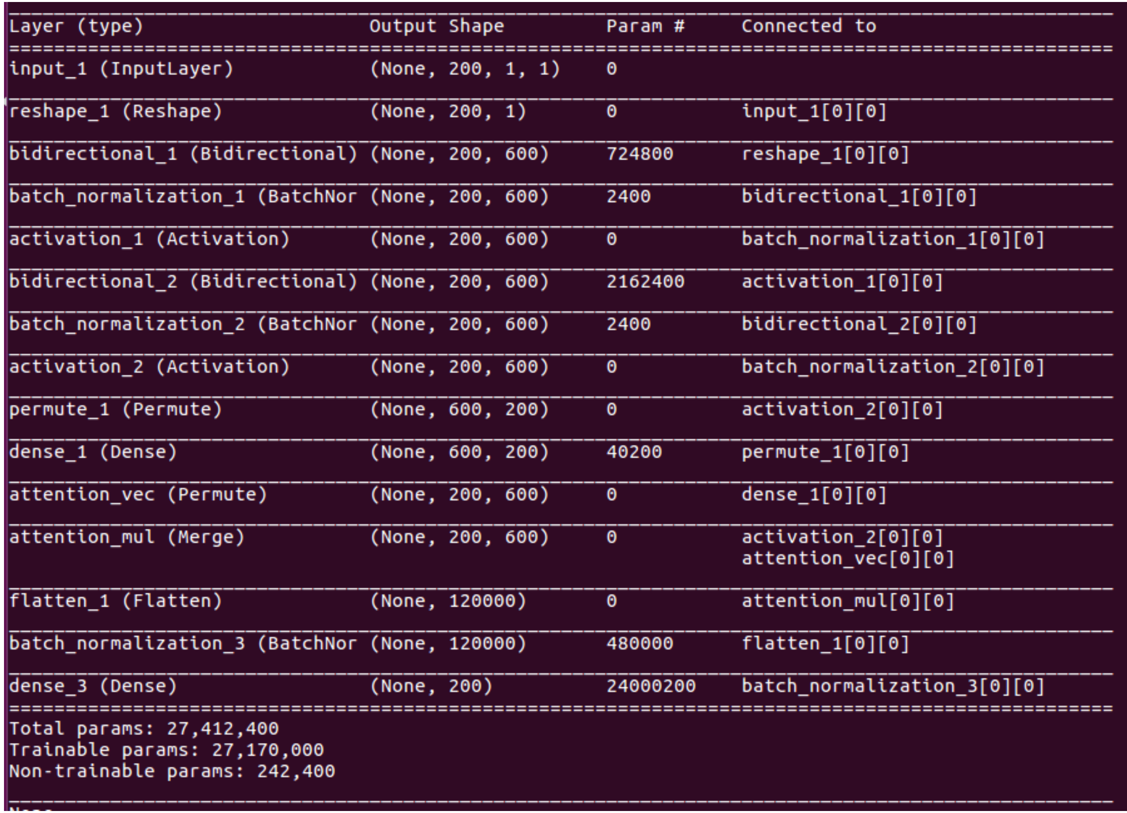


```
In [24]: ^Isklearn.metrics.pairwise.cosine_similarity([model_fasttext['Programs Manager']],[model_fas
    ...: ttext['Software Developer']])
    ...:
Out[24]: array([[0.40331721]])

In [25]: ^Isklearn.metrics.pairwise.cosine_similarity([model_fasttext['Programs Manager']],[model_fas
    ...: ttext['Project Manager']])
    ...:
Out[25]: array([[0.79746953]])

In [26]: ^Isklearn.metrics.pairwise.cosine_similarity([model_fasttext['Data engineer']],[model_fastte
    ...: xt['Software Developer']])
    ...:
Out[26]: array([[0.68383318]])

In [27]: ^Isklearn.metrics.pairwise.cosine_similarity([model_fasttext['Data engineer']],[model_fastte
    ...: xt['Data Developer']])
    ...:
Out[27]: array([[0.61384196]])

In [28]: ^Isklearn.metrics.pairwise.cosine_similarity([model_fasttext['Data scientist']],[model_fastt
    ...: ext['Accountant/Financial Analyst']])
    ...:
Out[28]: array([[0.61291311]])
```

Modeling description

Ideally, I should use split features separately.

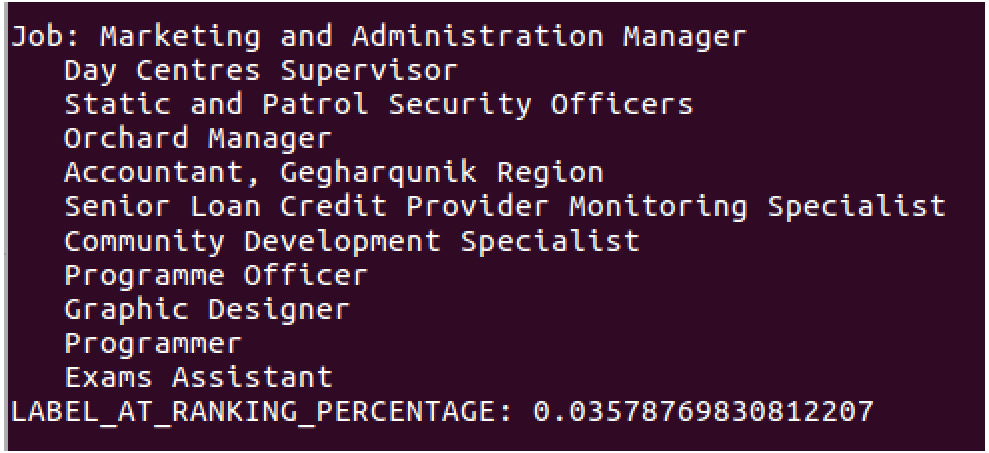
Given limited time, I fed the concatenated <qualification, responsibility> together as input to train LSTM model.



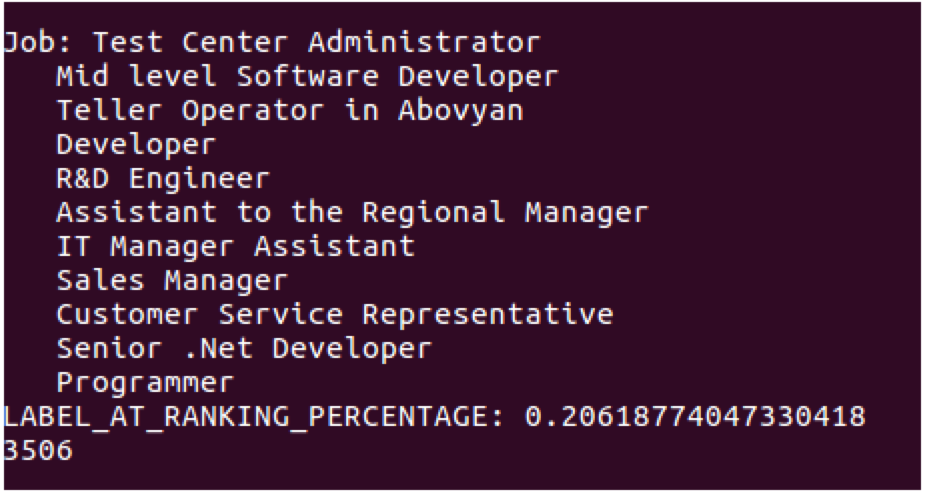
```
__________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
input_1 (InputLayer)            (None, 200, 1, 1)    0
__________________________________________________________________________________________________
reshape_1 (Reshape)             (None, 200, 1)       0           input_1[0][0]
__________________________________________________________________________________________________
bidirectional_1 (Bidirectional) (None, 200, 600)     724800      reshape_1[0][0]
__________________________________________________________________________________________________
batch_normalization_1 (BatchNor (None, 200, 600)     2400        bidirectional_1[0][0]
__________________________________________________________________________________________________
activation_1 (Activation)       (None, 200, 600)     0           batch_normalization_1[0][0]
__________________________________________________________________________________________________
bidirectional_2 (Bidirectional) (None, 200, 600)     2162400     activation_1[0][0]
__________________________________________________________________________________________________
batch_normalization_2 (BatchNor (None, 200, 600)     2400        bidirectional_2[0][0]
__________________________________________________________________________________________________
activation_2 (Activation)       (None, 200, 600)     0           batch_normalization_2[0][0]
__________________________________________________________________________________________________
permute_1 (Permute)             (None, 600, 200)     0           activation_2[0][0]
__________________________________________________________________________________________________
dense_1 (Dense)                 (None, 600, 200)     40200       permute_1[0][0]
__________________________________________________________________________________________________
attention_vec (Permute)         (None, 200, 600)     0           dense_1[0][0]
__________________________________________________________________________________________________
attention_mul (Merge)           (None, 200, 600)     0           activation_2[0][0]
                                                                 attention_vec[0][0]
__________________________________________________________________________________________________
flatten_1 (Flatten)             (None, 120000)       0           attention_mul[0][0]
__________________________________________________________________________________________________
batch_normalization_3 (BatchNor (None, 120000)       480000      flatten_1[0][0]
__________________________________________________________________________________________________
dense_3 (Dense)                 (None, 200)          24000200    batch_normalization_3[0][0]
==================================================================================================
Total params: 27,412,400
Trainable params: 27,170,000
Non-trainable params: 242,400
__________________________________________________________________________________________________
```

Evaluation

Still to be improved



```
Job: Marketing and Administration Manager
   Day Centres Supervisor
   Static and Patrol Security Officers
   Orchard Manager
   Accountant, Gegharqunik Region
   Senior Loan Credit Provider Monitoring Specialist
   Community Development Specialist
   Programme Officer
   Graphic Designer
   Programmer
   Exams Assistant
LABEL_AT_RANKING_PERCENTAGE: 0.03578769830812207
```



```
Job: Test Center Administrator
   Mid level Software Developer
   Teller Operator in Abovyan
   Developer
   R&D Engineer
   Assistant to the Regional Manager
   IT Manager Assistant
   Sales Manager
   Customer Service Representative
   Senior .Net Developer
   Programmer
LABEL_AT_RANKING_PERCENTAGE: 0.20618774047330418
3506
```

Improvement

Error analysis

Use extracted features

Complicated structure