A method to improve the response rate of an online forum

Guanlong Fu

Introduction

CareerVillage.org is a nonprofit that crowdsources career advice for underserved youth. Founded in 2011 in four classrooms in New York City, the platform has now served career advice from 25,000 volunteer professionals to over 3.5M online learners. The platform uses a Q&A style similar to StackOverflow or Quora to provide students with answers to any question about any career.

In U.S., there are almost 500 students for every guidance counselor. Underserved youth lack the network to find their career role models, making CareerVillage.org the only option for millions of young people in America and around the globe with nowhere else to turn.

To date, 25,000 volunteers have created profiles and opted in to receive emails when a career question is a good fit for them. This is where your skills come in. To help students get the advice they need, the team at CareerVillage.org needs to be able to send the right questions to the right volunteers. The notifications sent to volunteers seem to have the greatest impact on how many questions are answered.

The current recommendation system mainly recommends questions to professionals who shared the same or similar hash tags with the asked questions. Although according to the dataset offered, the answer rate is more than 90 percent, there are more than half of the professionals did not answer any questions and the time taken until questions are answered is somewhat long. The goal of the data challenge is: develop a method to recommend relevant questions to the professionals who are most likely to answer them, with the hope that the proposed new recommendation engine will overcome some of the shortcomings that the current one has.

The data

10 datasets are offered to do the analysis. The introduction of some main datasets is as follows. **Answers:** answers get posted by professionals in response to questions. There are totally 51123 answers corresponding to 23117 questions answered by 10169 professionals. **Emails:** Emails could be sent to either a professional to recommend a question or a student to remind him that his question is answered. There are totally 1850101 unique emails sent to 22168 recipients. **Professionals**: who answer questions. There are totally 28152 observations of professionals from 2583 different locations and

2425 industries. **Questions**: there are totally 23931 questions asked by 12331 students and each has a title, a body of content and one to several hashtags. **Hashtags:** there are 16269 unique hashtags. Professionals and students follow hashtags and each question has to be attached with at least one hashtag by students who ask them.

As shown in figure 1, questions can be linked to professionals through hashtags (and potentially the title and body of questions), answers (each answered_question_id corresponds to an answer_author_id). Questions can also be linked to professionals through students via their shared school, group membership, and location with professionals. However, the information of school, group membership and location is largely missed for students. Thus our analysis will be mainly focus on the link through question hashtag, question title and question body and through answers.

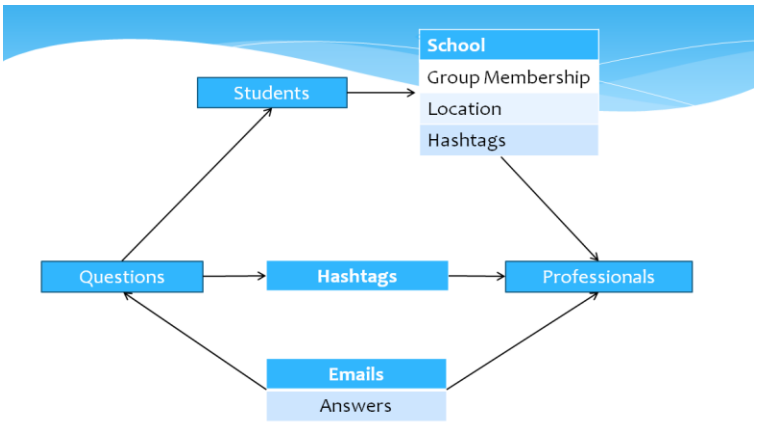


Figure 1: the relationship between datasets

Figure 2 show the barplots for hashtags of questions (left) and professionals (right). From this figure we can see that there are some miss-matches in the frequency distribution of hashtags between questions and professionals, so recommend only based on hashtags might not be accurate and might lead to nonresponse of professionals.

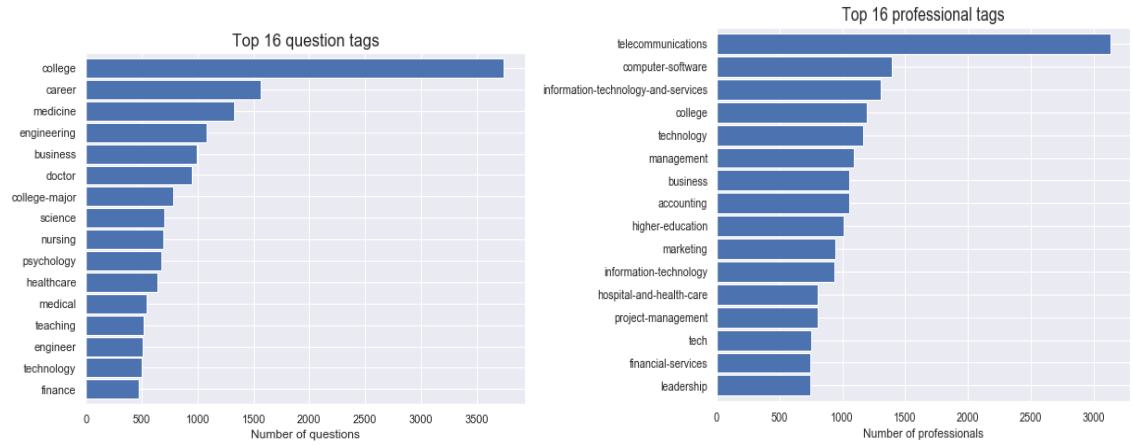


Figure 2: barplots for barplots for hashtags of questions (left) and professionals (right).

We first merge the question dataset with the hash tag dataset to get the full information of questions:

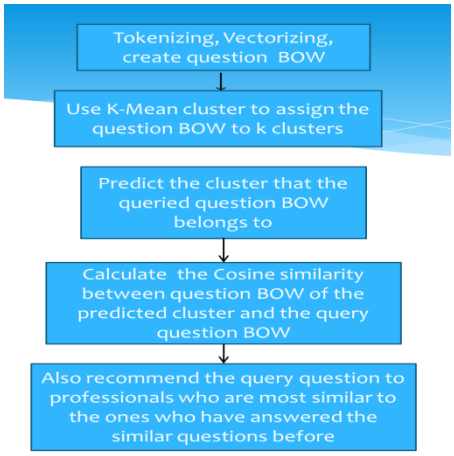
questions_id	questions_author_id	questions_date_added	questions_title	questions_body	tags_tag_name
485cf7a7a556a5e54	8f6f374ffd834d258ab69d376dd998f5	2016-04-26 11:14:26 UTC+0000	Teacher career question	What is a maths teacher? what is a ma...	lecture college professor
lcad8f16bc25aa2d9c	acccbda28edd4362ab03fb8b6fd2d67b	2016-05-20 16:48:25 UTC+0000	I want to become an army officer. What can I d...	I am Priyanka from Bangalore . Now am in 10th ...	military army

We then merge the questions dataset to the answers dataset and to the professionals dataset so as to get an idea which questions is answered by which professional:

	answers_question_id	answers_author_id
0	f6b9ca94aed04ba28256492708e74f60	9ced4ce7519049c0944147afb75a8ce3

Methodology

We will combine the mainstream recommendation algorithm, collaborative filtering, with k-mean cluster, to get better accuracy, performance, and response rate. The algorithm is illustrated in Figure 3 (a).



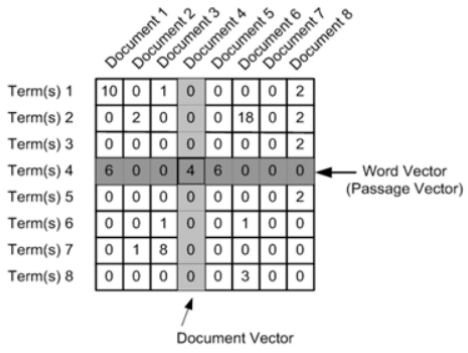


Figure 3(a): the algorithm of this recommendation engine; Figure 3(b): the words embedding sparse matrix (https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/)

First, we create the question bag of words (BOW) by combining the text of the question-title, the question-body and the question-hashtag of each of the 23931 questions.

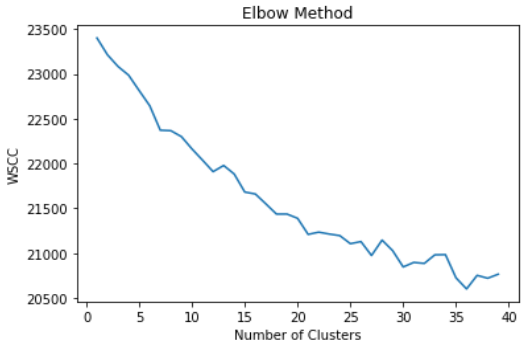
```
0    Teacher  career  question What  is  a  maths...
1    I want to become an army officer. What can I d...
2    Will going abroad for your first job increase ...
```

The above picture gives the example of the BOW of the first three questions.

Then we use the TFIDF vectorizer, term frequency- inverse document frequency, to implement the words tokenizing and embedding. The idea of TFIDF is to give the words those are specific to one document/question more weights and the words that occur in all documents/questions less weights. The tokenizing and embedding result, illustrated in Figure 3(b) would be a large sparse matrix whose rows represent each term/token in the corpus of the questions BOW and columns stand for each document/question in the corpus of the questions BOW. And each element in this matrix represents the TFIDF of that term/token in that document. The dimension of the resulted sparse matrix is 5391 by 23931, saying that this matrix has 5391 tokens/features and 23931 documents. Below are first few token/features of all the 5391 tokens:

```
'across',
'act',
'acting',
'action',
'actions',
'active',
'actively',
'activism',
'activist',
'activities',
'activity',
'actor',
```

After tokenizing and embedding is done, we can start the numerical analysis with the tokenized and vectorized questions BOW. We employ k-mean cluster on the resulted sparse embedding matrix to assign questions to k different clusters/topics. The minibach k-mean is implemented with k taking the value from 1 to 40. The WSCC versus k plot is as follows:



We can see from the above plot that after k=20, the WSCC is not decreasing quickly. Therefore, we select k=20 to fit our k-mean model. Each question will get a predicted cluster label with the fitted k-mean cluster model.

ons_id	questions_author_id	questions_date_added	questions_title	questions_body	tags_tag_name	cluster_id
a5e54	8f6f374ffd834d258ab69d376dd998f5	2016-04-26 11:14:26 UTC+0000	Teacher career question	What is a maths teacher? what is a ma...	lecture college professor	9
aa2d9c	acccbda28edd4362ab03fb8b6fd2d67b	2016-05-20 16:48:25 UTC+0000	I want to become an army officer. What can I d...	I am Priyanka from Bangalore . Now am in 10th ...	military army	9
0decff	f2c179a563024ccc927399ce529094b5	2017-02-08 19:13:38 UTC+0000	Will going abroad for your first job increase ...	I'm planning on going abroad for my first job....	overseas working-abroad	12

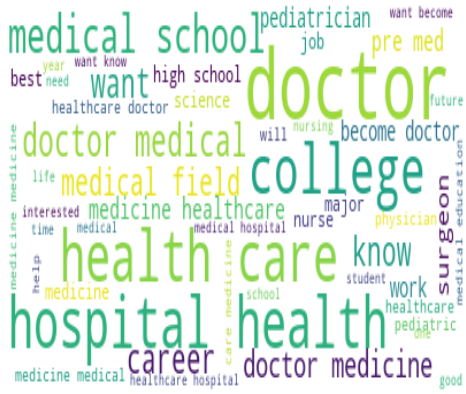
The clusters with the most questions are 9, 2, 17, 3, as illustrated in the below picture of question frequency distribution:

9	10162
2	2649
17	1110
3	1101
12	1023
14	849
4	845
10	765
15	724
13	657
8	635
7	595
5	513
0	506
19	441
18	386
11	360
16	331
6	278
1	1

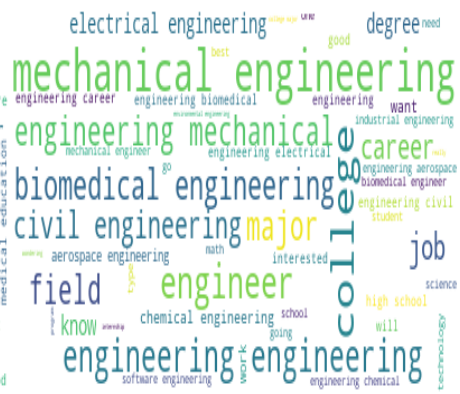
A wordcloud is created for each of these four clusters:

Cluster 9

Cluster 2



Cluster 17



Cluster 3

We can see very clearly from the wordclouds that the cluster 9 is mainly about career questions in college; the cluster 2 is about college major; the cluster 17 is about medical care; the cluster 3 is about engineering.

When we have a new question queried:

	questions_id	question_title	question_body	question_tag
0	dalfkjwlegjw1495uoriegh09hf3ofun36	how to become a successful Economicst?	I have recieved extensive academic training in...	Economics job

Its BOW will be first created and then appended to the existing question BOW (or a portion of it) and created a new tokenized and vectorized question BOW. The cluster id of the queried question will be predicted. The reason that we need to append the BOW of the queried question to that of the existing question (or a portion of it) is that the inverse document frequency, IDF, of the tokens in the queried question alone would be different than that of the tokens in the existing question BOW. So you have to at least give some corpus to the queried question to get the right TFIDF and thus to predict the correct cluster label.

After the cluster label of the queried question is predicted, 12 in this case, the cosine similarity between its BOW and the BOW of existing questions which have the same cluster id as the queried question is calculated. The ten existing question with the largest cosine similarity, sim_score, will be identified:

	questions_id	questions_title	questions_body	tags_tag_name	cluster_id	sim_score
5477	189d711122dc4718b64d0040f15812ac	What do I need to know to succeed in the field...	I'm planning on majoring in Economics and I'm...	career job skills economics knowledge	12	0.281496
16986	af81028cf7e14b42a0bfa918404e56a9	How hard is it to find a job after college?	I want to know the average how hard it was an...	jobs college graduate	12	0.217868
4824	79db1f06f09e48f082a059bad0e562f5	why is it so hard to find a job?	#job #career	career job	12	0.205694
2225	92dc5d56fa5b4129aa020e700e63d7fa	Can you become a CEO without a degree?	How to become a successful CEO without a colle...	job-search business ceo college	12	0.160849
17244	fd5f5364804247e8baa46df79da9d558	What kind of job can you get with a social wor...	Will a major in social work give me the educat...	volunteering social-work	12	0.156965
7231	85021058d28b494a8a65c7a1fa62fa17	Is it hard finding a job in college?	With school and studying, is it hard to find a...	jobs	12	0.141754
6160	e5e598ef1f7945f48cf26f87b0d054d1	Is academic research or internships better?	I am planning to go into aerospace engineering...	jobs internship research future	12	0.138386
22961	7fb1947aae67402d90eeb3918bf9c2d6	How hard is it to maintain a part-time job whi...	I'm always told that i'll have t work for coll...	job college	12	0.134389
15785	f8f5f47e5fbc431d8bb46c504ac1f5d7	What is the number one essential characteristi...	Is it personality? Is it the career itself? #...	college-jobs	12	0.117013
11983	87e052097d1a4b83b87d93f81e1c1861	How to write education in resume?	If i attend summer school or training in colle...	job-application resume-writing	12	0.116840

Professionals who answered these then most similar questions would be recommended to answer the query question. Below are the professional_id of these ten professionals.

```
596     066f10994b704178950d1645f8855247
595     36ff3b3666df400f956f8335cf53e09e
1858    b953b3a5cd564558a13b7bbe04dd0fd2
1856    3d8847ddc8d04185976a9e78b190ebdb
1857    23b6df4d73fb4d9bab3eafb6dd8bfb86
526     36ff3b3666df400f956f8335cf53e09e
238     f64df9b489864952917cb631be1ddac7
237     a1006e6a58a0447592e2435caa230f78
236     e3605142dffb4d87a0f8b950da727f73
1877    be5d23056fcb4f1287c823beec5291e1
Name: answers_author_id, dtype: object
```

Conclusion

Even though there is no way to evaluate the accuracy, performance and the response rate of this proposed recommendation engine in this stage (one has to actually implement it to test), my argument is that both the accuracy and speed performance will be greatly improved. First, this algorithm relies not only on the question-hashtag, but also on the whole BOW of question-title, question-body and question-hashtag to make recommendation. Therefore there is more information to be used to calculate the cosine similarity between the query question and the existing questions. Second, the use of k-mean cluster can greatly reduce the run time by saving the time used to calculate the cosine similarity between the queried question and the whole existing questions BOW. Instead, the cosine similarity only needs to be calculated between the queried question and the existing questions those are in the same cluster. The time cost of training the k-mean model is only about 45 seconds, which is a one-time cost and is much less compared with the query time that could be saved especially when the amounts of query booms.

Reference

An Intuitive Understanding of Word Embeddings: From Count Vectors to Word2Vec: https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/

Kaggle Kernel: https://www.kaggle.com/rdhnw1/triage-recommender-with-cold-start

Kaggle Kernel: https://www.kaggle.com/rblcoder/recommend-based-on-nearest-neighbors

Kaggle Kernel: https://www.kaggle.com/frissontek/recommendation-engine-for-career-village

Machine Learning A-Z, open courses in Udemy.

Rounak Banik, January 16th, 2018, Recommender Systems in Python: Beginner Tutorial

Shubham Jain, Februray 27, 2018: Ultimate guide to deal with Text Data (using Python) C for Data Scientists and Engineers: https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/

University of Michigan, Applied text mining with python, open courses in Cousera.

Wikipedia page for collaborative filtering: https://en.wikipedia.org/wiki/Collaborative_filtering