Data Science for Good

CareerVillage.org

AGENDA

- Business need
- Data visualization
- Data Preprocessing
- Model building
- Model evaluation

CareerVillage.org

CareerVillage.org

CareerVillage.org

CareerVillage.org

CareerVillage.org

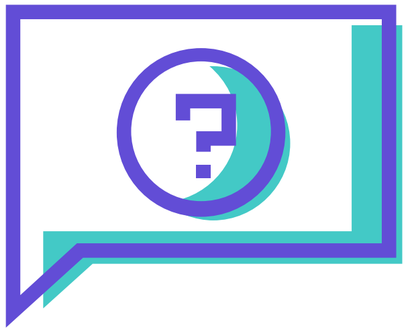
CareerVillage.org

3.5 M

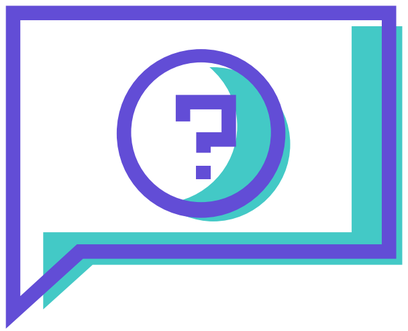
?

25,000

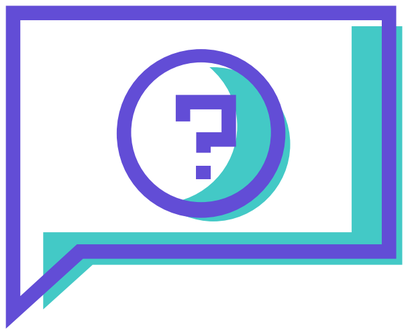
CareerVillage.org

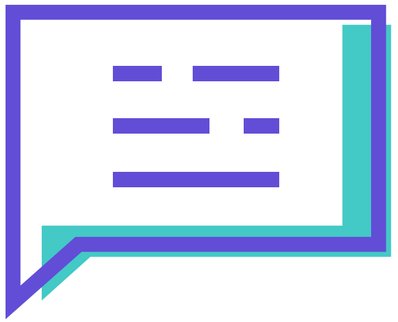


CareerVillage.org

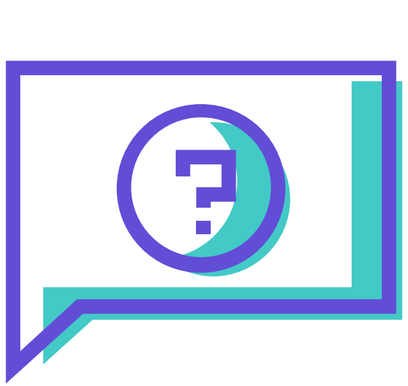


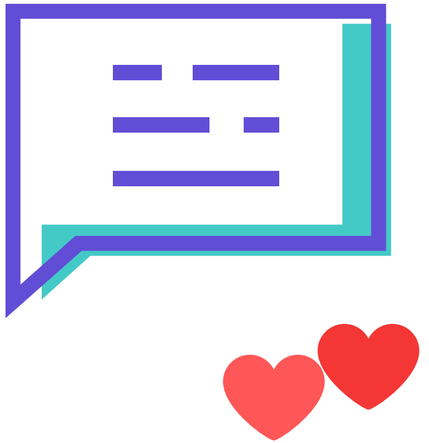
CareerVillage.org



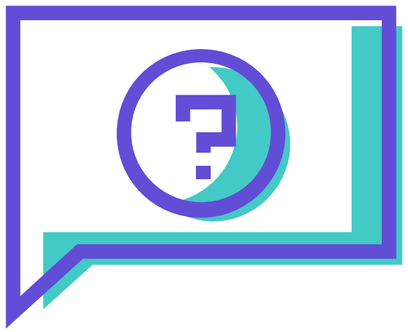


CareerVillage.org





CareerVillage.org

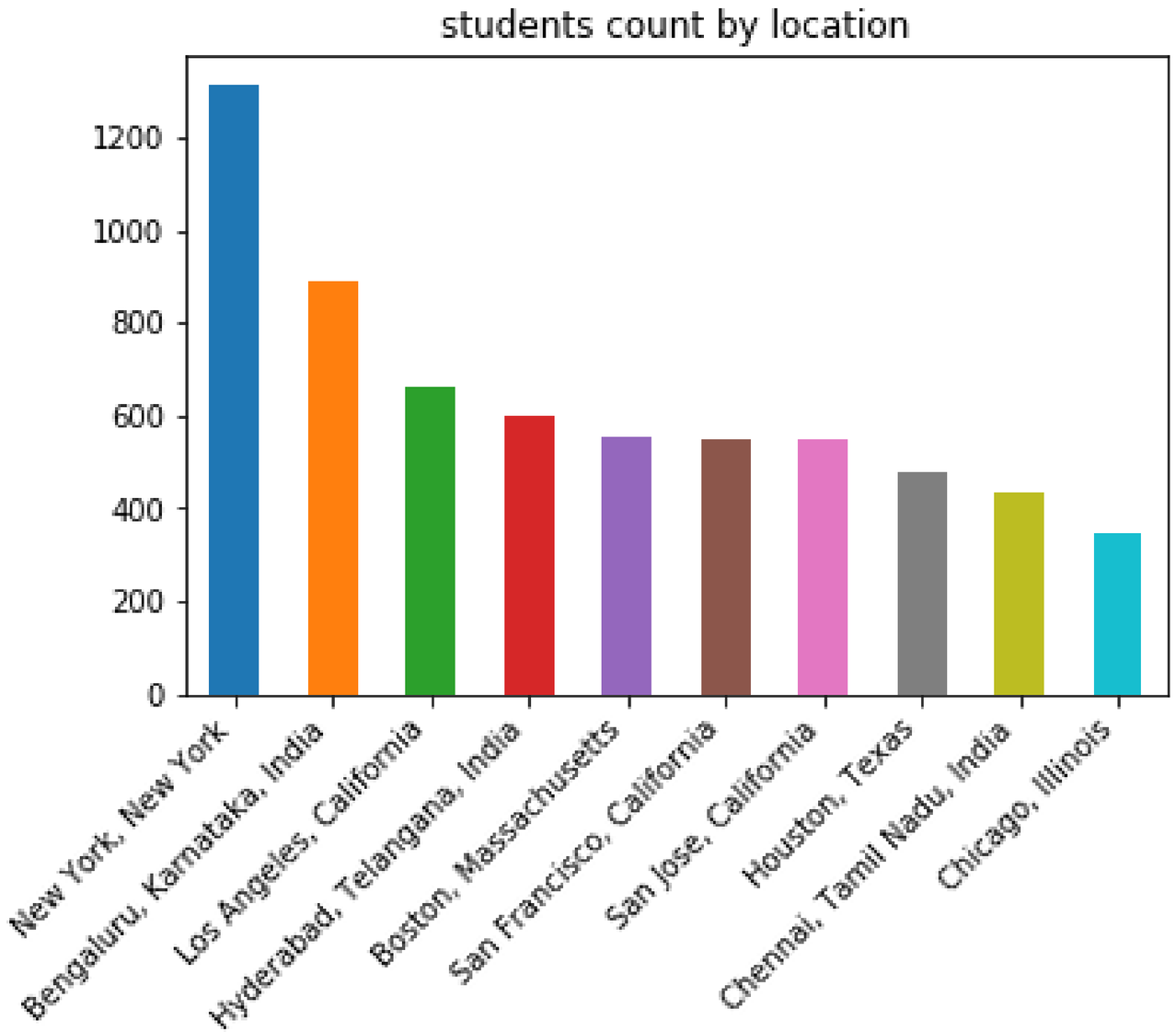


2

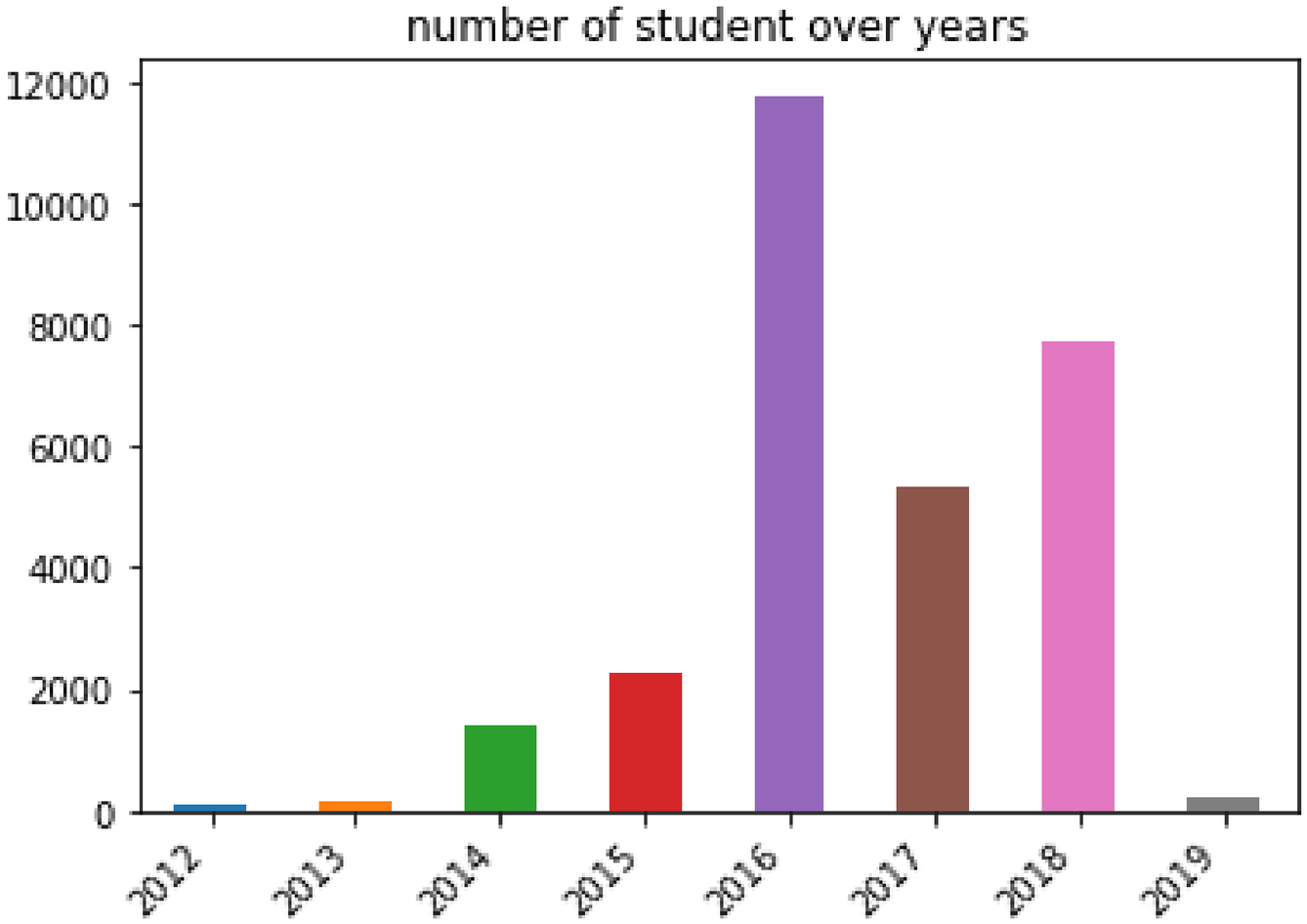
DATA VISUALIZATION

2.1

STUDENTS



New York is the largest region the students come from

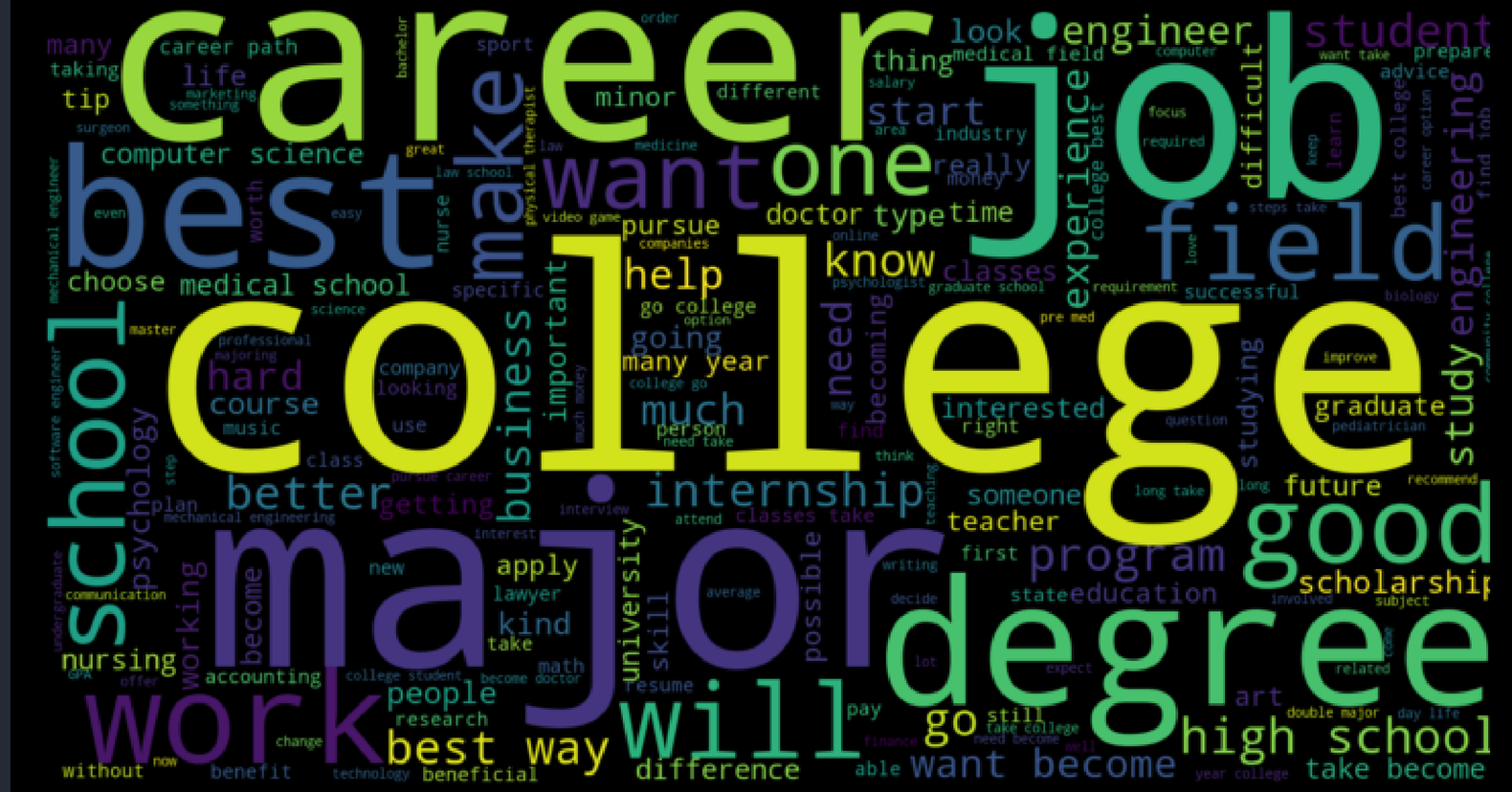


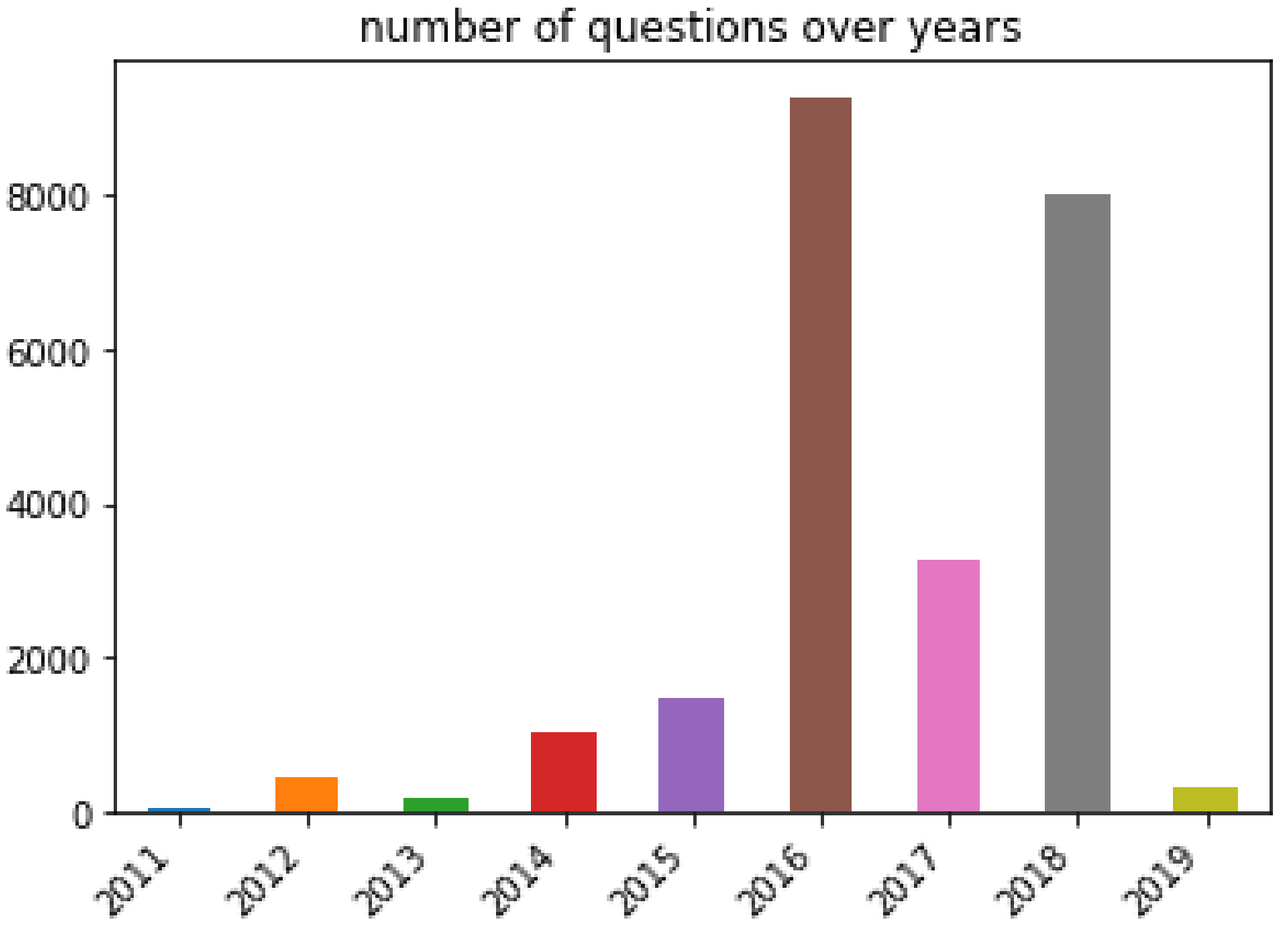
Seems that the community became most popular in 2016

2.2

QUESTIONS

Most of the questions' title about College , Career and jobs

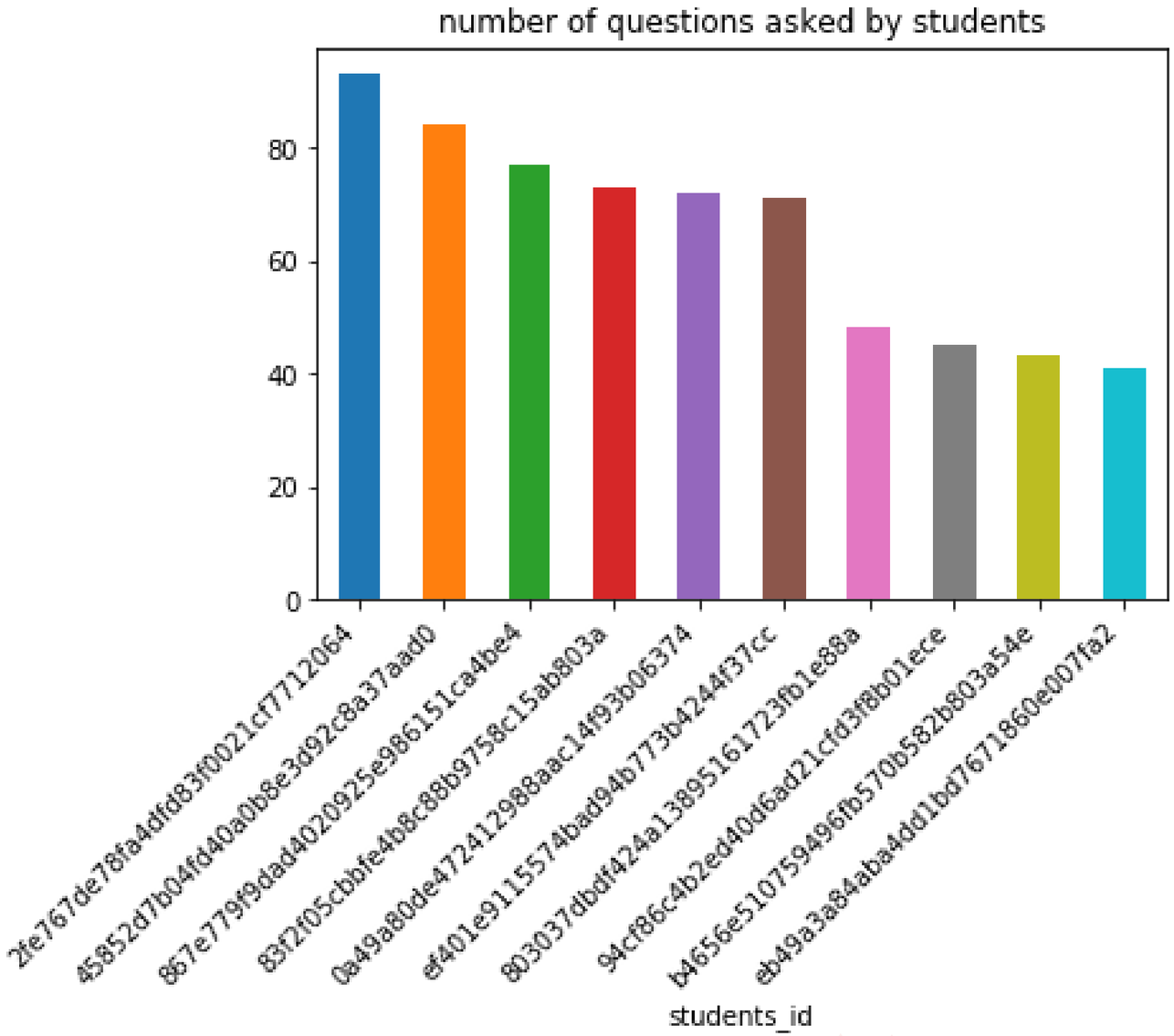




2016 is the year with the largest number of questions

2.3

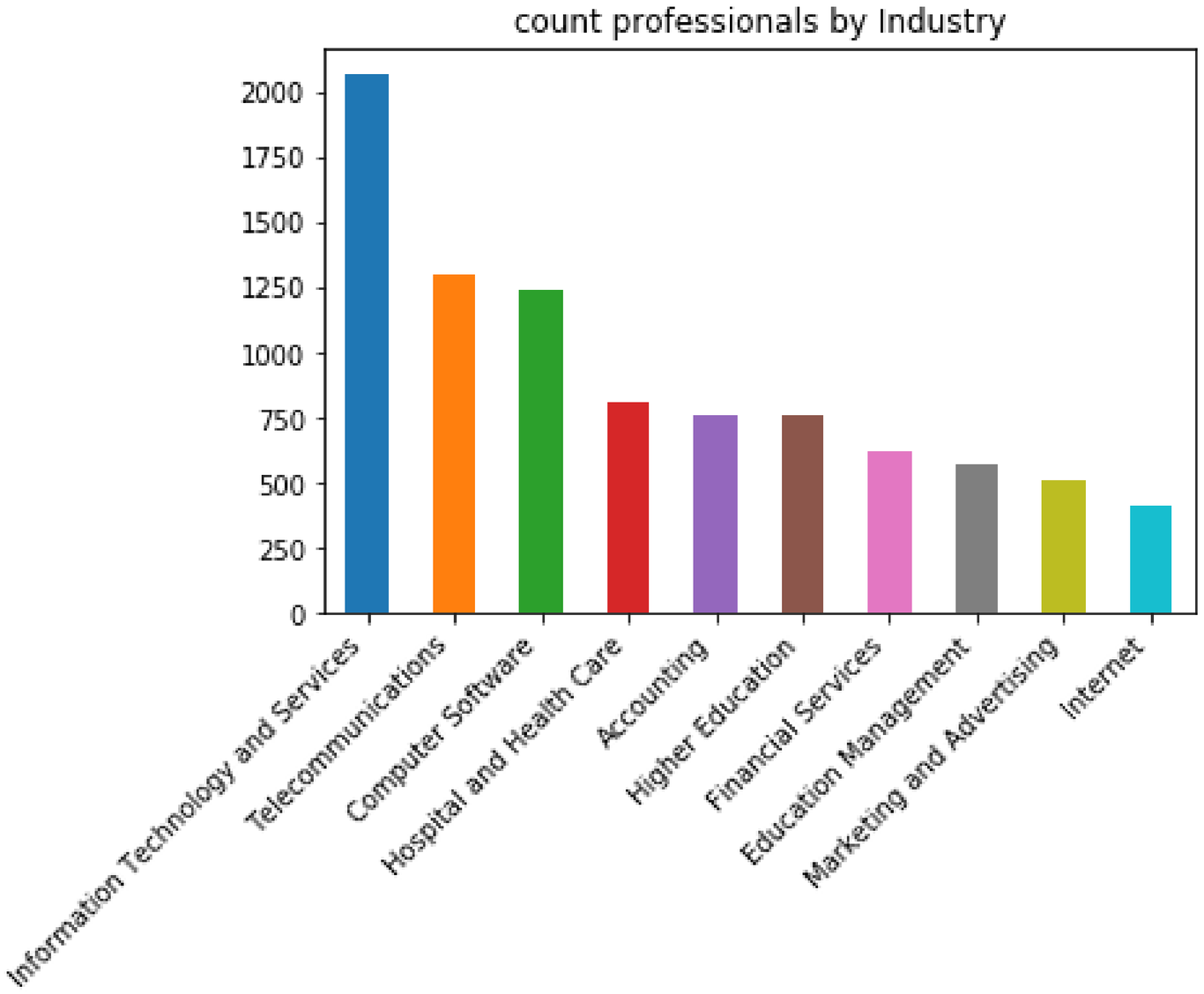
STUDENTS WITH QUESTIONS



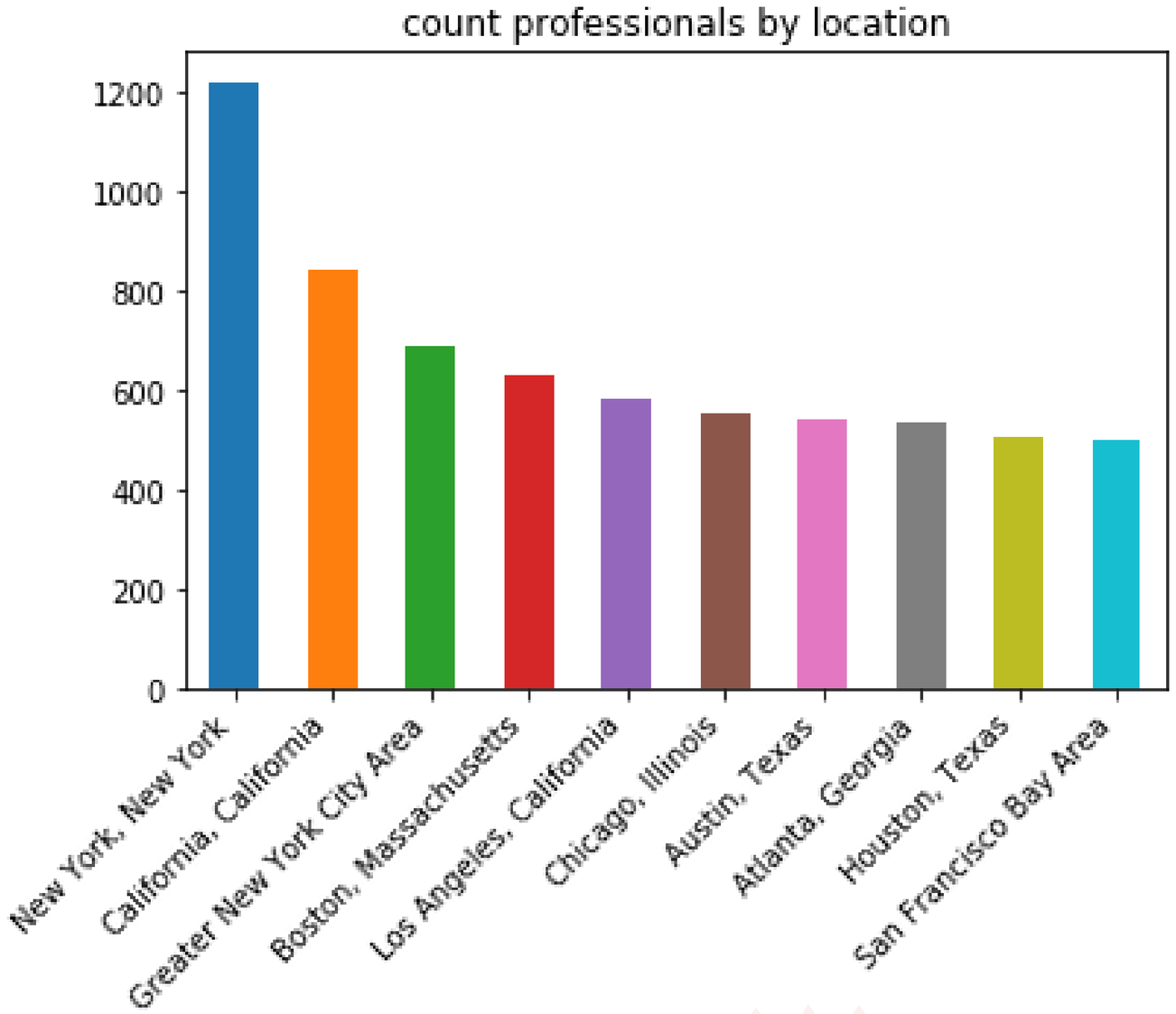
Who are the most active students ?

2.4

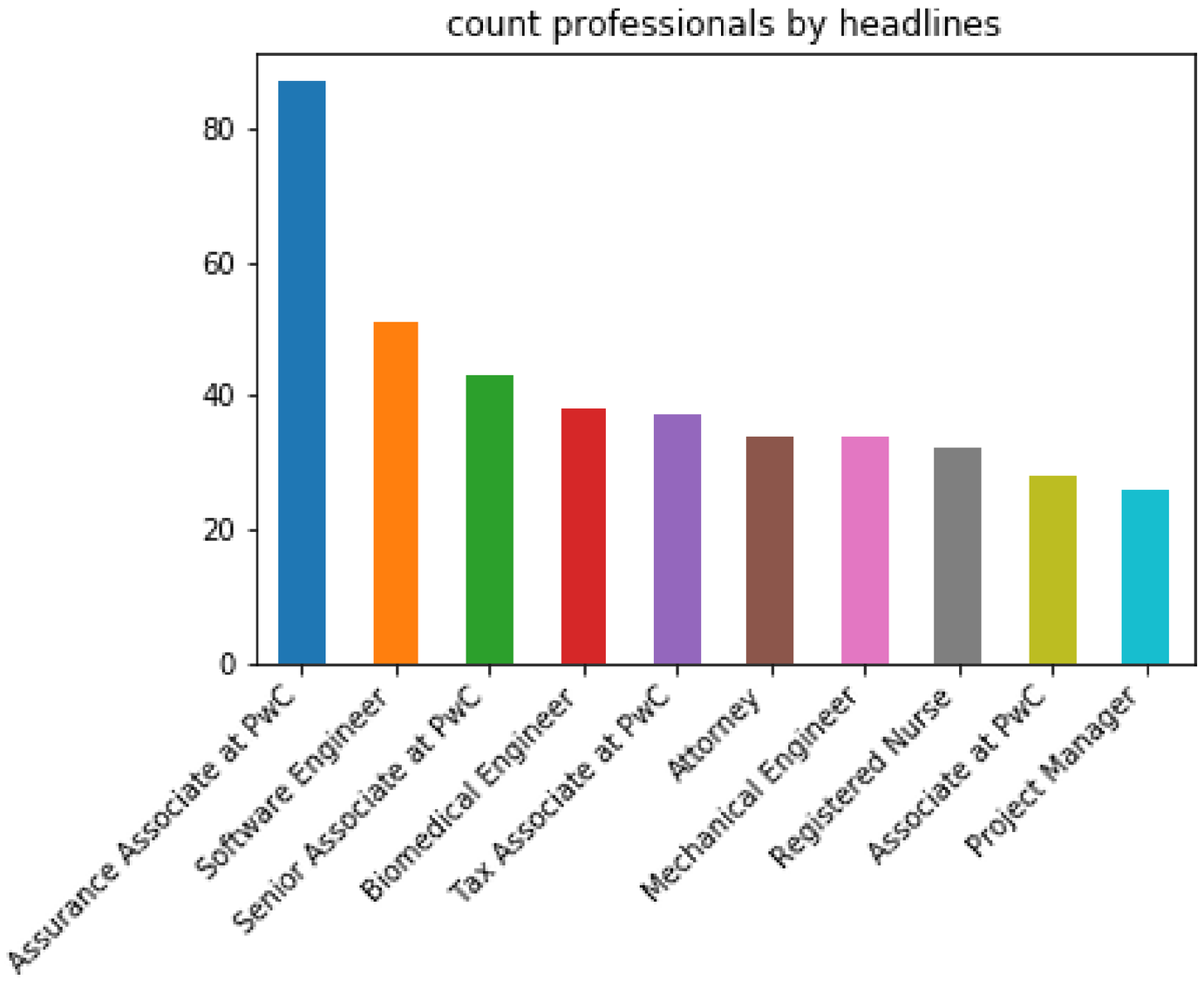
PROFESSIONALS



Information Technology and Services have the Largest number of professionals



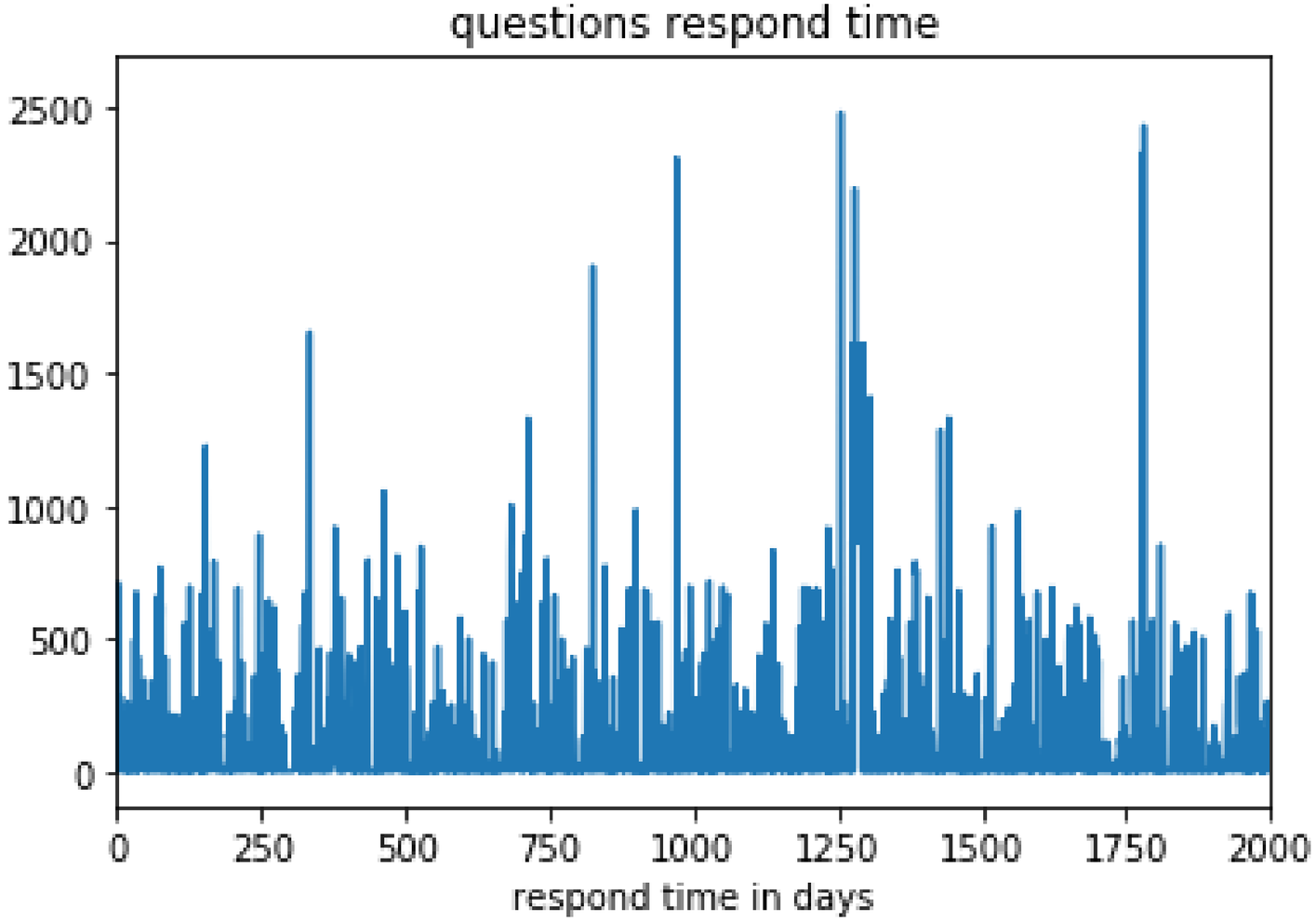
The Largest number of professionals from New York



Assurance associate at pwc and software engineer are the most frequent headlines

2.5

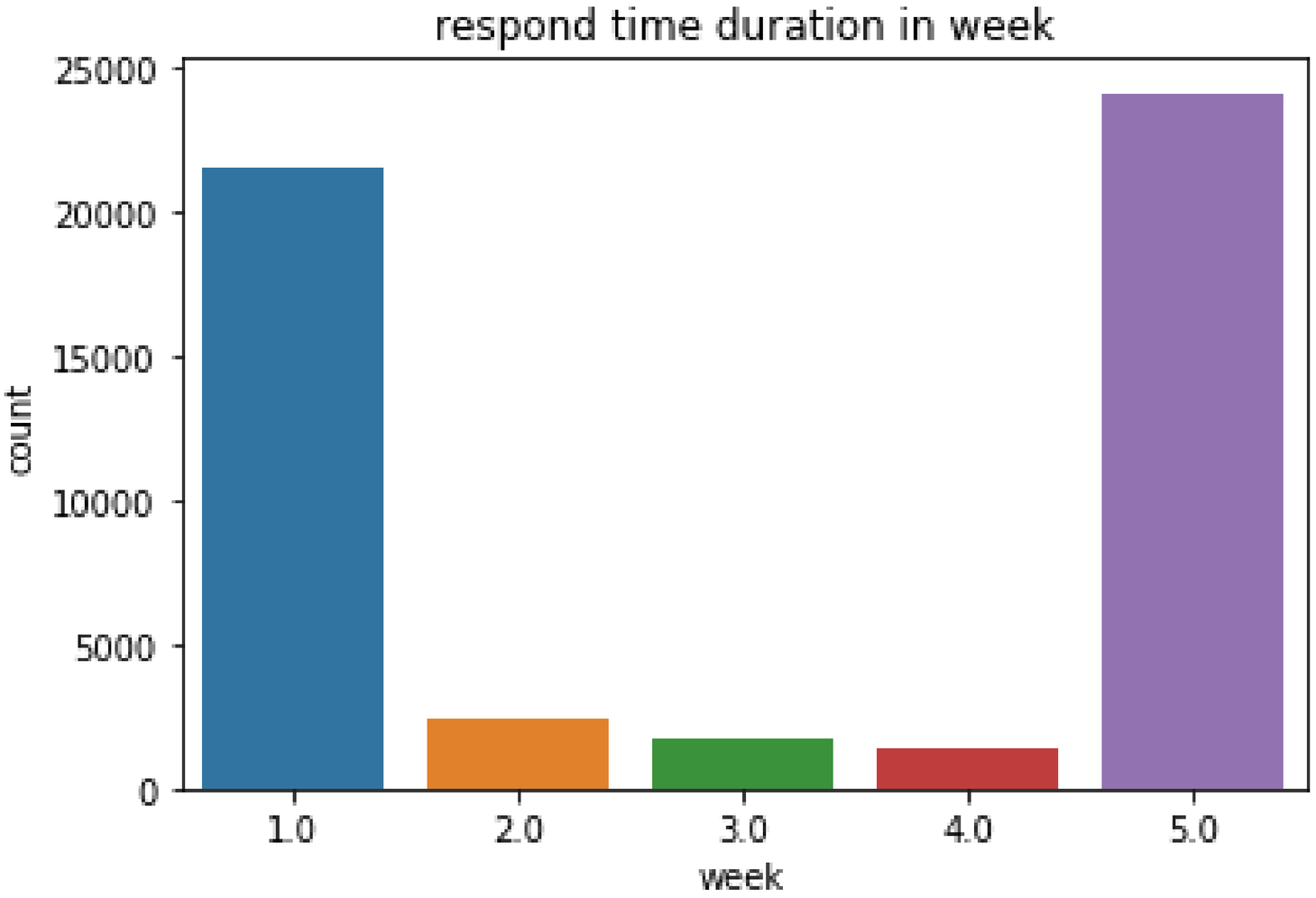
QUESTIONS WITH ANSWERS



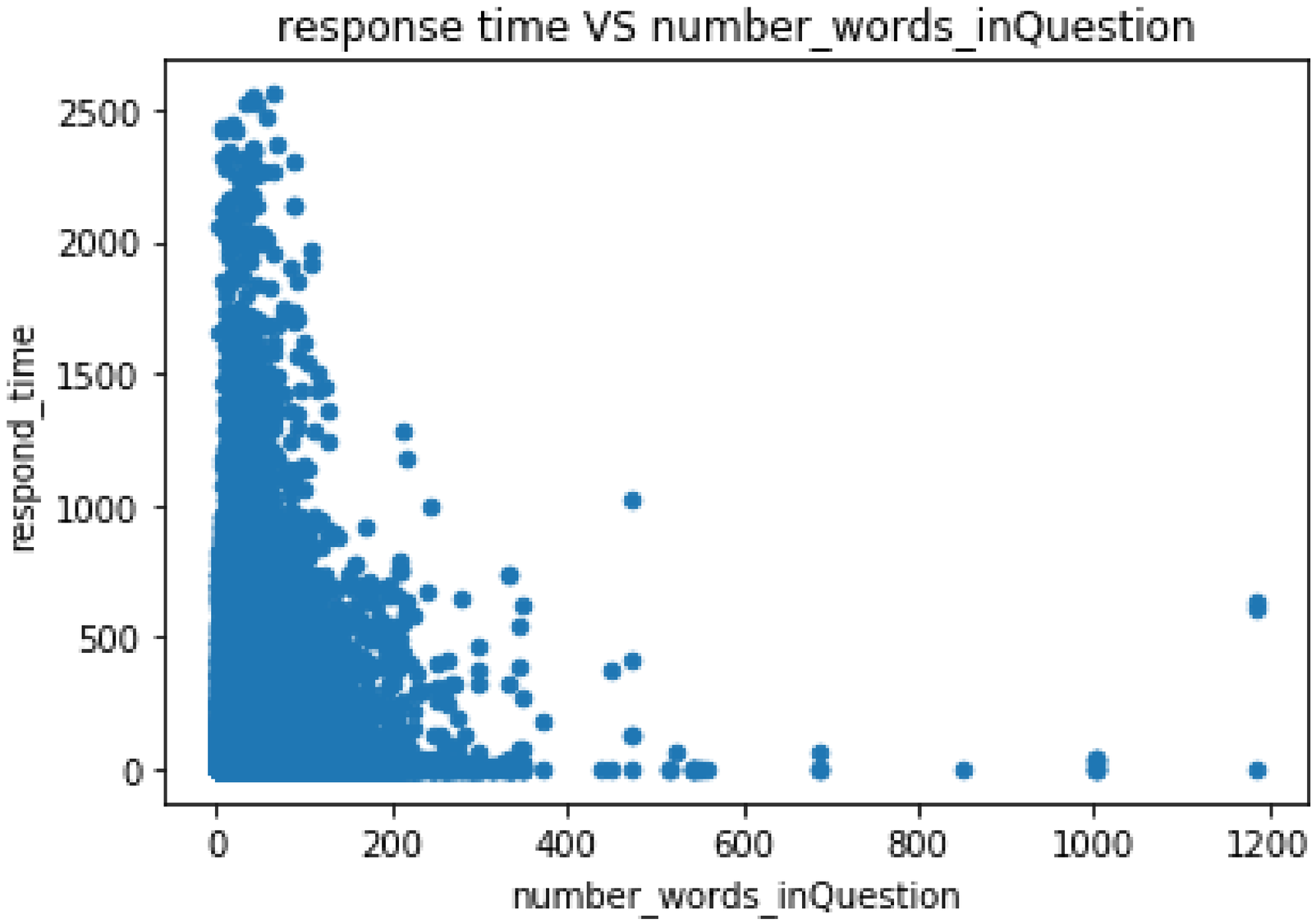
Response time of questions vary

But what makes a question be answered faster than the other ?

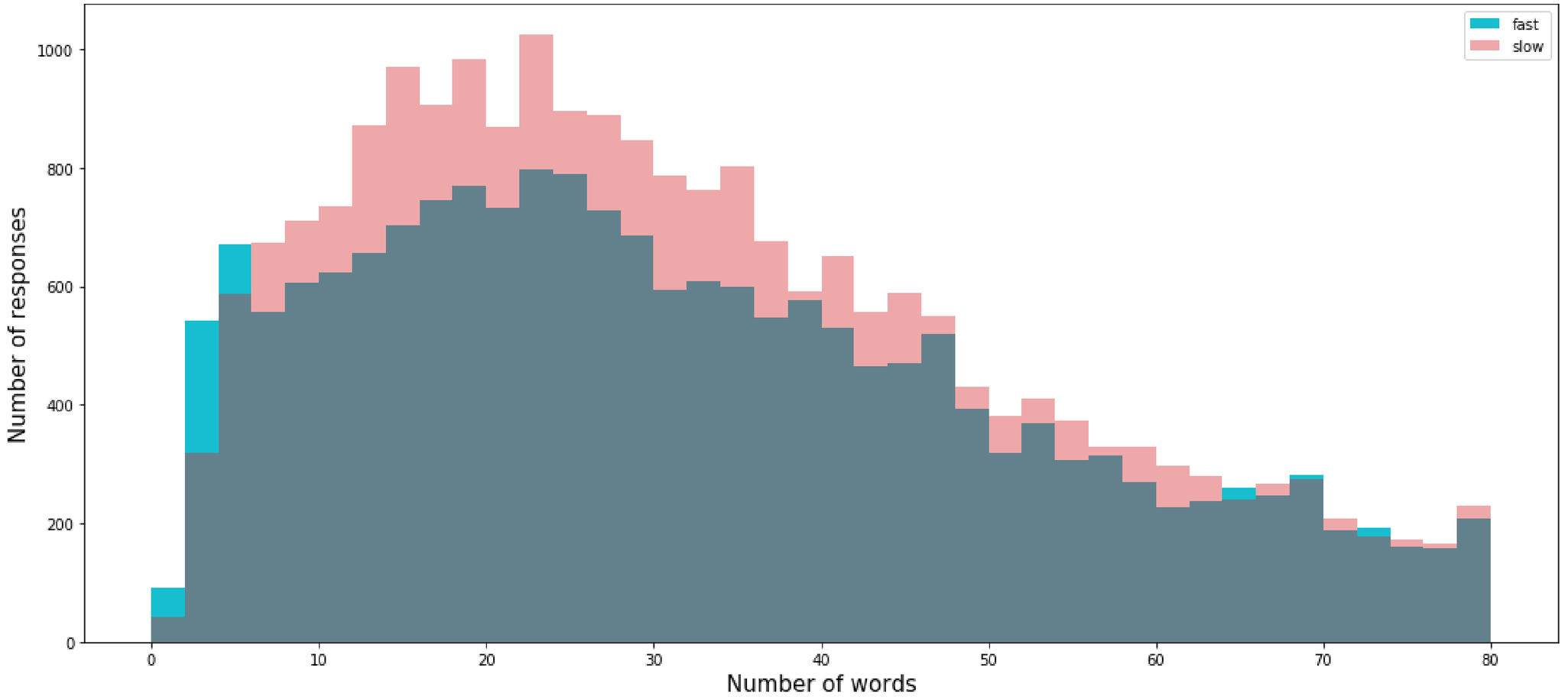
Is the length of question affects ?



24,092 questions took more than 1 month to be answered



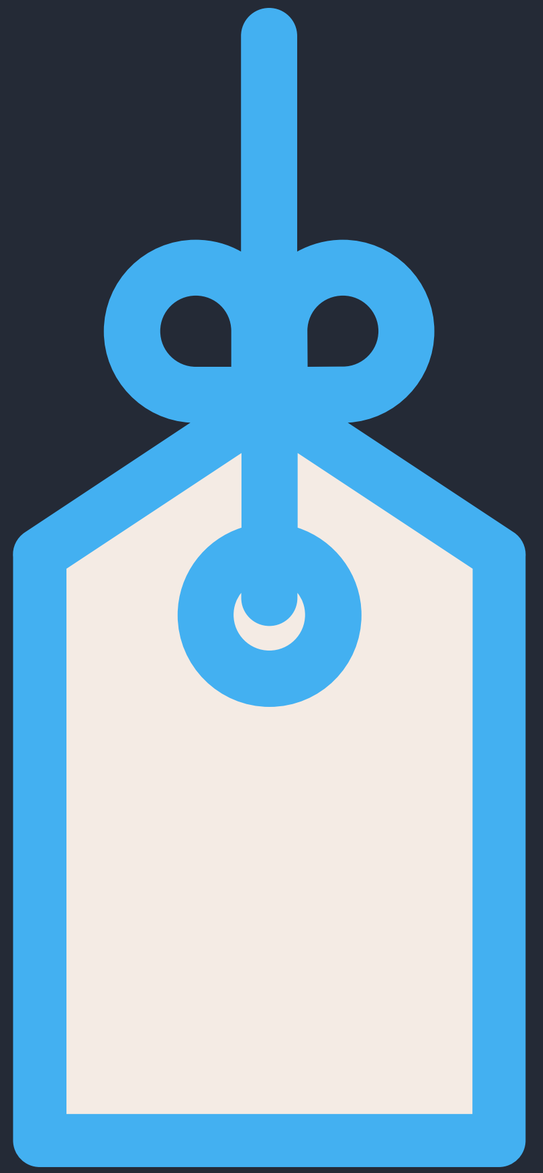
Seems there is no correlation at general view



Length of the question doesn't affect response time

2.6

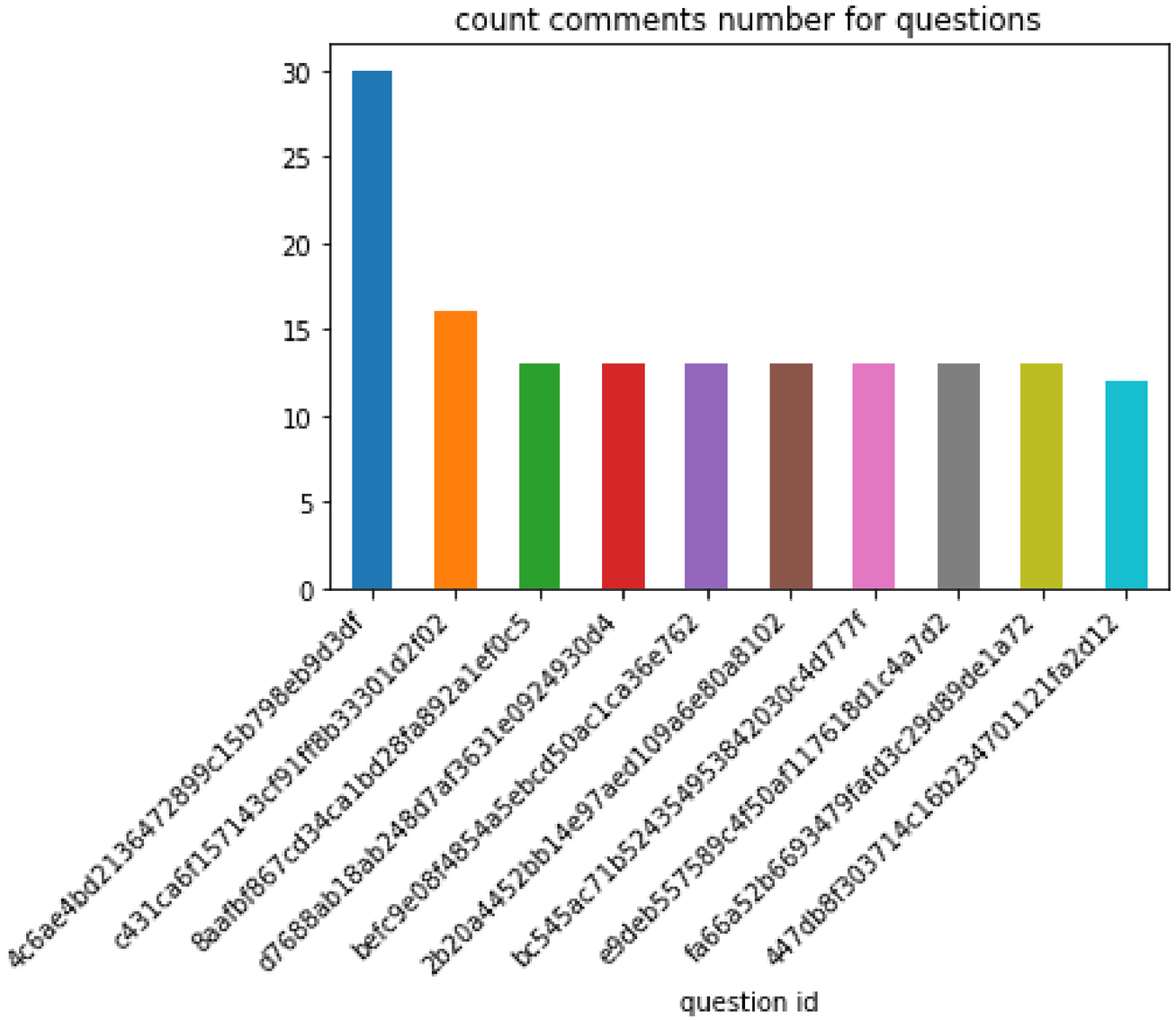
TAGS WITH TAG_USERS



Telecommunications , Information technology and College are the most followed Hashtags

2.7

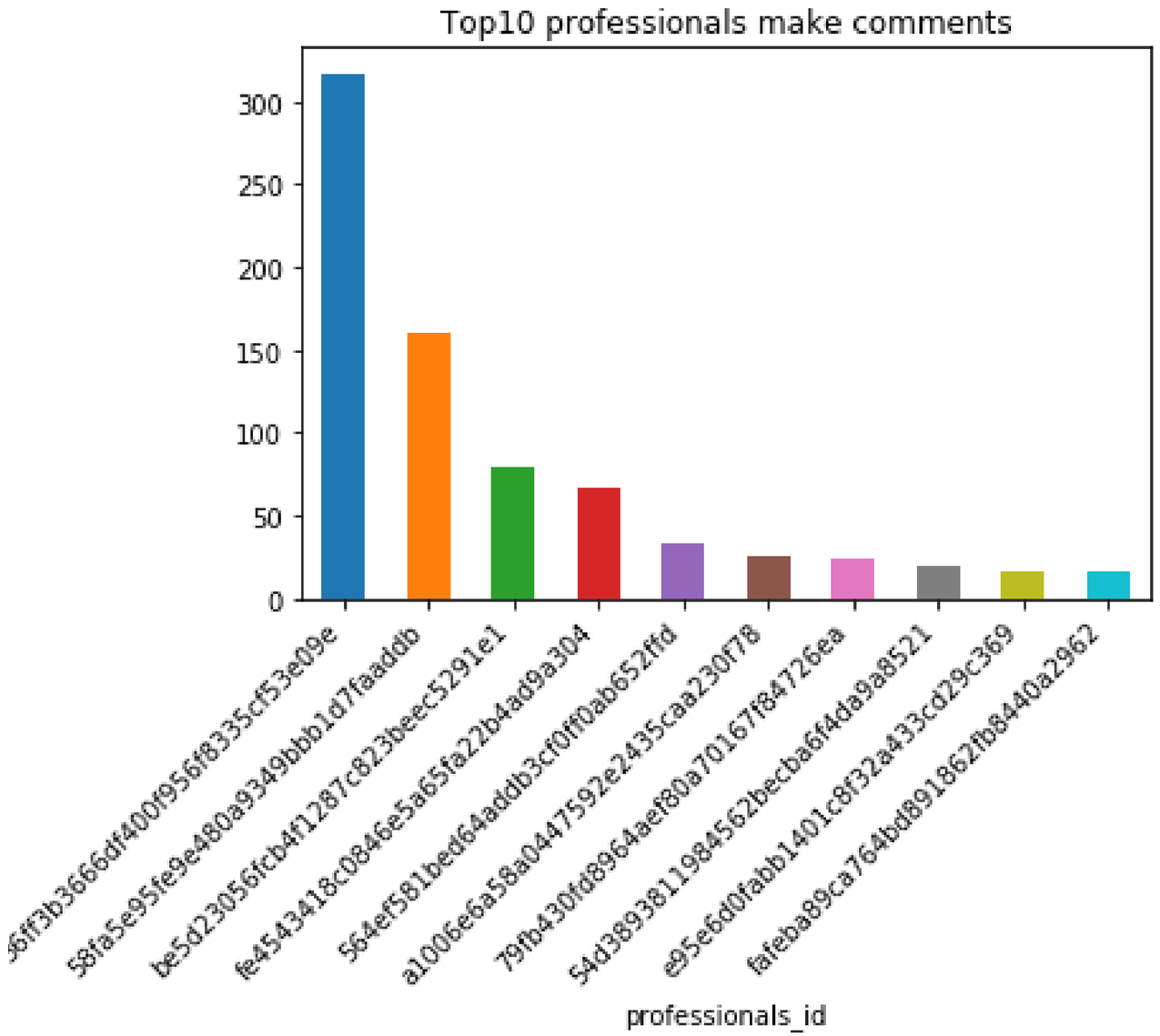
COMMENTS



What are the questions with the largest number of comments ?

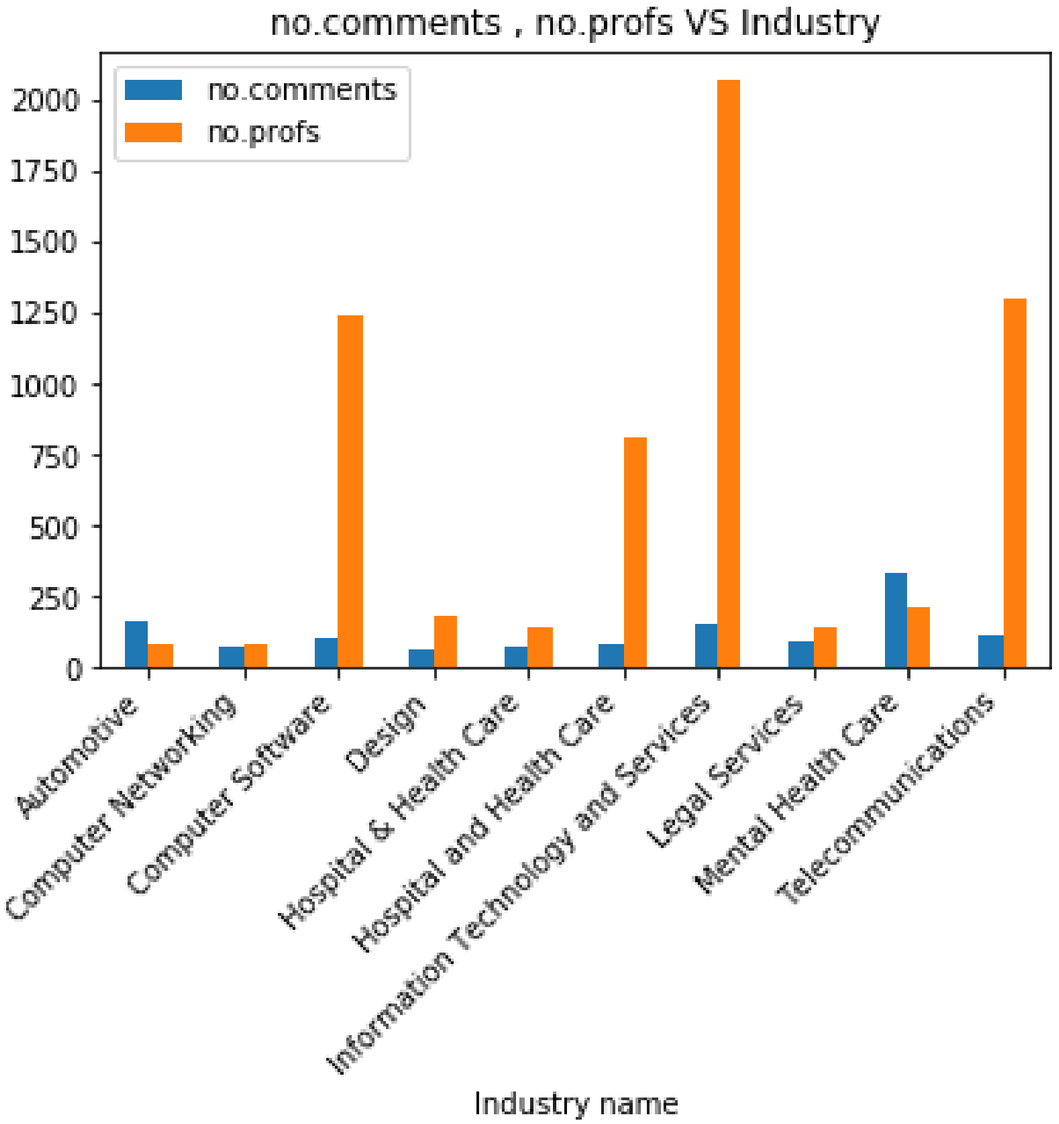
2.8

PROFESSIONALS WITH COMMENTS



Most active professionals in comments

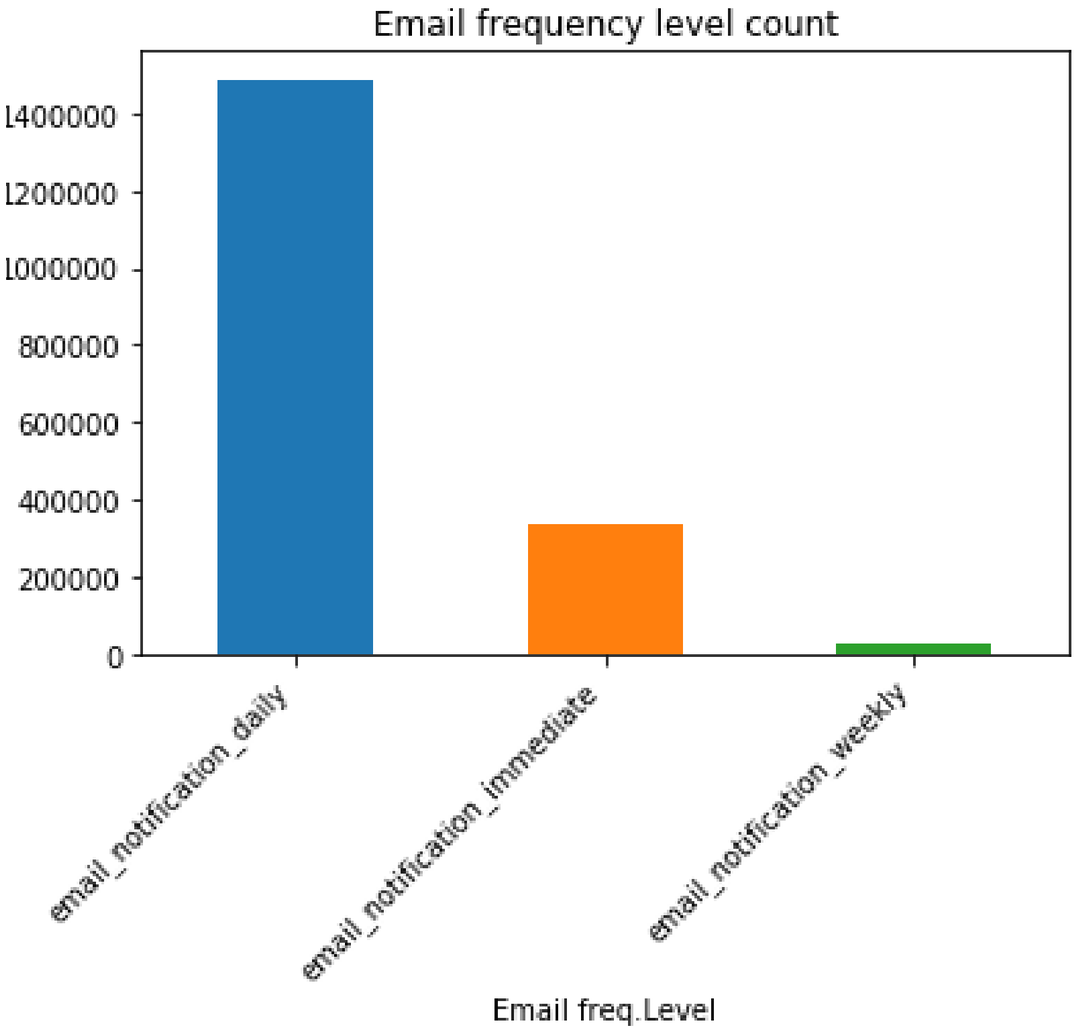
Is there a relations between number of professionals in an industry and number of comments in this industry ?



There is no relation between number of comments and professionals in the industry

2.9

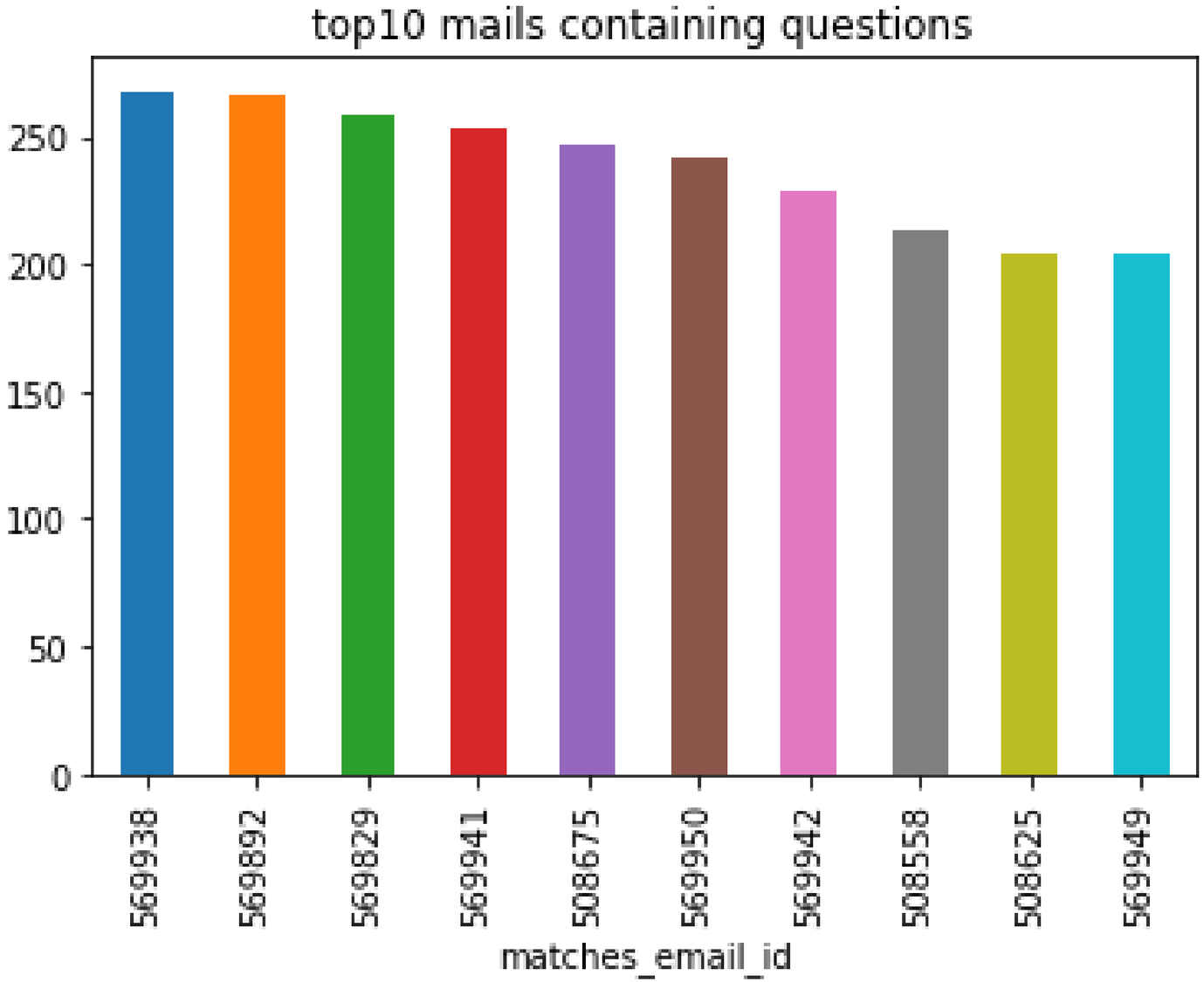
EMAILS



Most subscribers have daily notification

2.10

MATCHES



What are the top 10 mails containing questions ?

3

DATA CLEANING

Deleting Nan's

- **STUDENTS**

  2033 rows

- **PROFESSIONALS**

  5248 rows

Deleting unexpected values

- **PROFESSIONALS**

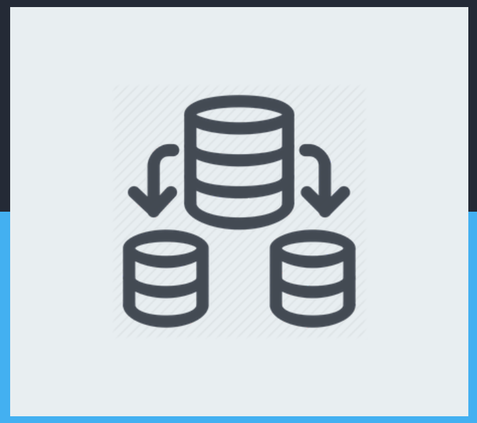
  167 professionals without headline

4

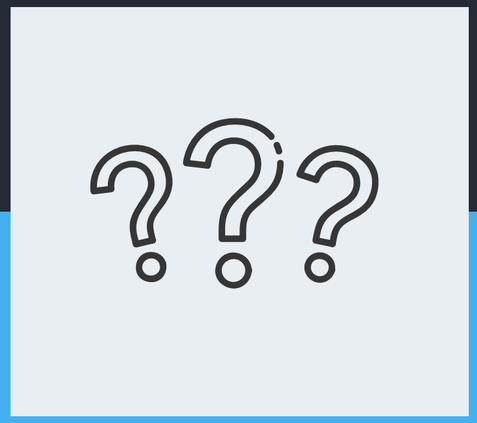
MODEL BUILDING

4.1

CONTENT BASED FILTERING

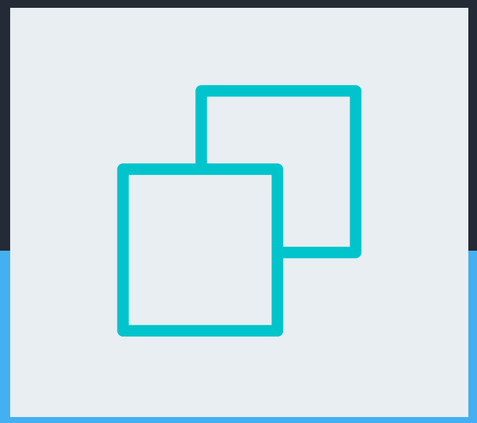


split data



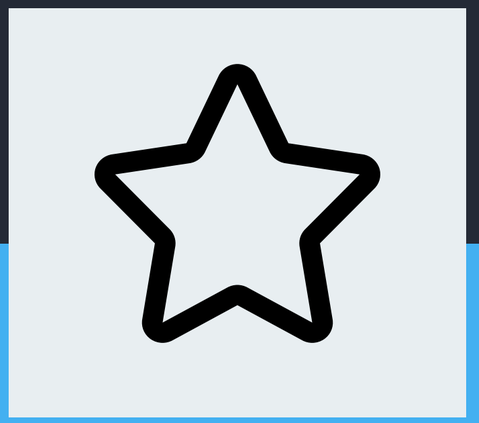
TF-IDF for all questions

TF-IDF for input question



Cosine similarity

Professionals answered similar questions

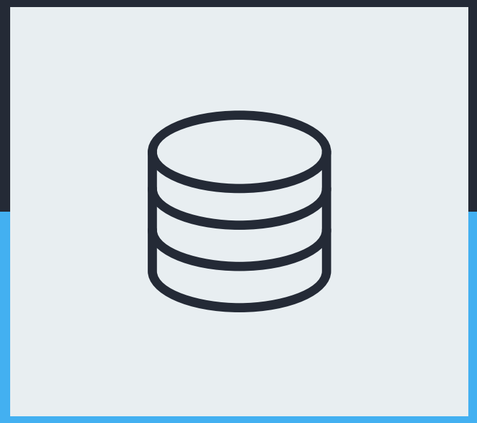


score based on activity and similarity

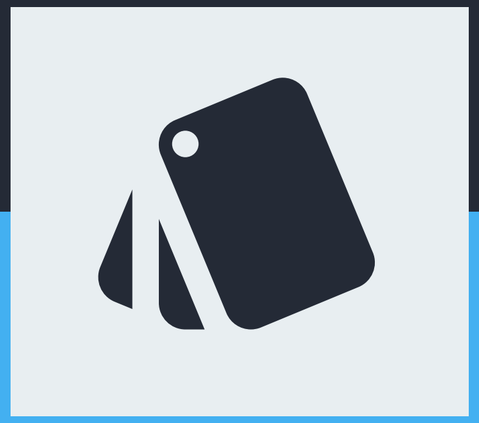
	answers_author_id	questions_id	Active	scores	Ques_ID	FinalScore
21	84bad21569ab43a7aab8e5585c4dfdbe	ed2c4ae96c24444cba1905a4c258eb91	True	0.576793	ed2c4ae96c24444cba1905a4c258eb91	4.883963
28	a7730a0f42cb4cd6805cecc03ec2df0a	ed2c4ae96c24444cba1905a4c258eb91	True	0.576793	ed2c4ae96c24444cba1905a4c258eb91	4.883963
43	eefd04d904cb4be7a441ec8bb2af84ba	38ef42d4cde74d8a9d52553e6b3d7455	True	0.343015	38ef42d4cde74d8a9d52553e6b3d7455	3.715075
29	b1f35632ef9f4cab95b12b9714e20c21	38ef42d4cde74d8a9d52553e6b3d7455	True	0.343015	38ef42d4cde74d8a9d52553e6b3d7455	3.715075
7	475a30894f03441e9431efad636a0365	c7b36b048401449995a9719b0b81099c	True	0.315523	c7b36b048401449995a9719b0b81099c	3.577617
38	d67ce930870945109a7ad86d29ba2035	e455abf42c4541cbb8b96eaa84908f41	True	0.281295	e455abf42c4541cbb8b96eaa84908f41	3.406473
16	70ce93bc0d2a43278a66e421604ea6e9	3e095e1c78a24c5d9616b095b39ea640	True	0.266740	3e095e1c78a24c5d9616b095b39ea640	3.333702
41	e3529e64e70643b6889353db79eeea3e	c6319b893390467d86c0477382f138c3	True	0.259353	c6319b893390467d86c0477382f138c3	3.296763
9	49b75499b16d4c888bc8850eced6a92e	9909cdbebb5c4b9ea585f405e28bacae	True	0.253289	9909cdbebb5c4b9ea585f405e28bacae	3.266445
46	fc8675180eee4e2985bb2ccfaf712241	f69e88c7cb104e56bf4c6c76601afd06	True	0.253062	f69e88c7cb104e56bf4c6c76601afd06	3.265310
39	d7f9afe721af42b1a03a993909e0568c	f69e88c7cb104e56bf4c6c76601afd06	True	0.253062	f69e88c7cb104e56bf4c6c76601afd06	3.265310
1	1713e8b9fe3b471e84567b5c0a2c4b45	f69e88c7cb104e56bf4c6c76601afd06	True	0.253062	f69e88c7cb104e56bf4c6c76601afd06	3.265310

4.2

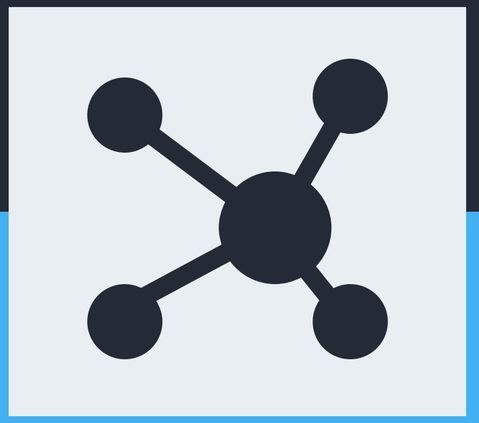
COLLABRATIVE FILTERING



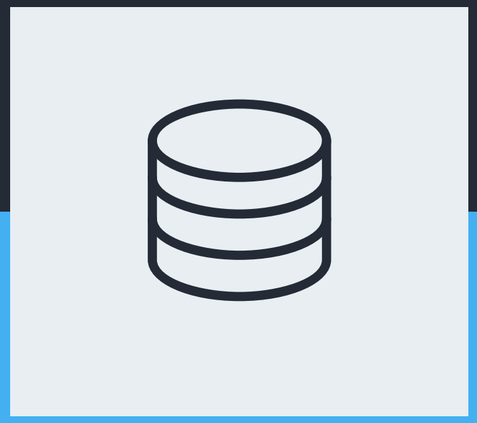
professionals, answers, tags



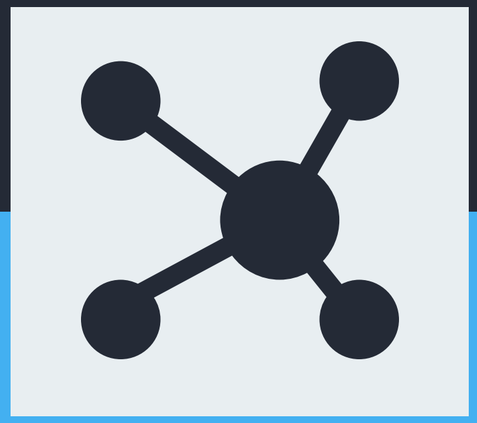
TF-IDF for all tags



K-means clustring



suggested professionals



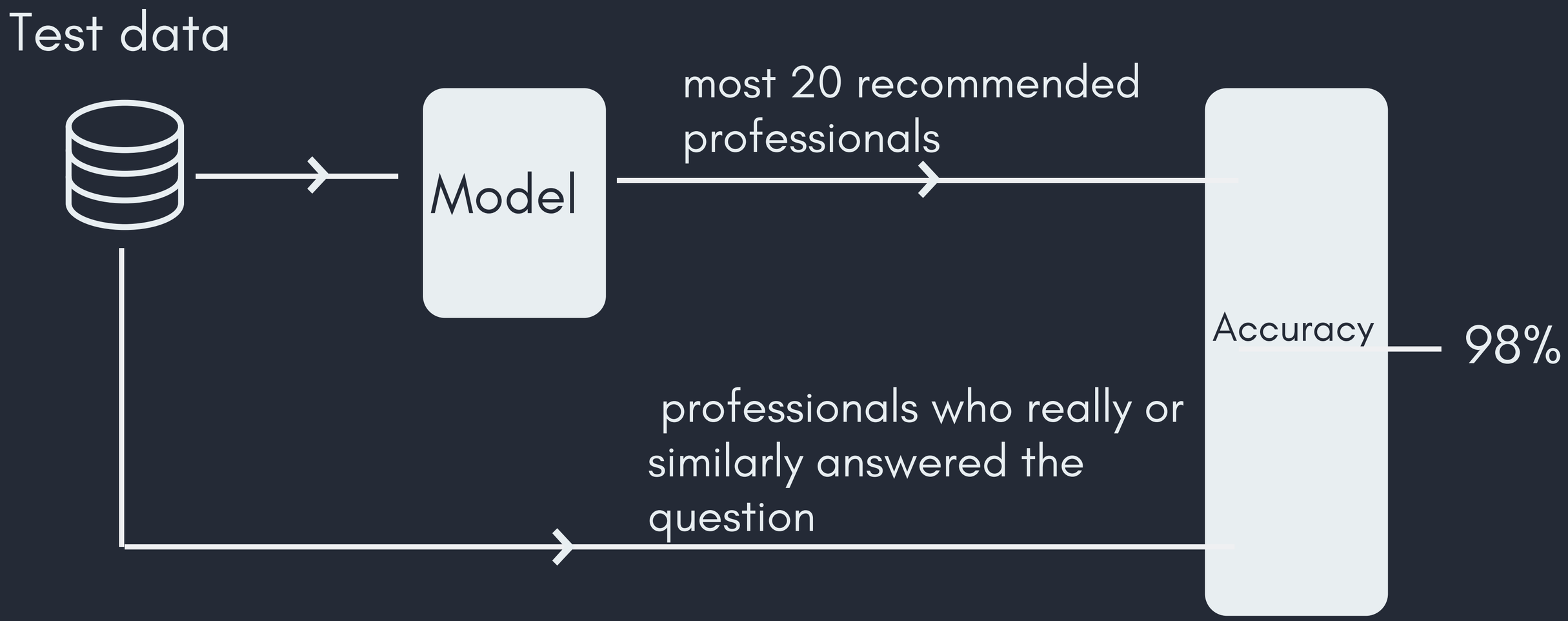
get each professional cluster

most 10 similar professionals

	Professinal_id	occurrence
0	36ff3b3666df400f956f8335cf53e09e	47
0	a1006e6a58a0447592e2435caa230f78	47
0	58fa5e95fe9e480a9349bbb1d7faaddb	47
0	369f1c8646b649f6997eae7809696bd5	47
0	be5d23056fcb4f1287c823beec5291e1	43
0	05ab77d4c6a141b999044ebbf5415b0d	43
0	a6d33c38902546849c36ea7e9e9f0870	43
0	d67ce930870945109a7ad86d29ba2035	28
0	4dc61581ec7b409bbd037e483f53ba0a	23
0	c3b4e11154f74a858779be7ba9b6f00c	21

5

MODEL EVALUATION


Test data
Model
most 20 recommended professionals
professionals who really or similarly answered the question
Accuracy
98%


Thanks