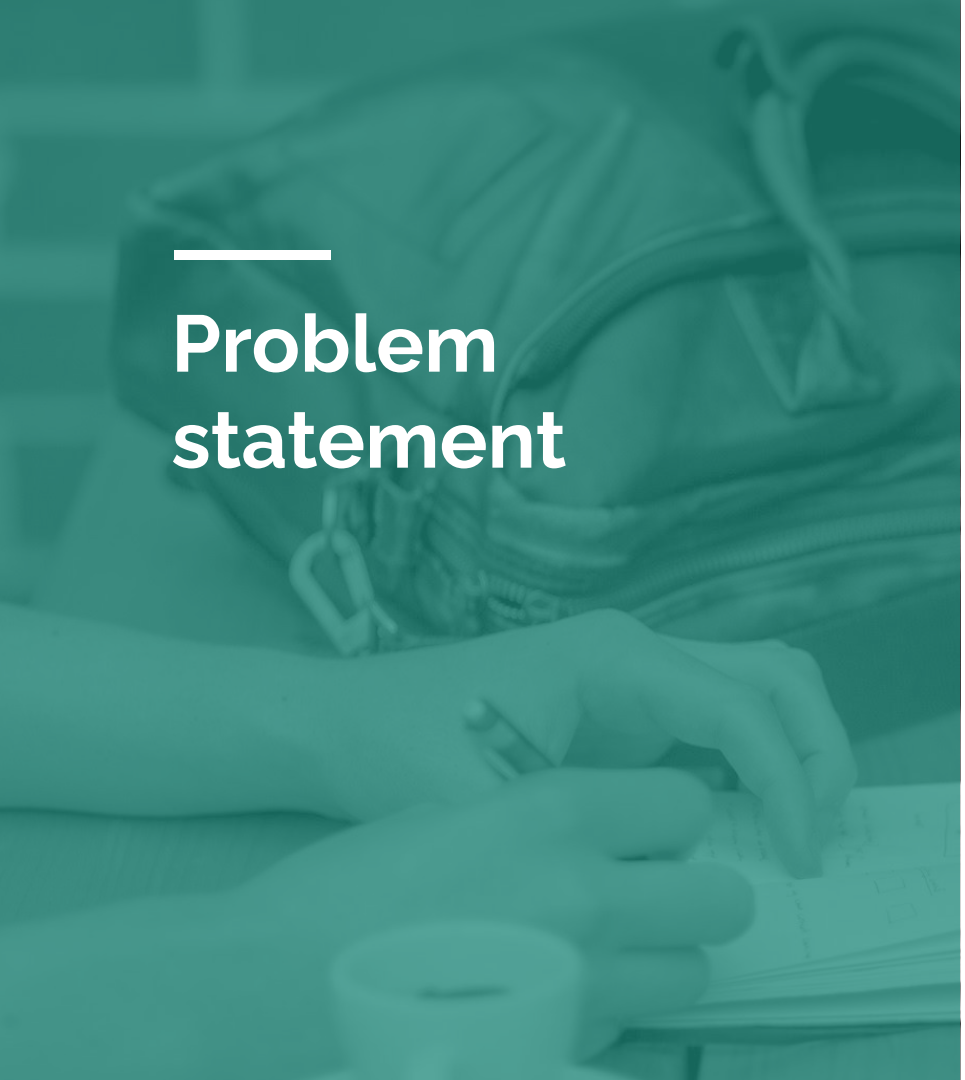
P3: Subreddit scraping and analysis
(Web APIs and NLP)

Analysing corpora in r/careerguidance **against** r/jobs

Models used: MultinomialNB,

Wordclouds are good decorations, let's keep it that way. :-)
(you will only see them here)

Special thanks to Reddit and redditors in the abovementioned subreddits, and credits to Balsamiq Wireframes for this Google Slides template



Problem statement

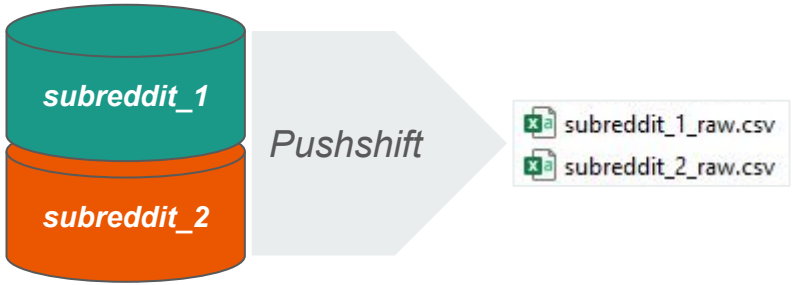
As a Redditor and data aficionado, train a classifier to tell two sub-reddits apart using evaluation metrics (below):

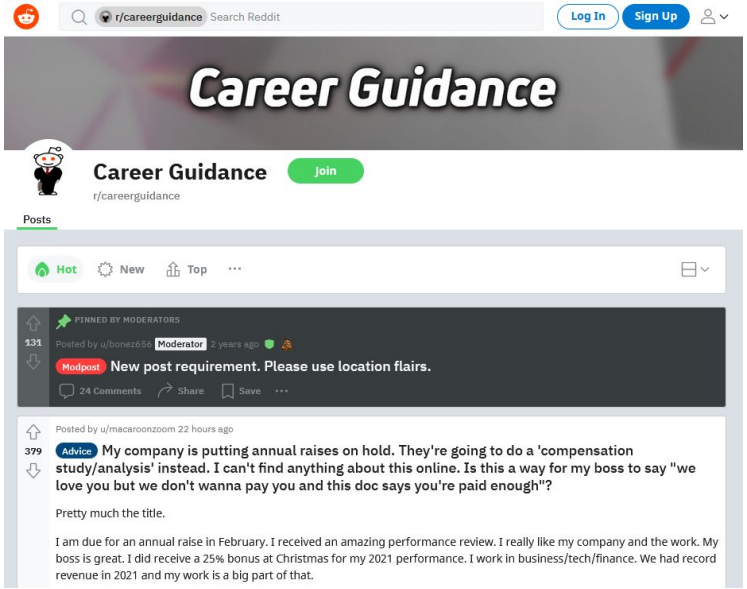
- *r/jobs*, which focuses more on immediate issues and getting a job, and
- *r/careerguidance*, which focuses more on longer-term (career) decisions.

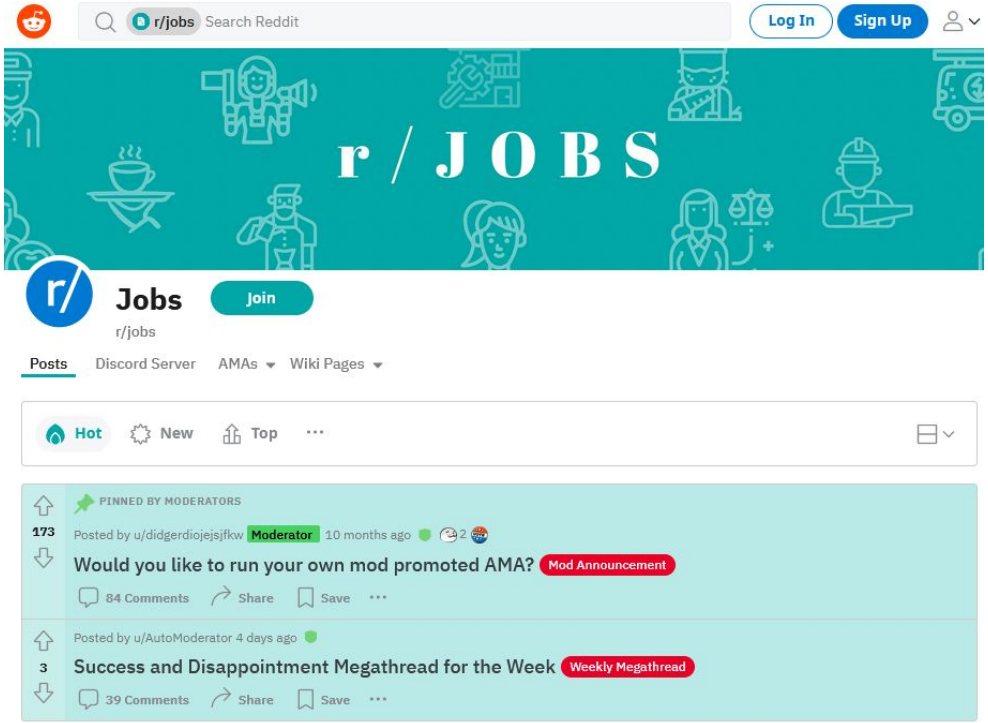
Administrators and moderators for these subreddits have a partnership to develop a feature to suggest to users to in which subreddit to post their content. This feature will simply take their content (*selftext*) and evaluate where their post belongs

Methodology and Workflow

1. Scrape data using Pushshift API



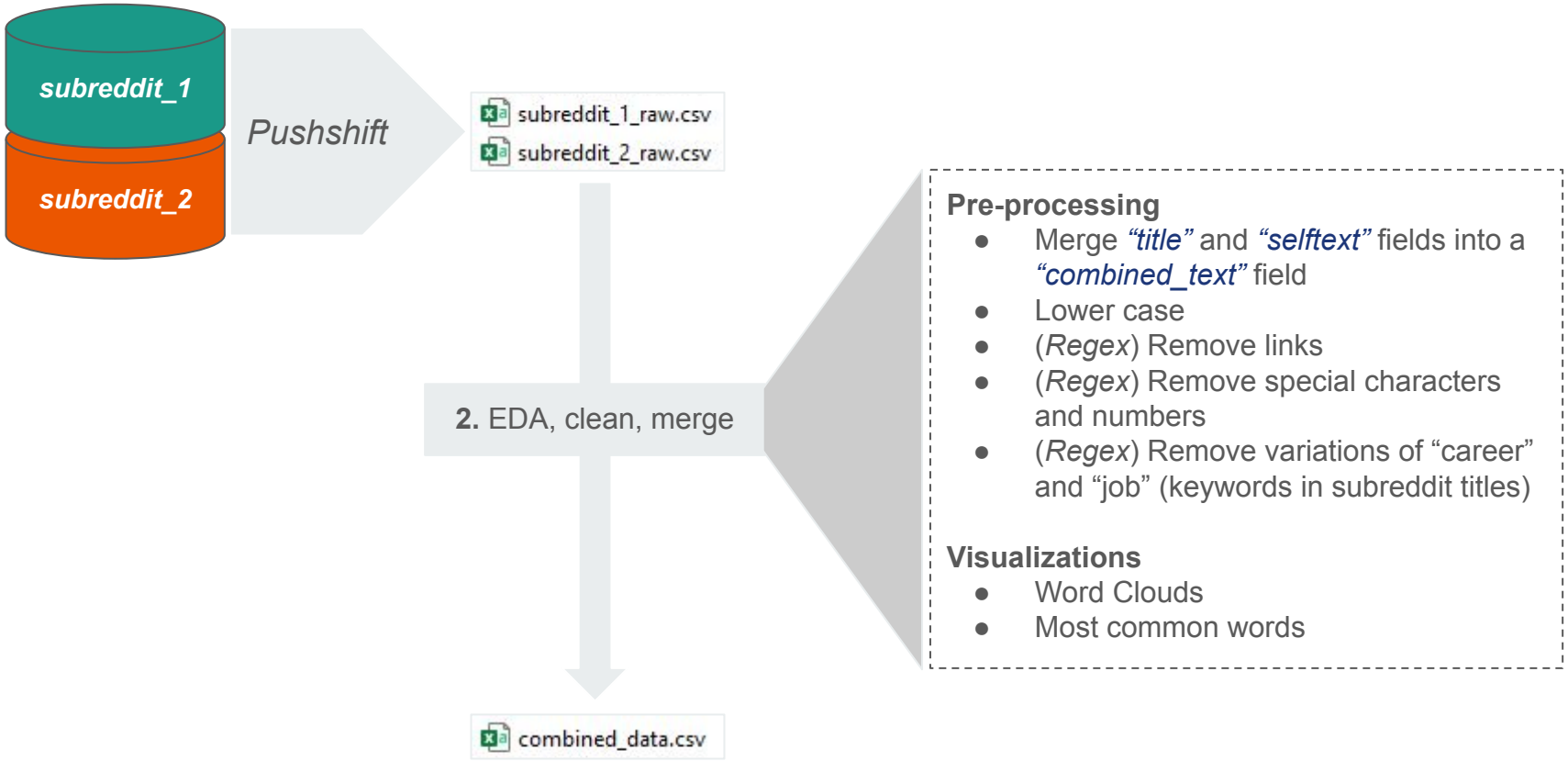




Source: www.reddit.com

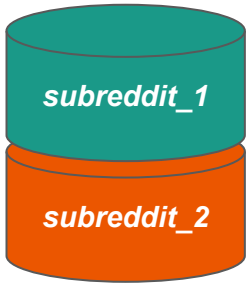
Methodology and Workflow

1. Scrape data using Pushshift API



Methodology and Workflow

1. Scrape data using Pushshift API



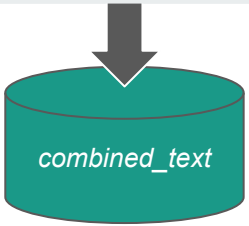
Pushshift

subreddit_1_raw.csv
subreddit_2_raw.csv

2. EDA, clean, merge

combined_data.csv

3. Iterative modelling, hyperparameter tuning and production



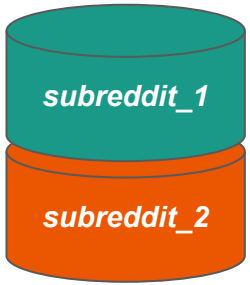
[cvec, tvec]

[nb, rf]

Choose Model

Methodology and Workflow

1. Scrape data using Pushshift API



Pushshift

subreddit_1_raw.csv
subreddit_2_raw.csv

2. EDA, clean, merge

combined_data.csv

```
Performing grid search...
pipeline: ['cvec', 'rf']
parameters:
{'cvec__ngram_range': ((1, 1), (1, 2), (1, 3)), 'cvec__min_df': [0.01, 0.1, 0.5], 'cvec
__max_df': [0.1, 0.5, 0.9], 'cvec__max_features': [None, 5, 20, 50, 100], 'rf__max_dept
h': [1, 5, 10, 20], 'rf__n_estimators': [1, 2, 4, 8, 16, 32, 64, 100, 200]}

C:\                                                    _search.py:918: UserWar
ning: One or more of the test scores are non-finite: [0.51742907 0.5265209  0.52864487
... 0.58622982 0.58653286 0.58683589]
  warnings.warn(


Best score: 0.690
Best parameters set:
        cvec__max_df: 0.9
        cvec__max_features: None
        cvec__min_df: 0.01
        cvec__ngram_range: (1, 2)
        rf__max_depth: 20
        rf__n_estimators: 200
Wall time: 1h 27min 36s
```

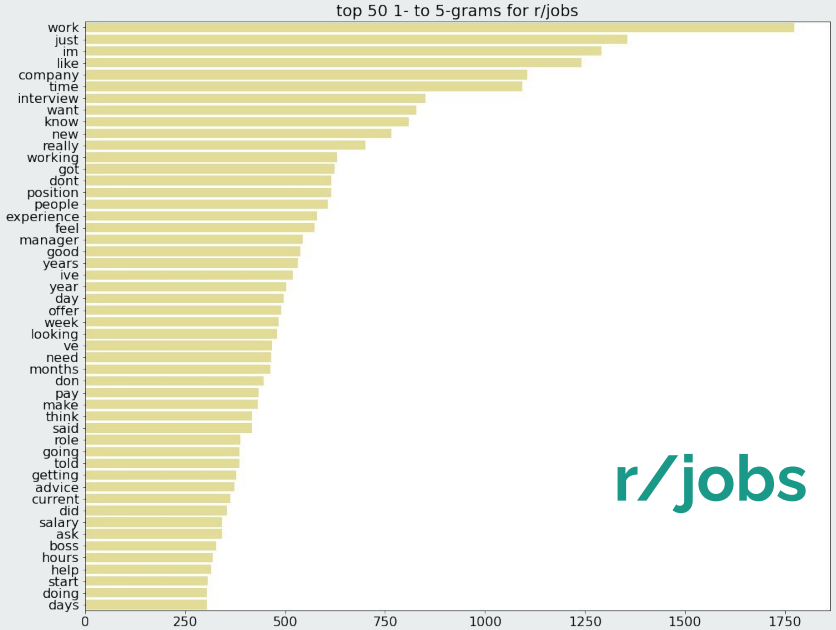
ONE AND A HALF HOURS!!!
有没有开玩笑！！！

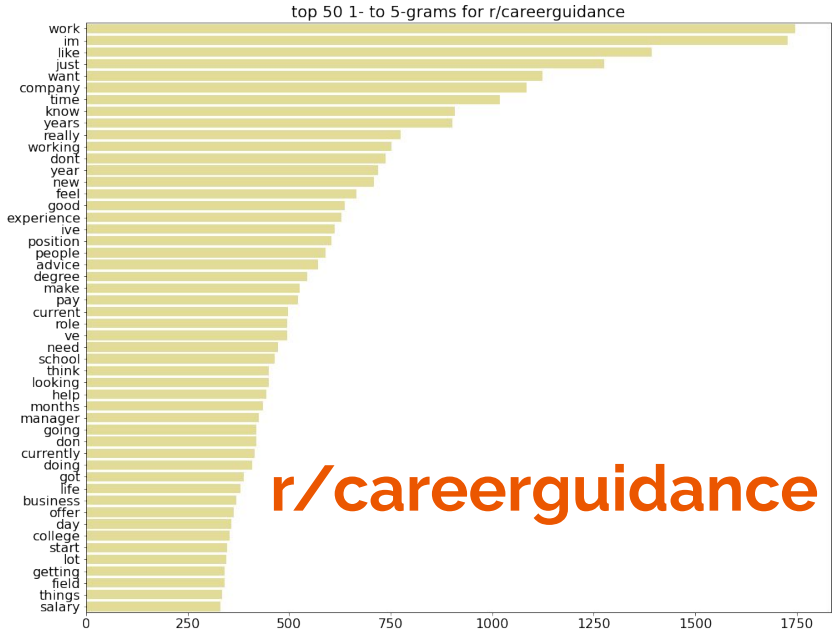
3. Iterative modelling, hyperparameter tuning and production

combined_text → *[cvec, tvec]* → *[nb, rf]* → **Choose Model**

Top 1- to 5-grams

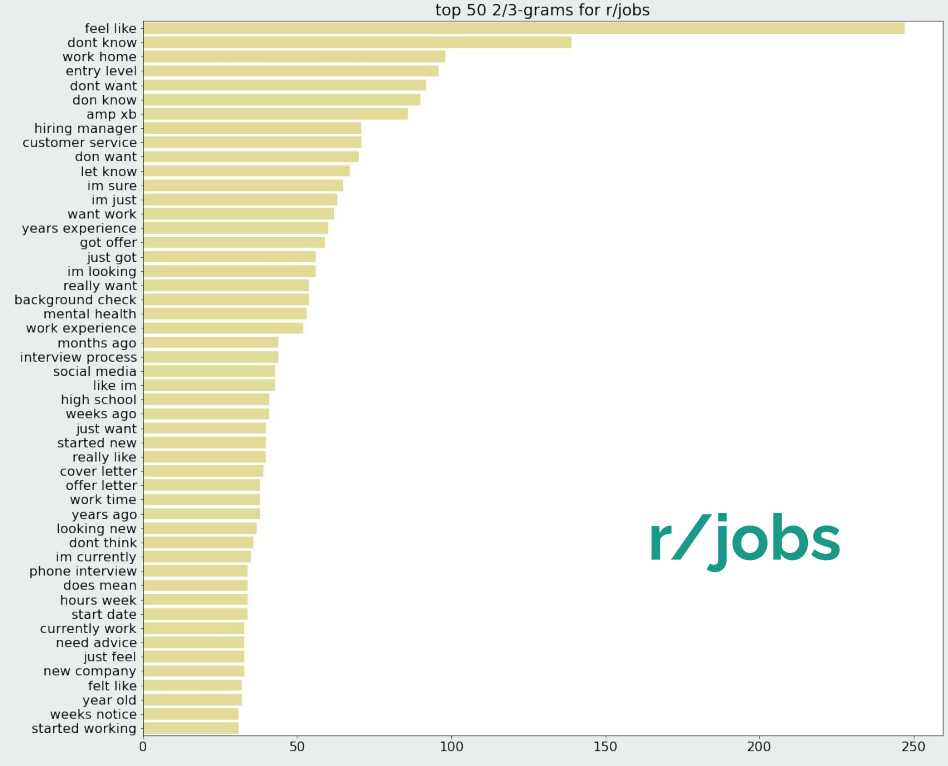
- Unigrams dominate
- Similar top words with different ranking / frequency

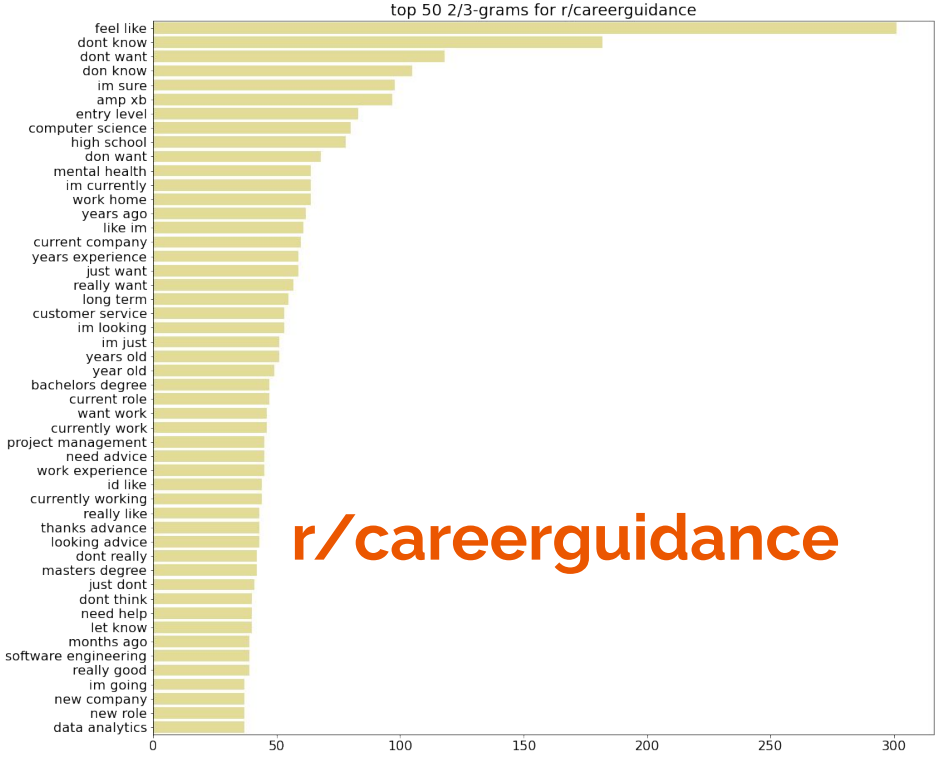




Top 2/3-grams

- No trigrams made the top 50 list
- Some similar top words with different ranking / frequency





After some tuning...

Model Evaluation

Metrics

No.	Predictor	Model	Train accuracy	Validation accuracy	Train F1 Score	Validation F1 Score	Wall Time
1	combined_text	DummyClassifier (Most frequent)	0.515	0.515	n.a.	n.a.	<1s
2	combined_text	RandomForest (cvec)	0.862	0.667	0.842	0.629	20min 46s
3	combined_text	RandomForest (tvec)	0.853	0.666	0.828	0.617	20min 43s
4	combined_text	MultinomialNB (cvec)	0.705	0.684	0.691	0.670	2min 32s
5	combined_text	**MultinomialNB (tvec)**	0.729	0.689	0.714	**0.674**	**2min 29s**

Misclassification

Analysing misclassified text indicates innate difficulty in categorization

(could a human do better?)

Multinomial Naive Bayes
(with tf-idf vectorizer)

	text	predicted	true
735	out of curiosity, what types of can i apply for if i have a bachelors in film and a bachelors in law if i combined these two degrees when looking for a whether it be in the business field or legal field, what could i possible search for? i have a bachelors degree in film and i have another bachelors degree in law i am currently looking for a and i got curious: if i were to look for a where i combined my search with both my film and law degree what type of could i expect to find i am very interested in the business or legal field i really dont mind but im just very curious since im trying to think out of the box here	0	1
3676	how to convince recruiter to transfer me to nyc? i have a super day coming up with a large company so obviously this question is just proactive in case i do get the offer lol when applying for the i had to rank the out of cities i preferred, and nyc was the only one i wanted i ended up ranking nyc and columbus i just got the email for the super day interview and it says it is for the columbus location i was wondering if this is something i should speak to the recruiter about before the interview or hopefully once i actually receive the offer? i don't want to bring it up too soon and look unenthusiastic or something also what do you even think the odds are of being moved to nyc as opposed to columbus? thanks all in advance	1	0
725	for uiux designers this is for the future, but id like to see what kind of companies id like to be at:	0	1
2625	new requires travel - concerns? hola, so i just accepted a new it position right before the new year with a increase in pay i spoke with the hiring manager, and they mentioned about a - travel rate, which seemed fine for me this morning i got connected with someone that i am going to be working with and they were able to give me more insight into the travel so basically, they said that travel varies sometimes it can be multiple times a month, once a month, or even every other month our work days are only monday - thursday, so we would fly out monday and fly back thursday he also said that if not always, you will be home for the weekends friday - sunday this is my first that has a travel aspect, and i have a baby boy coming in may i talked with the wife and she seems to be ok with this, and happy for me making great moves in my if we werent having a baby right now, i would have no reservations about this, but since we are, i just dont know how i feel about it those that do travel, how do you see it affecting your at home life, if at all? have you noticed that a little time away from each other has helped your relationship we both work from home right now and we tend to get a little agitated easier since we never get time apart	1	0
741	anyone get nervous when starting a new ? anxiety through the roof	0	1

0: r/careerguidance; 1: r/jobs

Feature Rankings

Top 10 predictive words under this model

Multinomial Naive Bayes
(with tf-idf vectorizer)

	feature names	feature_log_prob_0	feature_log_prob_1	log_prob_diff
609	path	-6.077097	-7.743109	1.666012
102	choose	-6.431733	-7.675785	1.244052
282	finance	-6.394366	-7.539469	1.145103
224	email	-7.120348	-5.997100	1.123248
738	science	-5.986337	-7.096724	1.110387
31	application	-7.310175	-6.208768	1.101407
128	computer science	-6.565098	-7.620112	1.055013
337	guidance	-6.490284	-7.543298	1.053014
521	masters	-6.179064	-7.231109	1.052045
700	references	-7.716201	-6.688933	1.027269

Next steps to take model further

- Preprocessing: iteratively lemmatization and augment stop words

- Review misclassification iteratively and add relevant steps in preprocessing to improve model scores

- Try other models and vectorizers

Thanks for listening!