Job recommendation system based on LinkedIn API

First Author: Yutong Chen
UNI: yc3993

Second Author: Yuqin Zhao
UNI: yz4131

Third Author: Shuai Ren
UNI: sr3849

Abstract

Nowadays, hundreds of millions of jobs are posted on LinkedIn, which is the home to countless job seekers and employers. How to recommend and notify these jobs to the job seekers who are interested and suited to correctly and efficiently can be non-trivial. To realize this, a classification distribution should be generated among all job seekers in order to personalize their needs. Thus, what we are working at is to use the personal description job seekers posted on LinkedIn to classify them in different domains. Then we can use the domain to recommend related jobs to them. In addition to recommend jobs to different people with different preferences, we also aim to provide suggestions to users on how to build their skills, how to improve themselves in order to gain a higher chance of being employed. In a word, we try to personalize recommendations and personalize suggestions.

1. Introduction

1.1. Focused Problems

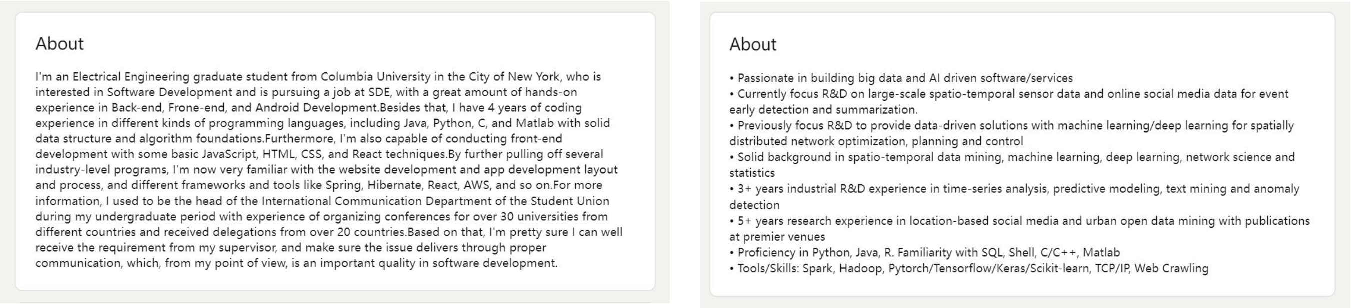
Nowadays, hundreds of millions of jobs are posted on LinkedIn, which is the home to countless job seekers and employers. How to recommend and notify these jobs to the job seekers who are interested and suited to correctly and efficiently can be non-trivial. To realize this, a classification distribution should be generated among all job seekers in order to personalize their needs. Thus, what we are working at is to use the personal description job seekers posted on LinkedIn to classify them in different domains. Then we can use the domain to recommend related jobs to them.

In addition to recommend jobs to different people with different preferences, we also aim to provide suggestions to users on how to build their skills, how to improve themselves in order to gain a higher chance of being employed. In a word, we try to personalize recommendations and personalize suggestions.

1.2. Methods

The first problem can be considered as a classification work, which should be solved by Machine Learning and Deep Learning Models by intuition. To train such models, getting our datasets is our first priority. Thus, by looking up some references[1], we managed to apply for the LinkedIn API and gain the datasets using LinkedIn API. We aim to use API to mine the personal description on LinkedIn.

Then in order to train the model, we need to label the personal description based on their domains as well. Since we are dealing we big data, labeling such datasets one by one by brute force can be impractical. As a result, we decided to use an unsupervised learning model, LDA, to label these personal description for us automatically. With description information and label, we complete our datasets and can be used to train the model. Using LDA to label the



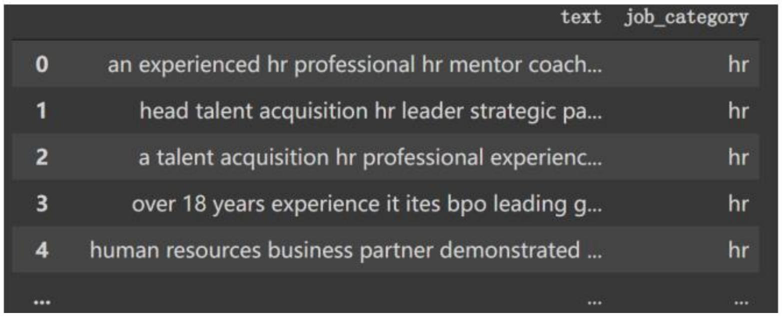
About

I'm an Electrical Engineering graduate student from Columbia University in the City of New York, who is interested in Software Development and is pursuing a job at SDE, with a great amount of hands-on experience in Back-end, Frone-end, and Android Development.Besides that, I have 4 years of coding experience in different kinds of programming languages, including Java, Python, C, and Matlab with solid data structure and algorithm foundations.Furthermore, I'm also capable of conducting front-end development with some basic JavaScript, HTML, CSS, and React techniques.By further pulling off several industry-level programs, I'm now very familiar with the website development and app development layout and process, and different frameworks and tools like Spring, Hibernate, React, AWS, and so on.For more information, I used to be the head of the International Communication Department of the Student Union during my undergraduate period with experience of organizing conferences for over 30 universities from different countries and received delegations from over 20 countries.Based on that, I'm pretty sure I can well receive the requirement from my supervisor, and make sure the issue delivers through proper communication, which, from my point of view, is an important quality in software development.

About

- Passionate in building big data and AI driven software/services
- Currently focus R&D on large-scale spatio-temporal sensor data and online social media data for event early detection and summarization.
- Previously focus R&D to provide data-driven solutions with machine learning/deep learning for spatially distributed network optimization, planning and control
- Solid background in spatio-temporal data mining, machine learning, deep learning, network science and statistics
- 3+ years industrial R&D experience in time-series analysis, predictive modeling, text mining and anomaly detection
- 5+ years research experience in location-based social media and urban open data mining with publications at premier venues
- Proficiency in Python, Java, R. Familiarity with SQL, Shell, C/C++, Matlab
- Tools/Skills: Spark, Hadoop, Pytorch/Tensorflow/Keras/Scikit-learn, TCP/IP, Web Crawling

Figure 1.1: Some Personal Description on LinkedIn



	text	job_category
0	an experienced hr professional hr mentor coach...	hr
1	head talent acquisition hr leader strategic pa...	hr
2	a talent acquisition hr professional experienc...	hr
3	over 18 years experience it ites bpo leading g...	hr
4	human resources business partner demonstrated ...	hr
...	...	...

Figure 1.2: Datasets

data is logical, since what people tend to write in their personal descriptions are, for example, skills they are proficient with, previous work experiences, our even the work domains they are interested in, as shown in the figure 1.1. In this case, description with similar skills tend to be labeled the same and this is exactly what we want to label the data. The more classes we use to label them, the more specific domains we can get.

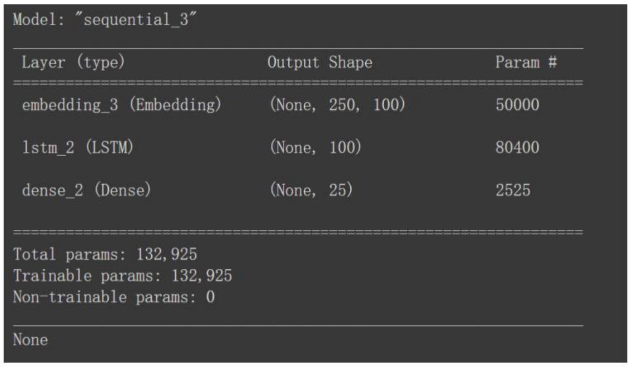
For the model, we currently are using RNN. RNN is suitable for modeling sequential data. It has loops and memories to remember former computations. Variants of RNNs including LSTM and GRU are deployed to overcome the vanishing gradient problem. In our task, since we will have several description of users requirements, so RNN is suitable to solve words classification problems.

The RNN is an extremely expressive model that learns highly complex relationships from a sequence of data. The RNN maintains a vector of activation units for each time step in the sequence of data, this makes RNN extremely deep; the depth of RNN leads to two well-known issues, the exploding and the vanish gradient problem. In order to deal with this problem, we introduced LSTM.

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can process not only single data points (such as images), but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such as unsegmented, connected handwriting recognition, speech recognition and anomaly detection in network traffic or IDSs (intrusion detection systems).

Our data comes from the results of using LDA classification, the data contains 25 categories, as shown in the figure 1.2. First, we need to do data embedding. An embedding is a relatively low-dimensional space into which you can translate high-dimensional vectors. Embeddings make it easier to do deep learning on large inputs like sparse vectors representing words. Ideally, an embedding captures some of the semantics of the input by placing semantically similar inputs close together in the embedding space. An embedding can be learned and reused across models. After completing the distinction between training and test sets, we start to build the model.

The first time the model is the embedding layer (Embedding), it uses a vector of length 100 to represent each word. The LSTM layer contains 100 memory cells. The output layer is a fully connected layer containing 25 categories. Since it is a multi-classification, the activation function is set to “softmax”. Because it is multi-classification, the loss function is



```
Model: "sequential_3"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_3 (Embedding)     (None, 250, 100)          50000

 lstm_2 (LSTM)               (None, 100)               80400

 dense_2 (Dense)             (None, 25)                2525

=================================================================
Total params: 132,925
Trainable params: 132,925
Non-trainable params: 0
_________________________________________________________________
None
```

Figure 1.3: Datasets

categorical_crossentropy. Figure 1.3 shows more detailed information.

Finally, for giving personal suggestions about gaining skills, we decide to use the labeled data we have to do a word count. Use the frequency of words in different domain as a indicator that the top 5 most frequently appeared words appeared in the description tend to be extremely important since so many people in such domain have these skills. We want to visualize the frequency of them using JavaScript and clearly shows what skills are important.

2. Related Work

2.1 Recommender Systems

Recommendation engines tend to have no clear purpose, or their purpose is vague. Generally speaking, users don't even know what they want. At this time, it is the user's place of the recommendation engine. The recommendation system sends the recommendation algorithm through the user's historical behavior, user's interest preference or user's demographic characteristics, Then the recommendation system uses the recommendation algorithm to generate the list of items that users may be interested in, and users are passive to the search engine.

In recent years, the recommender systems has dramatically increased. In the Recommendation algorithm, it classifies into four types: Content-based filtering, Collaborative filtering, Rule-based, and Hybrid approaches.

Collaborative Filtering (CF):

It works through the power of collective wisdom to filter out items that users are not interested in. Collaborative filtering is based on the assumption that a good way to find the content that a specific user is really interested in is to first find other users with similar interests to this user, and then recommend the content they are interested in to this user[2].

Content-based filtering (CBF):

Memory based recommendation system (CF) mainly makes recommendations through heuristic methods. One of the main steps is the selection of similarity function. How to select an appropriate similarity function to better measure the similarity of two items or users is the key; Another major step is how to make recommendations. The simplest recommendation method is based on most recommendation strategies, that is, recommend projects that most people have behavior but the target user has not[3].

Hybrid filtering (HF):

For the fusion of the above algorithms, Taobao has both content-based recommendation and collaborative filtering recommendation. The specific fusion should be combined with specific application scenarios, including feature fusion or algorithm level fusion.

2.2 Literature Review

The article[4] deal with the classification problem by using deep learning algorithms such as textcnn to classify the fields. They recognize text as images and provides advantages for extracting important features，then use a pre trained glove vector and a tokenization sequence to generate a sentence matrix. Finally, the soft Max layer is used as the output to classify 25 categories.

The advantage is that textcnn is used instead of the traditional machine learning method with high accuracy at the beginning, and textcnn is adopted to improve the accuracy.

The second article[5] build recommendation system by Collaborative Filtering. The disadvantages associated with CF relates to scenarios where users or projects do not have enough interactive data (for example, users with abnormal taste, new users or projects). The advantage of this article is to overcome this shortcoming is to develop another type of

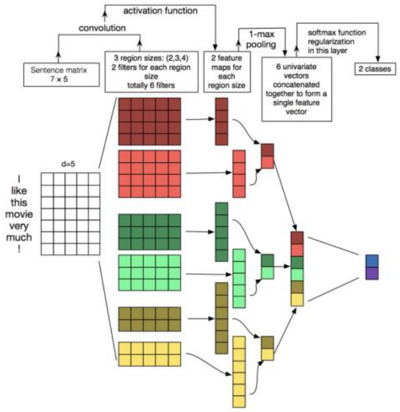

activation function
convolution
1-max pooling
softmax function regularization in this layer
Sentence matrix 7 × 5
3 region sizes: (2,3,4) 2 filters for each region size totally 6 filters
2 feature maps for each region size
6 univariate vectors concatenated together to form a single feature vector
2 classes
d=5
I
like
this
movie
very
much
!

Figure 2.1:job_CNN

recommendation system, the so-called content-based method, which uses the clear domain specific functions of users and projects to build models. The system combining CF and content-based method is called s hybrid recommendation system.

It proposed method combines content-based signals into CF based core system, it is considered as a hybrid method. In this article, content is out in two different scenarios by distinguishing between jobs and users' "cold start". In the case of new work, the content-based similarity measure is calculated between jobs by constructing the neural network embedded in learning work. In this way, even if jobs do not share any user interaction, there are associations between jobs. The problem this article solved is that for users without any activity, model their preferences by classifying their resume content into predefined fine-grained job categories, and then recommend popular jobs to them under these categories[5].

2.3 Current model

Text understanding based on neural network. We think it is meaningful to use recurrent neural networks. Other interesting techniques RNN and recursive convolutional neural networks that show interesting results.

The model uses circular structure to capture context information, and uses convolutional neural network to construct text representation. On the other hand, text understanding and text classification are most commonly used in texts composed of complete and grammatically correct sentences, such as news articles or short stories. Abstracts are structurally different because people often use key points and abbreviations to express their views more effectively.

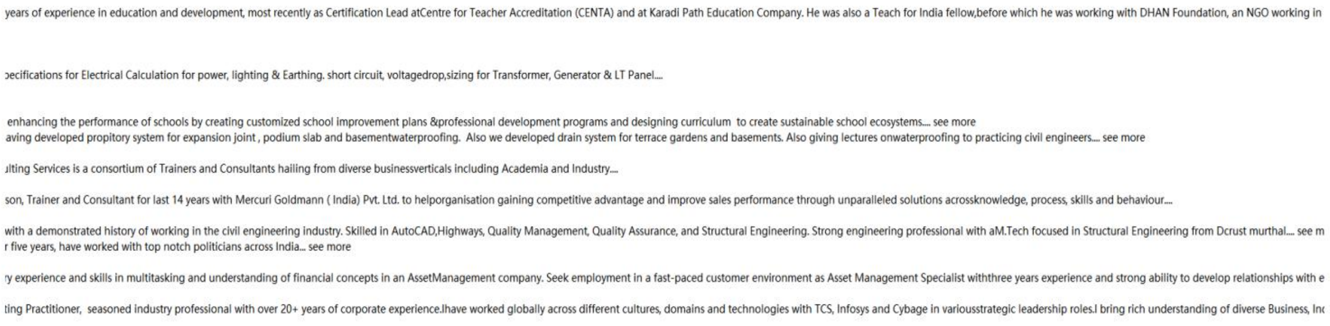
3. Current Progress

For the data mining part, now we have successfully got the data, which is the personal description information, as shown in Figure 3.1. As we can see, the raw data contains lots of unnecessary information: stop words. Since we decide to use LDA to label the data, such data with so many stop words can be detrimental to the result. Thus, we use nltk lib to filter them out[6]. With the pre-processed data, we use LDA to generate a topic distribution to label them, as shown in Figure 3.2.

After data prepossessing part, we start to build our model. First, we delete the Chinese and garbled characters in the job description. Then delete some words that are often used in daily life but do not affect the meaning of the text. Finally, group the data into the model for training.The size of our training data and model definition are shown in Figure 3.3, 3.4.

Since the time duration is very long for each epoch. We will only set our epochs = 5. The train accuracy is shown in Figure 3.5

Finally, we performed a simple test on our model, trying to find out if our model is capable of processing a description and give a domain. Results are shown in Figure 3.6.



years of experience in education and development, most recently as Certification Lead atCentre for Teacher Accreditation (CENTA) and at Karadi Path Education Company. He was also a Teach for India fellow,before which he was working with DHAN Foundation, an NGO working in

pecifications for Electrical Calculation for power, lighting & Earthing. short circuit, voltagedrop,sizing for Transformer, Generator & LT Panel....

enhancing the performance of schools by creating customized school improvement plans &professional development programs and designing curriculum to create sustainable school ecosystems.... see more
aving developed propitory system for expansion joint , podium slab and basementwaterproofing. Also we developed drain system for terrace gardens and basements. Also giving lectures onwaterproofing to practicing civil engineers.... see more

ulting Services is a consortium of Trainers and Consultants hailing from diverse businessverticals including Academia and Industry....

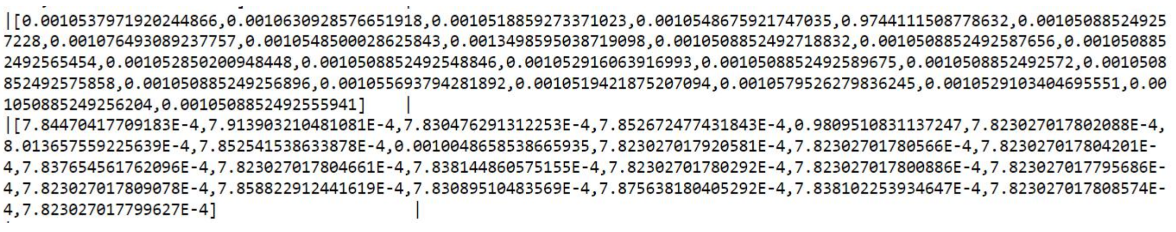
son, Trainer and Consultant for last 14 years with Mercuri Goldmann (India) Pvt. Ltd. to helporganisation gaining competitive advantage and improve sales performance through unparalleled solutions acrossknowledge, process, skills and behaviour....

with a demonstrated history of working in the civil engineering industry. Skilled in AutoCAD,Highways, Quality Management, Quality Assurance, and Structural Engineering. Strong engineering professional with aM.Tech focused in Structural Engineering from Dcrust murthal.... see m
r five years, have worked with top notch politicians across India... see more

y experience and skills in multitasking and understanding of financial concepts in an AssetManagement company. Seek employment in a fast-paced customer environment as Asset Management Specialist withthree years experience and strong ability to develop relationships with e

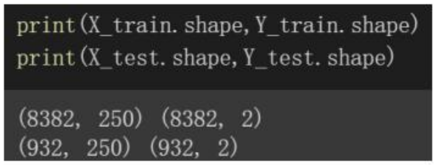
ting Practitioner, seasoned industry professional with over 20+ years of corporate experience.Ihave worked globally across different cultures, domains and technologies with TCS, Infosys and Cybage in variousstrategic leadership roles.I bring rich understanding of diverse Business, Inc

Figure 3.1: raw data



```
|[0.0010537971920244866,0.0010630928576651918,0.0010518859273371023,0.0010548675921747035,0.9744111508778632,0.00105088524925
7228,0.001076493089237757,0.0010548500028625843,0.0013498595038719098,0.0010508852492718832,0.0010508852492587656,0.001050885
2492565454,0.001052850200948448,0.0010508852492548846,0.001052916063916993,0.0010508852492589675,0.0010508852492572,0.0010508
852492575858,0.001050885249256896,0.001055693794281892,0.0010519421875207094,0.0010579526279836245,0.0010529103404695551,0.00
1050885249256204,0.0010508852492555941]     |
|[7.84470417709183E-4,7.913903210481081E-4,7.830476291312253E-4,7.852672477431843E-4,0.9809510831137247,7.823027017802088E-4,
8.013657559225639E-4,7.852541538633878E-4,0.0010048658538665935,7.823027017920581E-4,7.82302701780566E-4,7.823027017804201E-
4,7.837654561762096E-4,7.823027017804661E-4,7.838144860575155E-4,7.82302701780292E-4,7.823027017800886E-4,7.823027017795686E-
4,7.823027017809078E-4,7.858822912441619E-4,7.83089510483569E-4,7.875638180405292E-4,7.838102253934647E-4,7.823027017808574E-
4,7.823027017799627E-4]                  |
```

Figure 3.2: LDA results



```
print(X_train.shape, Y_train.shape)
print(X_test.shape, Y_test.shape)

(8382, 250) (8382, 2)
(932, 250) (932, 2)
```

Figure 3.3 data size

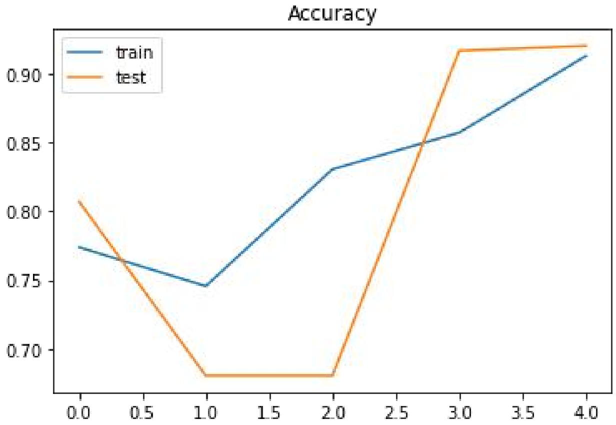


Figure 3.5 train accuracy

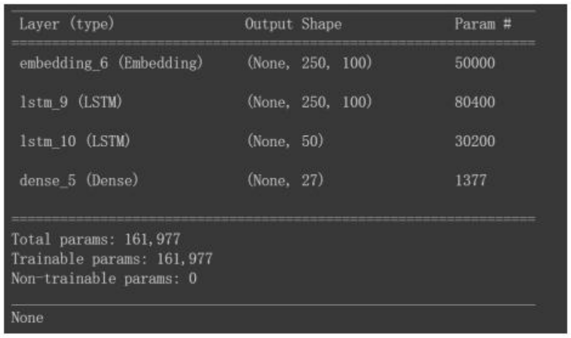
4. Planed Experiment

The experiment we planned is to predict relative job categories to several job description.

Job description	Job category

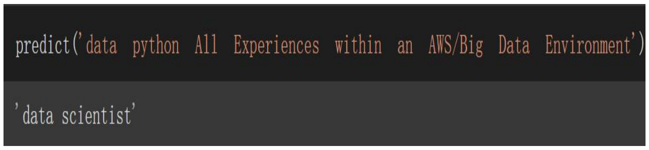
For example, we will give the model a txt: “Deep Learning, Python, R, Decision Tree, Bagging, Boosting, Random, Forest, NLP, Natural Language If you are a Data Scientist with 2+ years of experience with Deep Learning, please read on! Top Reasons to Work with Us This position is based in the downtown Chicago area within a great work location and comes with a highly competitive salary. We need your expertise with Deep Learning/Python/R to help some of our premier clients.” And the model will recommend data scientist to the user.

After that, we will put this on a website. And test it among several students.



```
Layer (type)                 Output Shape              Param #
=================================================================
embedding_6 (Embedding)      (None, 250, 100)          50000

lstm_9 (LSTM)                (None, 250, 100)          80400

lstm_10 (LSTM)               (None, 50)                30200

dense_5 (Dense)              (None, 27)                1377

=================================================================
Total params: 161,977
Trainable params: 161,977
Non-trainable params: 0
_________________________________________________________________
None
```

Figure 3.4 model definition



```
predict('data python All Experiences within an AWS/Big Data Environment')

'data scientist'
```

Figure 3.6 testing

References

[1] https://towardsdatascience.com/linkedin-api-python-programmatically-publishing-d88a03f08ff1
[2] https://en.wikipedia.org/wiki/Collaborative_filtering
[3] https://en.wikipedia.org/wiki/Recommender_system#Content-based_filtering
[4] https://medium.com/ai-techsystems/introduction-2c8fa6dc8924
[5] W. Shalaby et al., "Help me find a job: A graph-based approach for job recommendation at scale," 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 1544-1553, doi: 10.1109/BigData.2017.8258088.
[6] https://pythonspot.com/nltk-stop-words