PROJECT NYC JOBS ANALYSIS

User Guide

Meihao Chen (mc5283) · Po-Chih Lin (pcl322) · Yitong Wang (yw652)

1. Introduction

This project utilizes the NYC Jobs dataset from NYC Open Data to gain the deep insight of the posted jobs at NYC during 2012 to 2014. Several interactions between the program and the user are provided, including user's salary prediction, retrieval of keyterms in job description, and the plot drawing for graphical analysis with controllable options. This user guide is going to illustrate the basic UI flow for users and show some sample results of the program.

2. Project Structure

I. Project

project/ ->	*main.py*	*: The main program*
	test.py	*: Unit testing*
	lexicon.txt	*: A sample output for the lexicon configured by users*
	src/	*: The modules and external source codes*
	data/	*: The dataset*
	plots/	*: The plots generated for users*

II. Src

src/ ->	*genism/*	*: The* ***external source codes*** *of* ***genism*** *lib for Natural Language Processing*
	classLexicon.py	*: The class for lexicon*
	classSalaryPredictor.py	*: The class for predictor of salary interval*
	dataAnalysis.py	*: Algorithms for data analysis*
	dataProcessing.py	*: Data cleaning and pre-processing*
	exception.py	*: Exceptions*
	featureSummary.py	*: Data summarization and computation*
	functions.py	*: Featured functions of the project*
	__init__.py	*: Initialization for modules*
	LDA.py	*: The LDA (Latent Dirichlet Allocation) model training and pre-processing*
	lexFeature.py	*: Lexicographical feature generation*
	plot.py	*: Drawing plots*
	ui.py	*: User interfaces*
	utility.py	*: Validation and string processing*

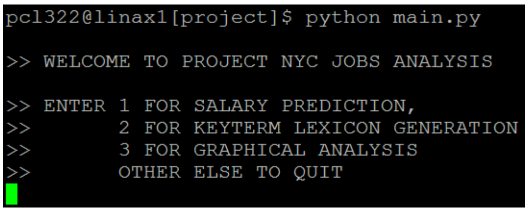
III. Data

data/ ->	*NYC_Jobs.csv*	*: The NYC Jobs from NYC Open Data*
	NYC_Jobs.lex	*: The default job keyterm lexicon (20 topics)*
	sample_resume_1.txt	*: A sample resume*
	sample_resume_2.txt	*: A sample resume*
	sample_resume_3.txt	*: A sample resume*
	stopwords/	*: The stopword list includes several language versions* *(**extracted externally from NLTK toolkit**)*

IV. Plot

plots/ ->	**.png*	*: The plots to be generated*

3. Start the Program

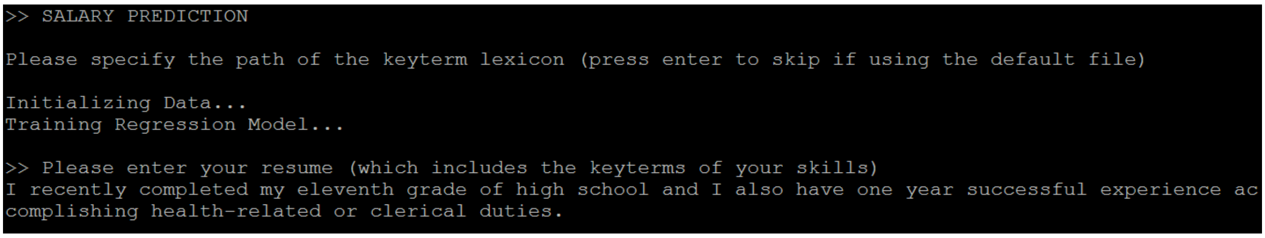


```
pcl322@linax1[project]$ python main.py

>> WELCOME TO PROJECT NYC JOBS ANALYSIS

>> ENTER 1 FOR SALARY PREDICTION,
>>       2 FOR KEYTERM LEXICON GENERATION
>>       3 FOR GRAPHICAL ANALYSIS
>>       OTHER ELSE TO QUIT
```

Just by simply typing “python main.py” we can start the whole project with the above welcome message and instructions. And then press 1, 2, or 3, to enter each different functions.

4. Salary Prediction

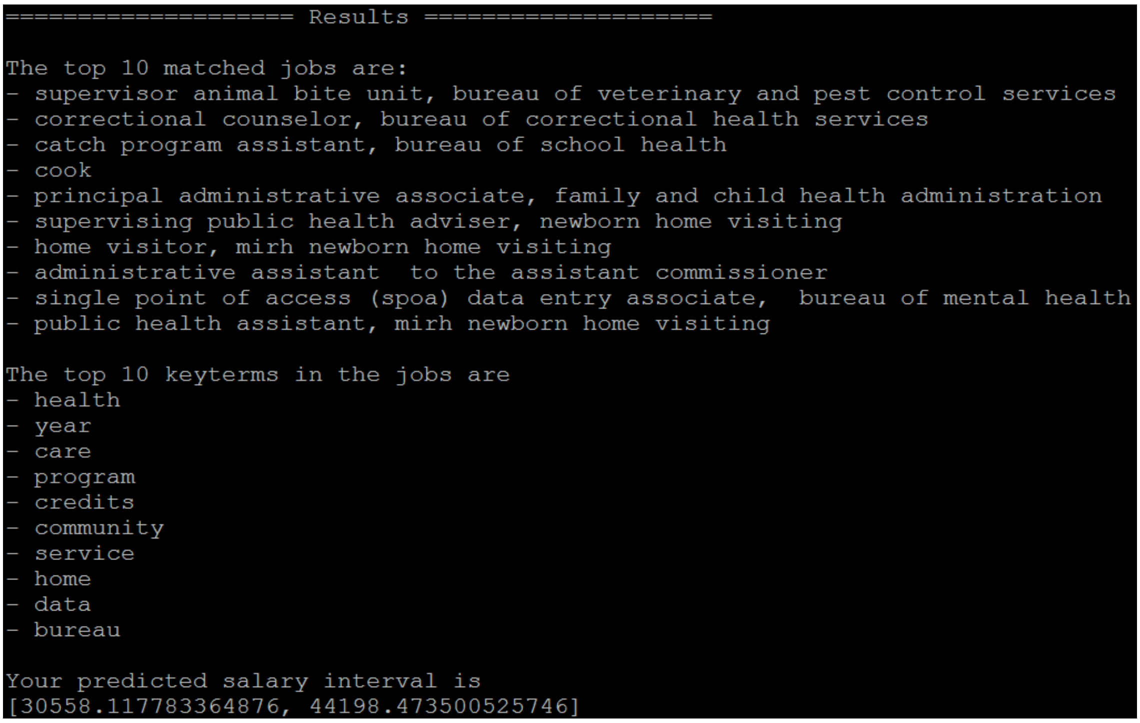
The purpose of this function is to find the matched jobs and the key terms shared by these jobs given your resume and the key term lexicon. Furthermore, it will predict your salary interval using the pre-computed Stochastic Gradient Descent Regression model.



```
>> SALARY PREDICTION

Please specify the path of the keyterm lexicon (press enter to skip if using the default file)

Initializing Data...
Training Regression Model...

>> Please enter your resume (which includes the keyterms of your skills)
I recently completed my eleventh grade of high school and I also have one year successful experience ac
complishing health-related or clerical duties.
```

First of all, you need to enter the file path of the keyterm lexicon. This lexicon can be generated by function 2 which will be introduced as follows. If you skip specifying the lexicon, the program will use the default one which is train by the LDA model with 20 topics.

Secondly, enter your resume. The algorithm of matching degree of your resume and the jobs are based on the keyterms. Please remember to include the keyterms (e.g., construction, programming, biological, and law) indicating your skills. (Start from trying the “sample_resume_1~3.txt” in project/data/)

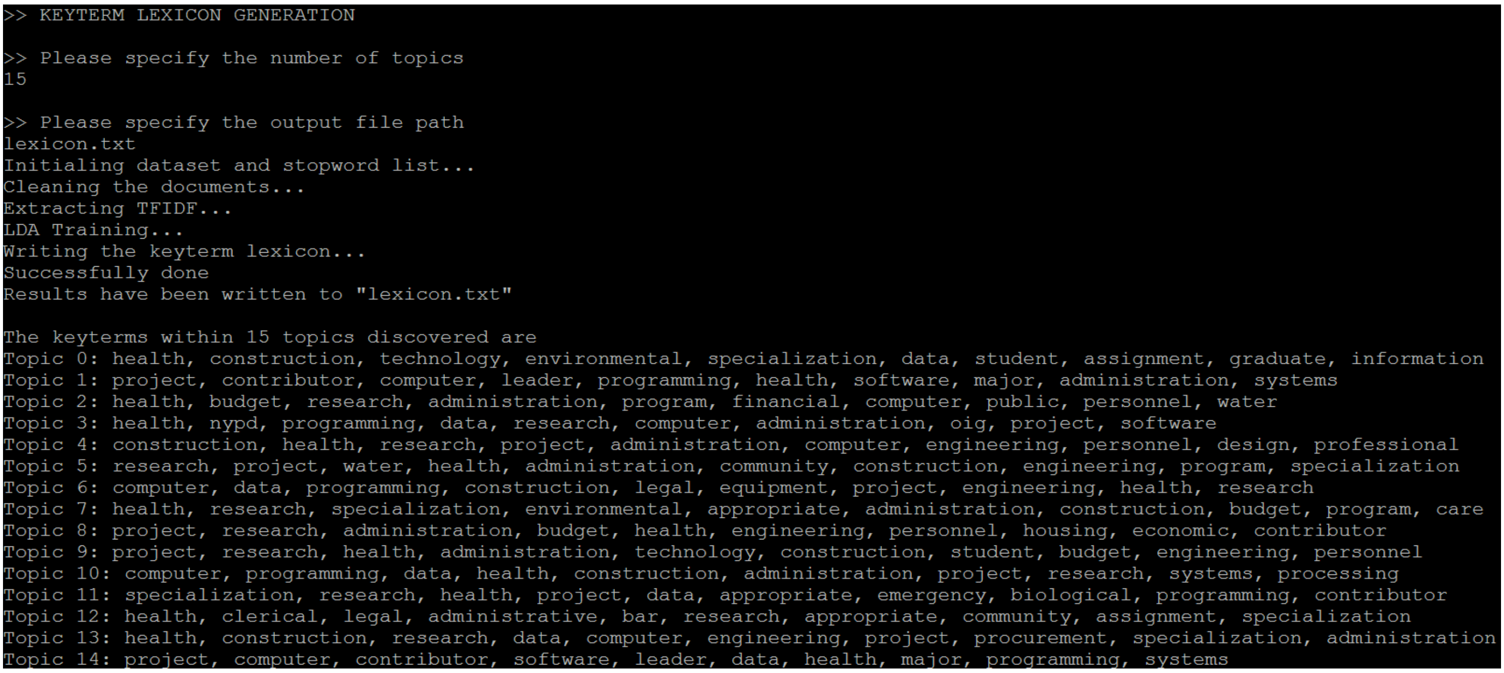


```
=================== Results ===================

The top 10 matched jobs are:
- supervisor animal bite unit, bureau of veterinary and pest control services
- correctional counselor, bureau of correctional health services
- catch program assistant, bureau of school health
- cook
- principal administrative associate, family and child health administration
- supervising public health adviser, newborn home visiting
- home visitor, mirh newborn home visiting
- administrative assistant  to the assistant commissioner
- single point of access (spoa) data entry associate,  bureau of mental health
- public health assistant, mirh newborn home visiting

The top 10 keyterms in the jobs are
- health
- year
- care
- program
- credits
- community
- service
- home
- data
- bureau

Your predicted salary interval is
[30558.117783364876, 44198.473500525746]
```

The result is shown like this. The program retrieves top 10 matched jobs followed by the top 10 keyterms shared by these jobs, and predicts the salary interval using the SGD regression model.

5. Keyterm Lexicon Generation

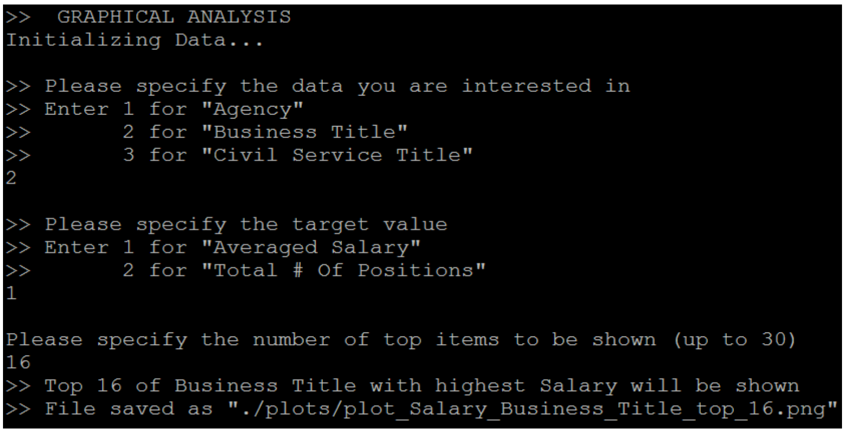
This function trains the LDA model and combines all the top 10 keyterms within each topic to build a keyterm lexicon. For example, the user trained a 20-topic LDA model, so there would be at most 20*10=200 keyterms among all these topics. However, several keyterms are usually shared by different topics. Therefore the size of the lexicon is usually less than the number of topics * 10. In the following example, we specify the number of topics to be equal to 15, and the lexicon file is saved to ./lexicon.txt.



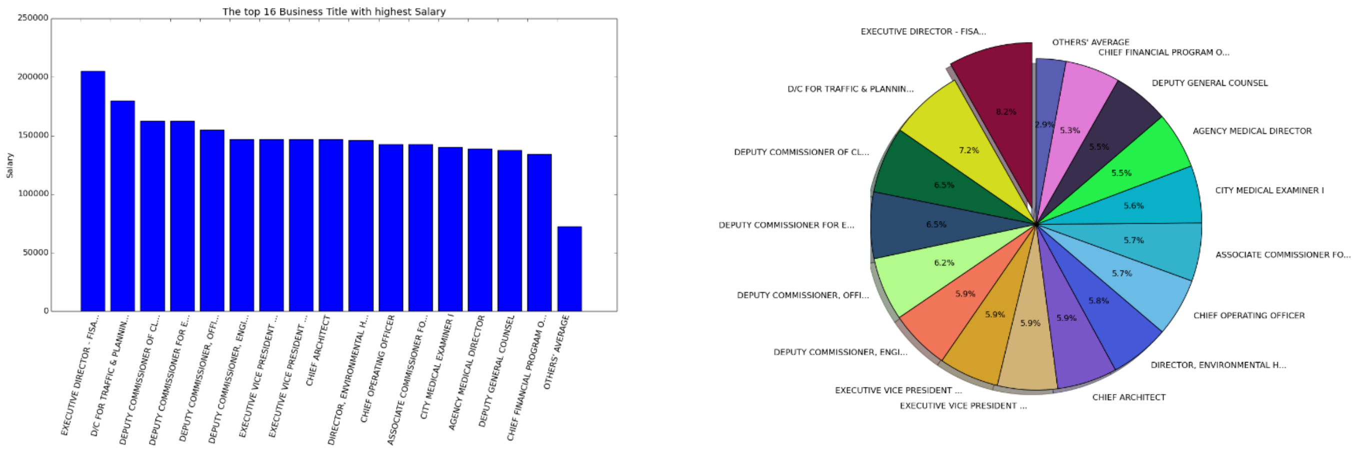
```
>> KEYTERM LEXICON GENERATION

>> Please specify the number of topics
15

>> Please specify the output file path
lexicon.txt
Initialing dataset and stopword list...
Cleaning the documents...
Extracting TFIDF...
LDA Training...
Writing the keyterm lexicon...
Successfully done
Results have been written to "lexicon.txt"

The keyterms within 15 topics discovered are
Topic 0: health, construction, technology, environmental, specialization, data, student, assignment, graduate, information
Topic 1: project, contributor, computer, leader, programming, health, software, major, administration, systems
Topic 2: health, budget, research, administration, program, financial, computer, public, personnel, water
Topic 3: health, nypd, programming, data, research, computer, administration, oig, project, software
Topic 4: construction, health, research, project, administration, computer, engineering, personnel, design, professional
Topic 5: research, project, water, health, administration, community, construction, engineering, program, specialization
Topic 6: computer, data, programming, construction, legal, equipment, project, engineering, health, research
Topic 7: health, research, specialization, environmental, appropriate, administration, construction, budget, program, care
Topic 8: project, research, administration, budget, health, engineering, personnel, housing, economic, contributor
Topic 9: project, research, health, administration, technology, construction, student, budget, engineering, personnel
Topic 10: computer, programming, data, health, construction, administration, project, research, systems, processing
Topic 11: specialization, research, health, project, data, appropriate, emergency, biological, programming, contributor
Topic 12: health, clerical, legal, administrative, bar, research, appropriate, community, assignment, specialization
Topic 13: health, construction, research, data, computer, engineering, project, procurement, specialization, administration
Topic 14: project, computer, contributor, software, leader, data, health, major, programming, systems
```

6. Graphical Analysis

Here we do some graphical analysis on the data that is interested by the user. User is allowed to choose either “Agency,” “Business Title,” or “Civil Service Title.” For each feature (category) user can decide to see the distribution of either salary or the number of position within that category.



```
>>  GRAPHICAL ANALYSIS
Initializing Data...

>> Please specify the data you are interested in
>> Enter 1 for "Agency"
>>       2 for "Business Title"
>>       3 for "Civil Service Title"
2

>> Please specify the target value
>> Enter 1 for "Averaged Salary"
>>       2 for "Total # Of Positions"
1

Please specify the number of top items to be shown (up to 30)
16
>> Top 16 of Business Title with highest Salary will be shown
>> File saved as "./plots/plot_Salary_Business_Title_top_16.png"
```



In this example, the user has chosen to illustrate the salary distribution of top 16 Business Title, and the plot is save as “./plots/plot_Salary_Business_Title_top_16.png.” The sub-figure on the left of the plot is a bar chart of and top 16 business title plus others (those ranked below 16) against their salary. The sub-figure on the right is the pie chart of the same data but shown to reveal more relative information between each category.

7. Reference

- NYC Open Data:
  https://nycopendata.socrata.com/
- Gensim:
  https://radimrehurek.com/gensim/
- Natural Language Toolkit:
  http://www.nltk.org/
- Latent Dirichlet Allocation (LDA):
  http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation
- Stochastic gradient descent :
  http://en.wikipedia.org/wiki/Stochastic_gradient_descent