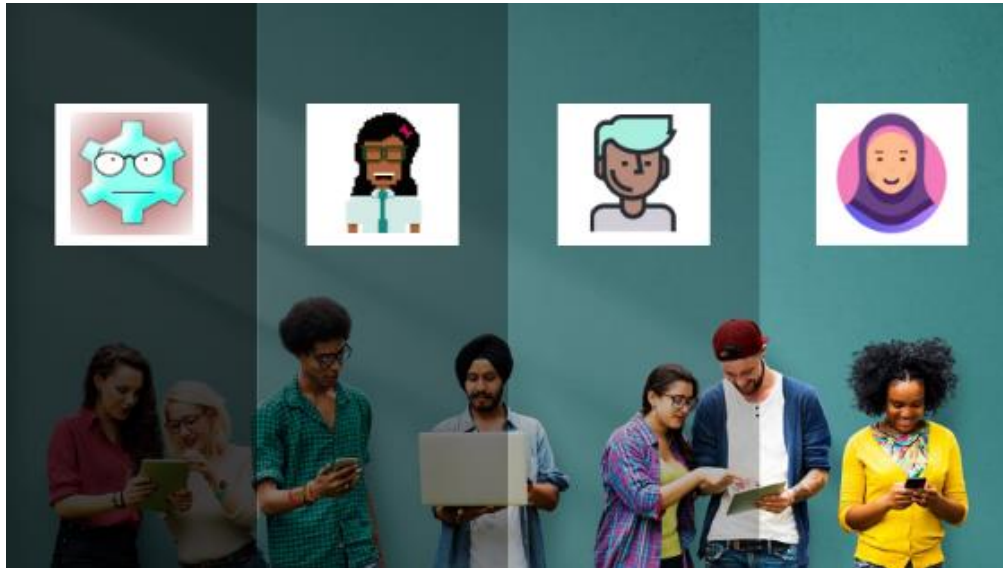# CAREER VILLAGE RECOMMENDATION SYSTEMS

By (Dinesh Reddy, Sashank Mandiga and Kavya Natta)

Class: Big Data Experience (CIS8395)

Professor: VijayKumar Gundapodi

pinnacle

Table of Contents

pinnacle

Introduction

Did you ever have questions about your future?

Did you ever thought about what career path to choose?

Ever had questions about how a career path look like?

These important open-ended questions bring us to this company CareerVillage.org a nonprofit organization that has developed a website to help answer questions of students about their college and careers by working professionals. To get most responses to a question CareerVillage.org needs to be able to send the right questions to the right volunteers. Notifications sent to volunteers have the greatest impact on how many questions are answered. In this project, we would like to develop a Recommendation System to recommend relevant questions to the professionals.

"Today, maybe more than ever, our youth - and in particular at-risk youth - need exposure to and guidance from leaders of industry, entrepreneurs, and professionals succeeding in the real world. CareerVillage provides that and more. After seeing it in action, it could be a game-changer"-

Robert Piercey - Regional Director, NFTE

The website that gets your questions about college and careers answered online by actual working professionals.

With a mission to democratize access to career information and advice for underrepresented youth, CareerVillage crowdsources the answers to every question from every student about every career. Together, they're building a massive open-access reference source that every online learner can access anytime, anywhere. CareerVillage has over 5,000,000 learners and over 80,000 volunteers

Milestones of CareerVillage.org

8K	5M	190	98%
Student-Defined Career Topics	Online learners served	Countries served	Answered rate

pinnacle

Problem Background

It is quite difficult to prepare for a career. Curiosity becomes knowledge, worry becomes drive, and paralysis becomes action when instant on-demand answers are available. Most young people believe they are on their own in their job search; with CareerVillage.org, they no longer have to be. Unfortunately, far too many people do not have access to career guidance. Young individuals in low-income communities, young people of color attempting to break into the financial services industry, young women attempting to break into STEM, young Asian Americans attempting to break into show business, and many other underrepresented and under-served demographics confront significant obstacles.

There are currently 5.5 million underserved students in USA. Now you might be interested in to know which groups are counted towards underserved Populations. According to Law Insider.com *"Underserved Populations means populations who face barriers to accessing and using victim services, and includes populations underserved because of geographic location or religion, underserved racial and ethnic populations, populations underserved because of special needs (such as language barriers, disabilities, alienage status, or age), and any other population determined to be underserved by the Attorney General or the Secretary of Health and Human."*

Have you felt that the definition of underserved populations is very complicated to understand? May be that's where the whole complexity starts. Underserved communities face many systemic and institutional barriers few of which include limited accessibility to basic needs like healthcare and education institutions. The 5.5mn underserved student has a very limited access to career advice either in the form of student counsellors or in any form of organized way.

The word "student counsellors" bring me to the next major contributor to the problem background which is Student Counsellor Ratio. The student-to counselor ratio of United States map clearly shows the poor state of the number of student counselors in USA. There are only 2 states which meet the Recommended ratio of 250. The

MAP A
Student-to-Counselor Ratios
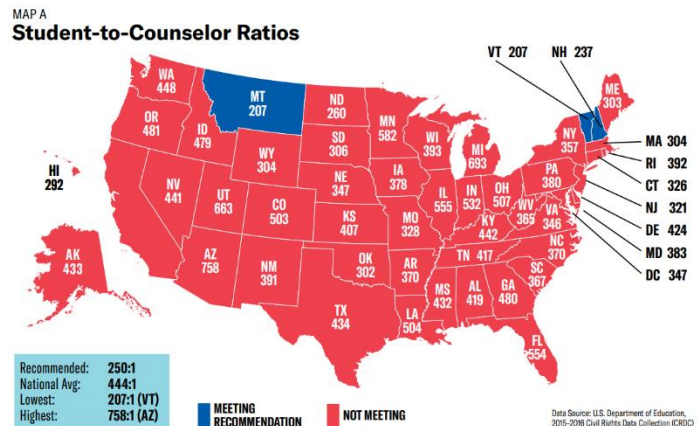
national average is 444:1 which is almost double the recommended ratio. This shows that there is a huge gap in the current system which needs to be addressed and a lot of students out there are looking for help with the guidance about the career opportunities.

CareerVillage.org does a perfect job addressing the both the problems helping students to connect with right professionals to get the right career advice. One of the technical challenges faced by the Careervillage.org team is to match the professionals to the questions they would be motivated to answer. If the question that matches with the professional's interest is recommended to the right professional, then the chances of getting an answer to that question highly increases. For example, if we consider the real-time example taken from the CareerVillage website we can clearly see that there are few questions with 0 answers. There are some basic questions like "What is it being a pediatric nurse?" not being answered. And this question is posed

almost 5 weeks ago with 75 views and still not being answered. There is a high possibility that the

pinnacle

right professional is not aware of this question. Our objective is to decrease the scenarios like this by developing a Recommender Engine.

Data Sources:

The data has been taken from Kaggle website consisting of 15 files with a size of 440MB in '.csv' format. The data consists of various information related to users in our case students and professionals who volunteered and has a verified tag. We collected five years of data that has 30,972 students and 25000 professionals from various industries such as Engineering, medical, Business Analyst, teaching etc., There are around 16,270 unique tags and more than one million students had subscribed on updates like daily updates, weekly digest etc.,

Description about data files:

- answers.csv: Answers are what this is all about! Answers get posted in response to questions. Answers can only be posted by users who are registered as Professionals. However, if someone has changed their registration type after joining, they may show up as the author of an Answer even if they are no longer a Professional.
- comments.csv: Comments can be made on Answers or Questions. We refer to whichever the comment is posted to as the "parent" of that comment. Comments can be posted by any type of user.
- emails.csv: Each email corresponds to one specific email to one specific recipient. The frequency_level refers to the type of email template which includes immediate emails sent right after a question is asked, daily digests, and weekly digests.
- group_memberships.csv: Any type of user can join any group. There are only a handful of groups so far.
- groups.csv: Each group has a "type". For privacy reasons, group names are encrypted.
- matches.csv: Each row tells you which questions were included in emails. If an email contains only one question, that email's ID will show up here only once. If an email contains 10 questions, that email's ID would show up here 10 times.
- professionals.csv: "Professionals" are volunteers here, They're the grown-ups who volunteer their time to answer questions on the site.

pinnacle

- questions.csv: Questions get posted by students. Sometimes they're very advanced. Sometimes they're just getting started. It's all fair game, as long as it's relevant to the student's future professional success.
- school_memberships.csv: Just like group_memberships, but for schools instead.
- students.csv: Students are the most important people on CareerVillage.org. They tend to range in age from about 14 to 24. They're all over the world, and they're the reason we exist!
- tag_questions.csv: Every question can be hashtagged. We track the hashtag-to-question pairings and put them into this file.
- tag_users.csv: Users of any type can follow a hashtag. This shows you which hashtags each user follows.
- tags.csv: Each tag gets a name.
- question_scores.csv: "Hearts" scores for each question.
- answer_scores.csv: "Hearts" scores for each answer.

Dataset Overview:

File Name	Columns	File Size	No. of Rows
answer_scores	id, score	1.7 MB	51,138
answers	answers_id, answers_author_id, answers_question_id, answers_date_added, answers_body	51.2 MB	51,137
comments	comments_id, comments_author_id, comments_parent_content_id, comments_date_added, comments_body	4.3 MB	14,966
emails	emails_id, emails_recipient_id, emails_date_sent, emails_frequency_level	172.2 MB	1,048,576
group_memberships	group_memberships_group_id, group_memberships_user_id	67 KB	1,038
groups	groups_id, groups_group_type	3 KB	49
matches	matches_email_id, matches_question_id	172 MB	1,048,575
professionals	professionals_id, professionals_location, professionals_industry, professionals_headline,	3.8 MB	28,152

pinnacle

	professionals_date_joined		
question_scores	Id, score	819 KB	23,928
questions	questions_id, questions_author_id, questions_date_added, questions_title, questions_body	9.2 MB	23,931
school_memberships	school_memberships_school_id, school_memberships_user_id	220 KB	5,638
students	students_id,students_location, students_date_joined	2.5 MB	30,971
tag_questions	tag_questions_tag_id, tag_questions_question_id	2.8 MB	76,553
tag_users	tag_users_tag_id, tag_users_user_id	5.2 MB	136,663
tags	tags_tag_id, tags_tag_name	318 KB	16,269

Data Architecture Diagram:
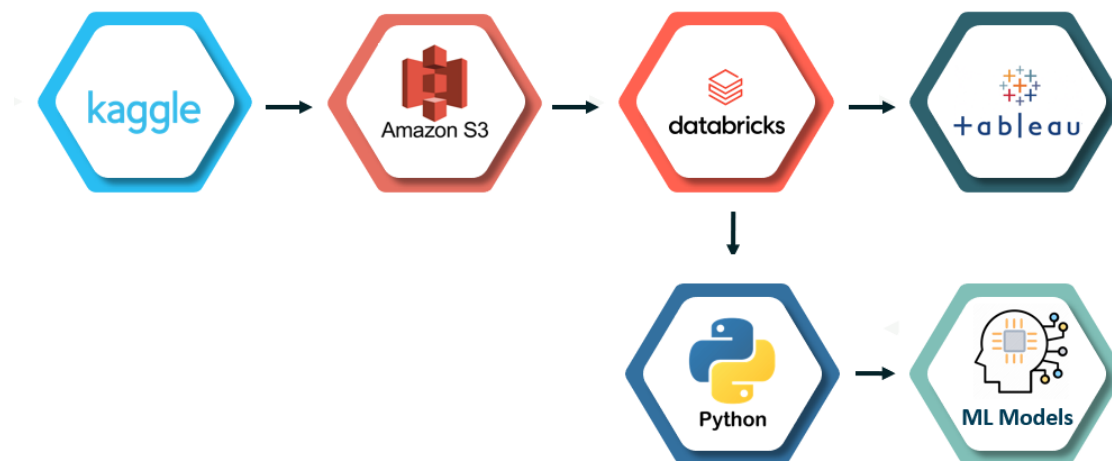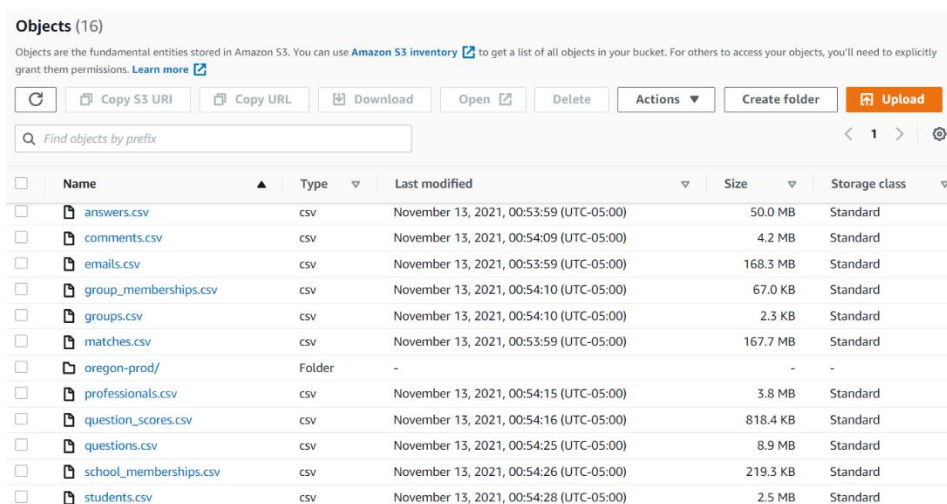
After we collect all the data files from Kaggle, we are storing the data consisting of 440MB in Amazon S3 (simple storage services) and performing ETL operations in Databricks. The programming language we will be using is python. Once the data is cleaning and transformed, Insights will be shared using Tableau and we will perform Machine learning algorithms and Recommendation Engines to analyze the patterns and recommendations to the students.

Why Amazon S3?

Amazon S3 is an object storage service with industry-leading scalability, data availability, security, and performance. Amazon S3 allows customers of all sizes and sectors to store and safeguard any amount of data for a variety of use cases, including data lakes, websites, mobile applications, backup and restore, archive, business applications, IoT devices, and big data analytics. We chose this because AWS provides 5GB of S3 standard storage for 12 months with the AWS free tier.

Importance of HDFS and Databricks:

HDFS stands for Hadoop Distributed File System. The function of HDFS is to operate as a distributed file system designed to run on commodity hardware. HDFS is fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets and enables streaming access to file system data in Apache Hadoop. HDFS is used to replace costly storage solutions by allowing users to store data in commodity hardware vs proprietary hardware/software solutions. Since Databricks manages Spark clusters, it requires an underlying Hadoop Distributed File System (HDFS). This is exactly what DBFS (Databricks file system) is in Databricks. Basically, HDFS is the low cost, fault-tolerant, distributed file system that makes the entire Hadoop ecosystem work. We will be Mounting the data from S3 bucket, and the data will be stored in DBFS.

ETL Process:

Full form of ETL is Extract, Transform and Load. ETL is a process that extracts the data from different source systems, then transforms the data (like applying Data cleaning, calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. Databricks, is a fully managed service which provides powerful ETL, analytics, and machine learning capabilities. A properly designed ETL system extracts data from the source systems, enforces data quality and consistency standards, conforms data so that separate sources can be used together, and finally delivers data in a presentation-ready format so that application developers can build applications and end users can make decisions. ETL systems commonly integrate data from multiple applications (systems), typically developed and supported by different vendors or hosted on separate computer hardware. The separate systems containing the original data are frequently managed and operated by different employees. We will be extracting the raw data from S3, perform data cleaning and transformations in Databricks and load.

Data Cleaning:

The practice of correcting or deleting incorrect, corrupted, improperly formatted, duplicate, or incomplete data from a dataset is known as data cleaning. Data cleaning will vary from dataset to dataset. In our dataset, there are numerous ways for data to be duplicated or mislabeled when merging multiple data sources. As a first option, we will drop observations that have missing values, but doing this might drop or lose important information. In that case, we will fill missing values with mean, median or mode based on other observations.
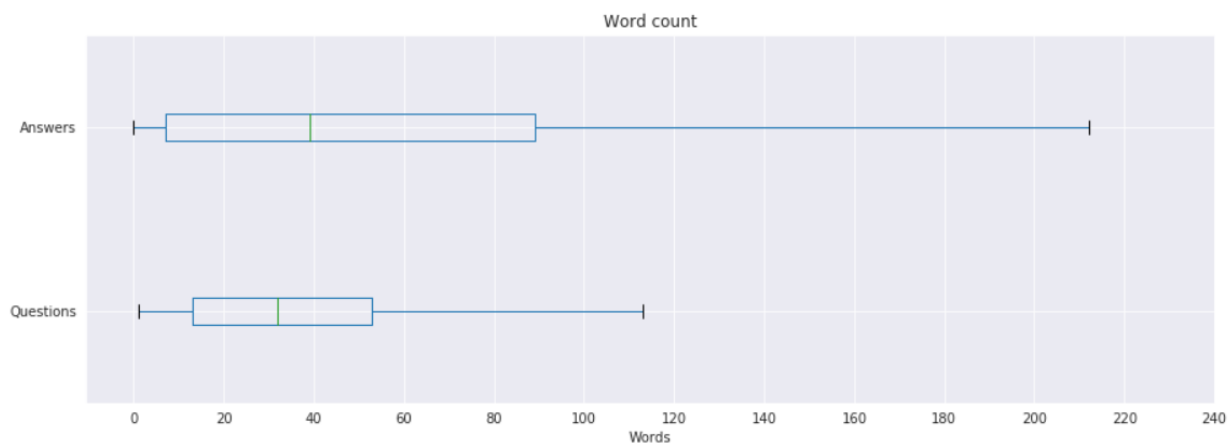
After performing Data validation on questions and answers body columns, the average word count for each question asked by the student ranges between 16-56 with a minimum of 0 till a maximum of 115 words. Whereas the average word count for each answer by a professional is in between the range of 5-90 with a minimum of 0 till a maximum of 215 words.

Data Joins:

Data merging is an important step in our solution. We have professionals, students, q&a and tags data as separate datasets. All tags (q & a) are stored in a separate dataset. The plan is to merge those tags with questions and answers datasets. Following this we will remove the outliers to improve the performance of the data. Following this, we will merge answers with questions since one question can relate to multiple answers.
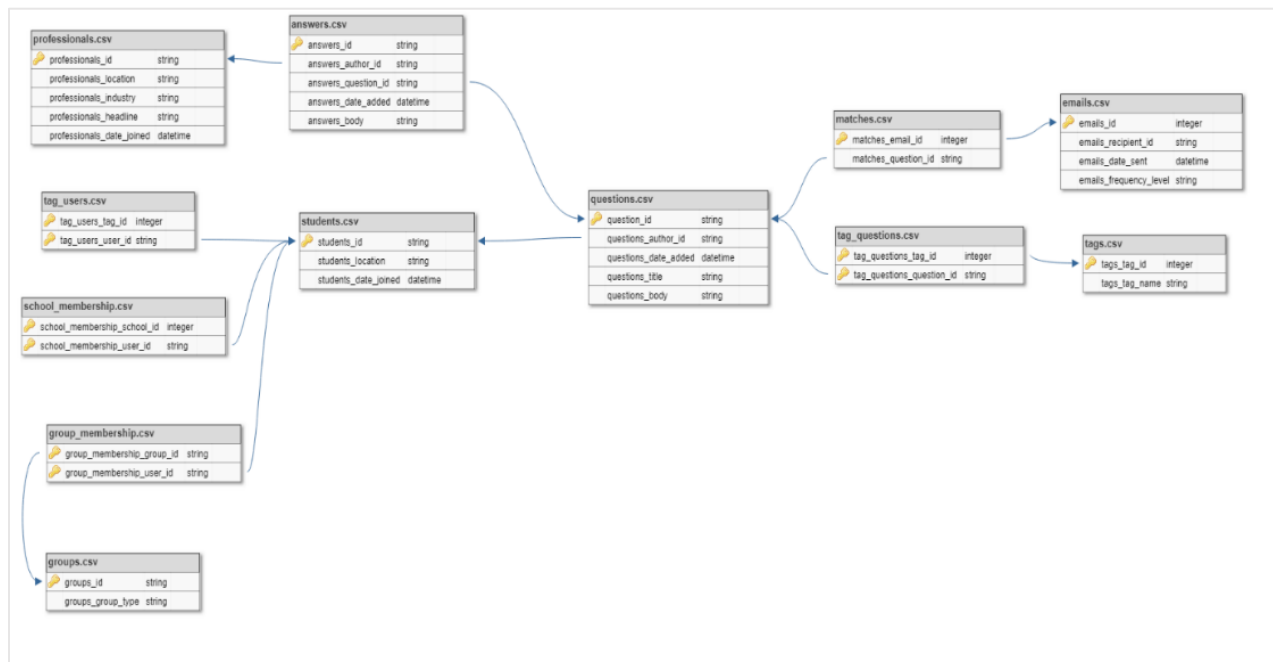
To simplify the complexity of our data, we created an E-R (Entity-Relationship) model to find a connection between all the 15 data files and merge them to a single data frame.

Visualizations and Insights using Tableau:

Tableau Software is an American interactive data visualization software company focused on business intelligence. Tableau can help anyone see and understand their data. After the Data is

cleaned and transformed, we use Tableau to provide insights on our data with interactive visualizations.
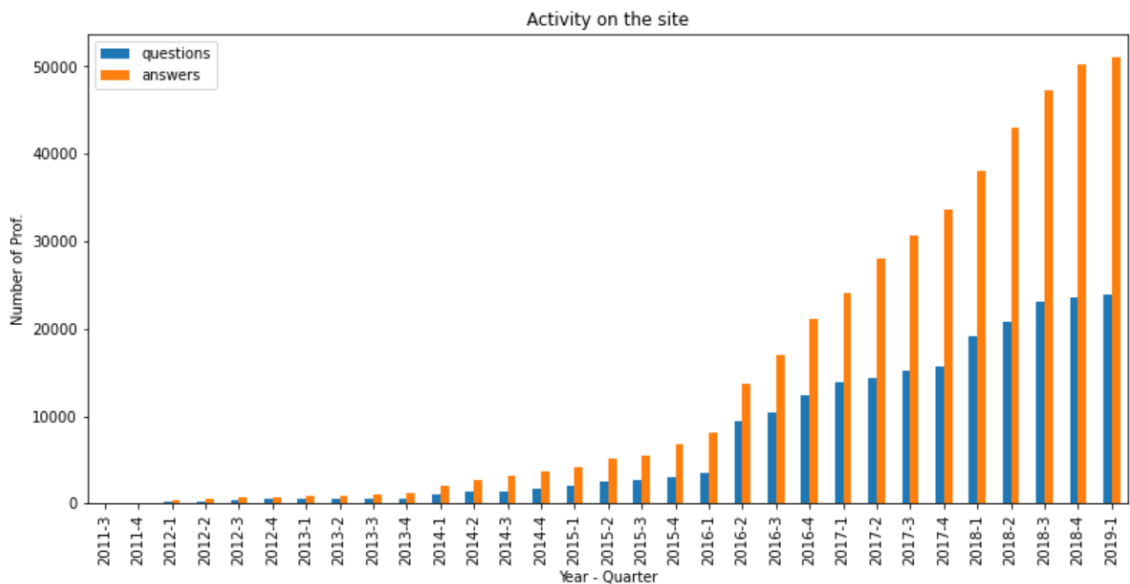
Fig. Number of questions professionals answered per year and quarter

From the above figure, the platform was able to stick to the policy of answering all questions during the years 2011-2015. Year 2018 has seen the lowest response rate of 91%; 712 questions went unanswered.
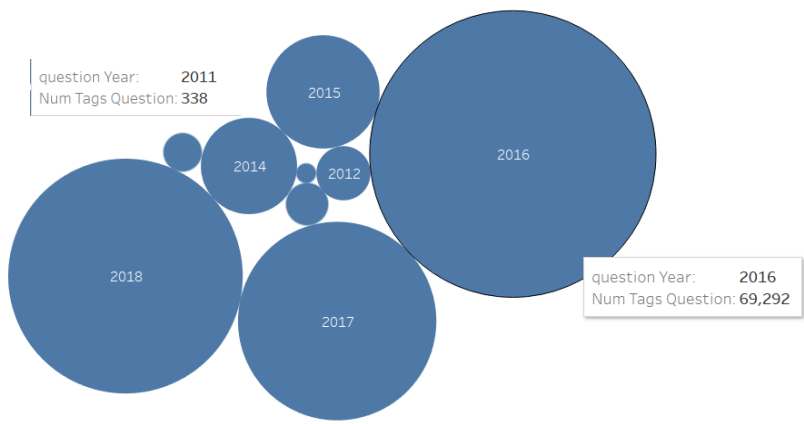
Fig. Number of questions Tags per year

pinnacle

Above visualization is the number of question tags posted per year. As the bubble size varies with number of tags, 2016 has recorded highest tags which is 69,292 and the least tags recorded is 2011 with 338 tags assuming that is the initial years creating the awareness to students.
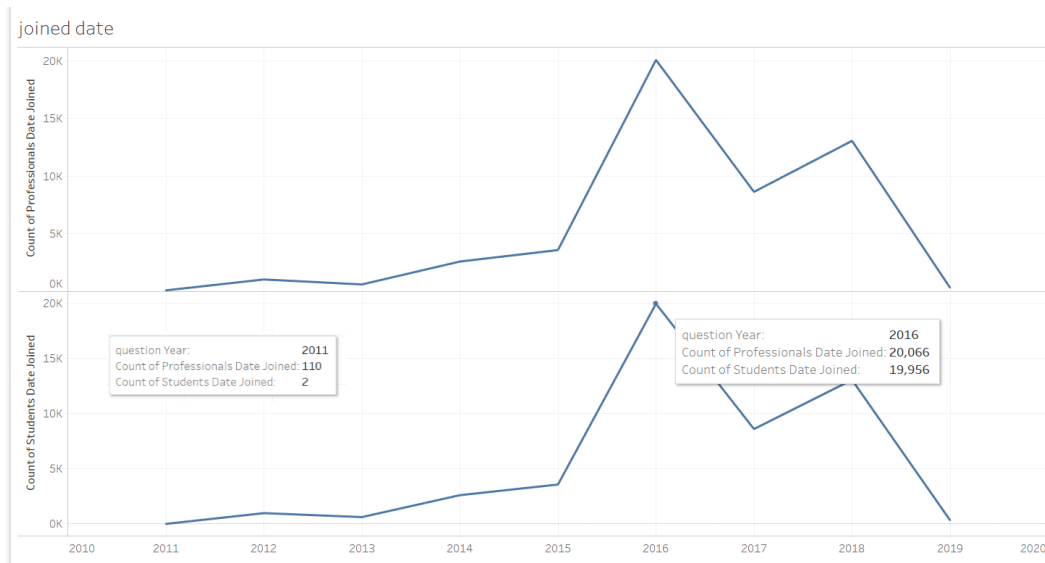
Fig. Professionals and students joined year

Above figure interprets the number of students and professionals joined per year. The highest is recorded in the year 2016 with total students to 19,956 and professionals to 20,056. The least number of students were 2 and professionals were 110 in the year 2011.
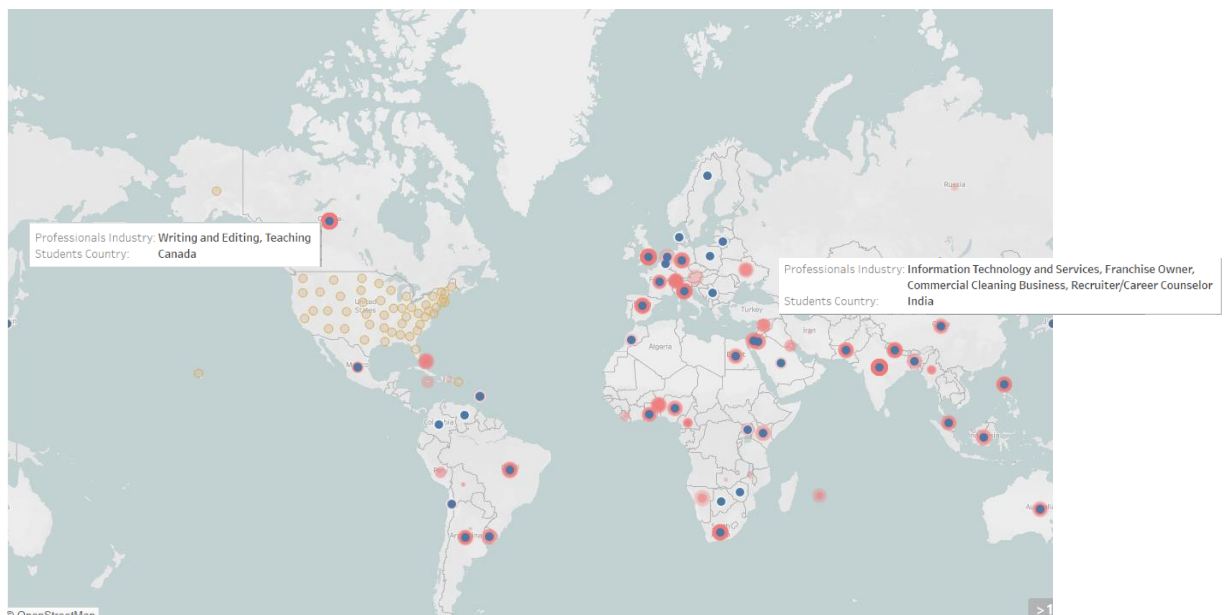
Fig. Geographical Location using Tableau

Above figure interprets the geographical map plotting the students and professionals' location along with the professional industry around the world. Glimpse of it are the countries India and Canada where the professional industry from India is vast with "information technology and services, Franchise owner, Commercial Cleaning business, Recruiter/career counselor" and in Canada, the professionals are from Writing and editing, Teaching background.

Recommendation Engines:

Whether consumers are shopping online, streaming a movie or song, or reading news stories, recommendation systems recommend products, content, or services they would enjoy. Recommender systems analyze relevant items based on existing context, such as user information, time, and location. It is thus possible to successfully recommend items from the tails of the popularity distribution. The majority of today's E-Commerce firms, such as eBay, Amazon, Alibaba, and others, make use of their own recommendation algorithms to better offer customers with products they are likely to like. Recommendation engines are essentially data filtering technologies that use algorithms and data to propose the most relevant items to a certain user.

What are the different types of recommendations?

There are basically three important types of recommendation engines:

- Collaborative filtering
- Content-Based Filtering
- Hybrid Recommendation Systems

Collaborative Filtering

This filtering strategy is typically focused on gathering and analyzing data on a user's behaviors, activities, or preferences, and then predicting what they will like based on their resemblance to other users. The collaborative filtering strategy has the advantage of not relying on machine analyzable content, which allows it to accurately recommend complicated objects like movies without requiring a "knowledge" of the item. Collaborative filtering is founded on the idea that people who have agreed in the past will agree again in the future, and that they will enjoy comparable products. The key trick is to find the similarity between users. This similar set of users is called neighborhood.

If a person A enjoys movies 1, 2, and 3 and a person B enjoys 2,3,4, they have comparable interests, and A should enjoy movie 4 and B should enjoy movie 1.

COLLABORATIVE FILTERING

Watched by both users

Similar users

Watched by her, Recommended to him!
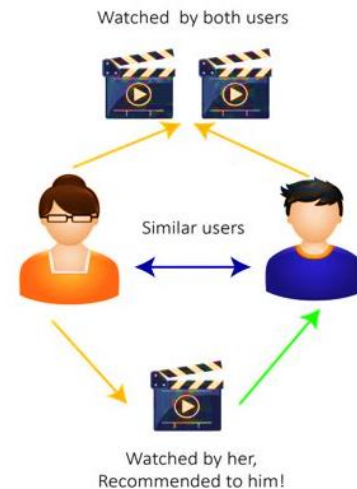
Content-Based Filtering

The main idea behind content-based filtering is to recommend movies to a user similar to his/her previous movies rated highly by that user. For example, recommending movies with same actors, director, genre; articles with similar content; recommend people with many friends. Recommendations by content-based filtering are generated by matching keywords and attributes supplied to database objects (such as items in an online marketplace) to a user profile. Purchases, ratings (likes and dislikes), downloads, products searched for on a website and/or added to a cart, and clicks on product links are all used to form the user profile.

Take, for example, Taylor Swift's "The Last Time," Shakira's "Can't Remember to Forget You," and Beyoncé's "Me, Myself, and I."

CONTENT - BASED FILTERING

Watched by User

Similar Movies

Recommended to user

pinnacle
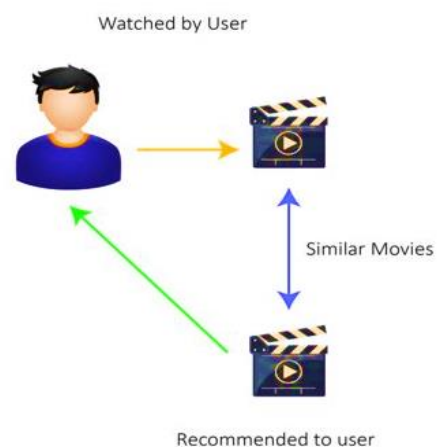
You could appreciate female pop musicians and breakup songs, according to a recommender system. More breakup songs by these and other female pop musicians, such as Miley Cyrus' "Slide Away," are likely to be recommended.

Hybrid Recommendation Systems

Hybrid recommender system is a special type of recommender system that combines both content and collaborative filtering method which could be more effective in some cases. Making content-based and collaborative-based predictions independently and then merging them is one way to apply hybrid techniques. Furthermore, by incorporating content-based capabilities into a collaborative-based method and vice versa; or by combining the two techniques into a single model.
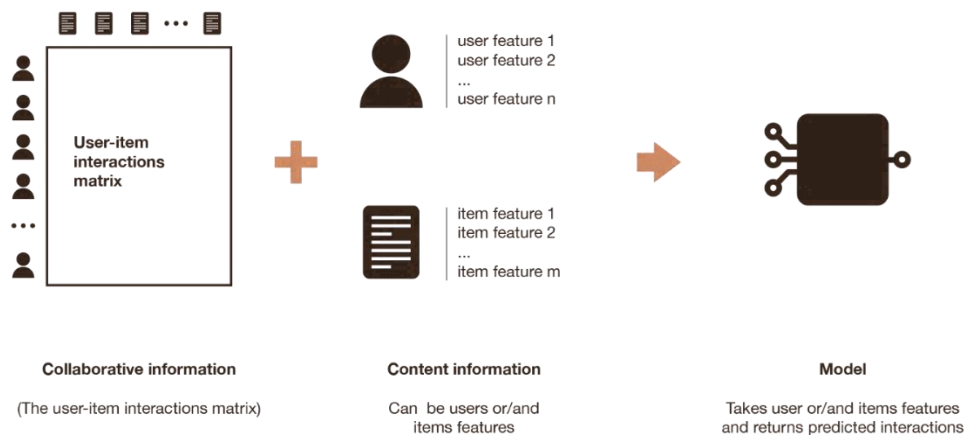
Several empirical studies compared the effectiveness of hybrid methods with pure collaborative and content-based methods, revealing that hybrid methods can deliver better recommendations than pure approaches. Some of the most prevalent recommender system issues, such as cold start and sparsity, can be solved using these strategies.

Reasons for choosing Hybrid Recommendation System

Career Village dataset is all about professionals, professional answered questions, students posted questions. The dataset doesn't involve any data related to any ratings or equivalent matrix that helps determine how professionals like a question. So, we will use implicit feedback data such as tags and location for this dataset.

pinnacle

It is not advisable to use collaborative filtering alone because of the cold start problem as majority of the professionals are either new or haven't answered any questions yet. We cannot use content-based filtering alone due to diversity problem. It will provide only safe recommendations and creating a user-profile matrix for new professionals will be difficult.

The above problems can be resolved if we use hybrid model as it will combine both *content-based and collaborating filtering* to make a robust recommendation engine.

Recommendations for Professionals

Why LightFM?

A hybrid recommender is a type of recommender that makes recommendations using both collaborative and content-based filtering. LightFM is a *hybrid matrix factorisation* model in which people and objects are represented as linear combinations of the latent factors of their content attributes. Users and things are represented as latent vectors in LightFM, just as they are in a collaborative filtering paradigm. The LightFM model learns user and item embeddings in a way that encodes user preferences over things. These representations are multiplied together to provide scores for each item for a certain user; items with high scores are more likely to be appealing to the user. An embedding is estimated for each feature, and these features are then aggregated to provide user and item representations.

- The sum of the features' latent vectors gives the latent representation of user u:

$$q_u = \sum\nolimits^{j \in f_u}.$$

- The same for the following items:
- The dot product of user and item representations, adjusted for user and item feature biases,

$$\hat{r_{ui}} = f(q_u \cdot p_i + b_u + b_i)$$

yields the model's prediction for user u and item I

LightFM outperforms pure content-based models in both *cold-start* and low-density circumstances, and significantly outperforms them when collaborative information is provided in the training set or user features are included in the model. Our CareerVillage recommendation system has many fresh questions and students, creating an ideal scenario for the cold start problem.

pinnacle

Benefits of LightFM

- The most important benefit is that our data is implicit feedback, so the LightFM model we used implements **WARP** (Weighted Approximate-Rank Pairwise) loss for implicit feedback learning-to-rank.

- Pick a negative item at random from all the available items for a given (user, positive item pair). Calculate forecasts for both items; if the negative item's prediction is greater than the positive item's plus a margin, apply a gradient update to raise the positive item and lower the negative item. If there isn't a rank violation, keep sampling negative items until one is discovered.

- If you find a violating negative example on the first try, perform a large gradient update: this suggests that, given the current state of the model, a huge number of negative things are rated higher than positive ones, and the model must be updated by a substantial amount. If finding a violating example requires a lot of sampling, do a minor update: the model is probably close to the optimum and should be updated at a low pace.

- Many developers have already put it to the test.

- Many well-known brands have already used it in their products.

- It has a very developer-friendly API.

- Provide assessment matrices for assessing the model's performance.

- LightFM handles cold start problems exceptionally well. LightFM makes recommendations based on item and user characteristics. When no connection is found for new users, LightFM will revert to content or collaborative methods based on the data available.

- Rankings of questions: The goal of emailing queries to professionals is for students to receive a timely response. We must ensure that underserved questions receive a higher ranking.

- We built a model to handle this $p_i = \sum^{j \in f_i}$ problem by weighting the model so that it gives less weight to the questions with higher value.

- Professionals and questions can have several tags, and most of the time the tags are identical save for a few letters. LightFM produces users and products with embedded characteristics that recognize tag similarity.

- Our model can detect that student frequently added tags to questions that differ slightly from those used by professionals.

- Our model is based on the LightFM library and employs WARP loss, so it is fast.

Kernel Overview:
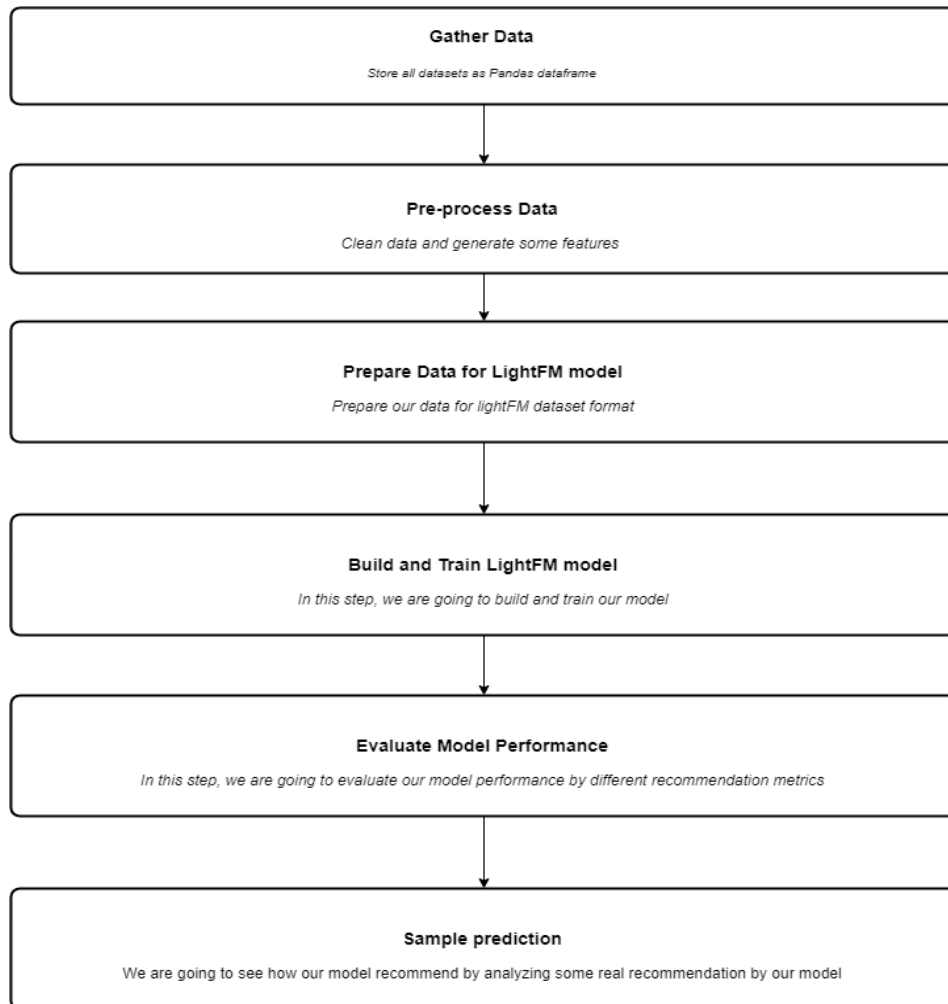
Recommendations for Students

Using NLP:

- Stopwords Removal
- Punctuation Removals
- TF -IDF Vectorizer

Stopwords Removal:

The most prevalent words in any natural language are stopwords. These stopwords may not contribute much value to the meaning of the document when evaluating text data and constructing NLP models.

The most prevalent terms in a book are "the," "is," "in," "for," "where," "when," "to," "at," and so on. When stopwords are removed from a dataset, the size of the dataset shrinks, as does the time it takes to train the model. Because there are fewer and only useful tokens left once stopwords are removed, performance may improve. As a result, it has the potential to improve classification accuracy. Stopwords are being removed from search engines like Google to allow for faster and more relevant data retrieval from databases.

Punctuation Removals:

Text preprocessing is a technique for cleaning text data and preparing it for use in a model. Text data comprises noise in the form of emotions, punctuation, and text in a different case, among other things. The removal of punctuations is another text processing approach. There are a total of 32 primary punctuations that must be addressed. We may use a regular expression and the string module to replace any punctuation in text with an empty string. We utilized a sub-method with three primary parameters: the first is a pattern to search, the second is the by which we must replace, and the third is the string or text that must be changed. So, after passing all of the punctuation, we check to see if anyone is present, and if so, we replace it with an empty string.

TF-IDF vectorizer:

To convert the Document Term Matrix (DTM) into numerical arrays, use the Count vectorizer or the TF-IDF vectorizer.

Why Gensim?

The final tokenization method we will cover here is using the Gensim library. **It is an open-source library for unsupervised topic modeling and natural language processing** and is designed to automatically extract semantic topics from a given document. Gensim is quite strict with punctuation. It splits whenever a punctuation is encountered. In sentence splitting as well, Gensim tokenized the text on encountering "\n" while other libraries ignored it. Gensim is a useful library
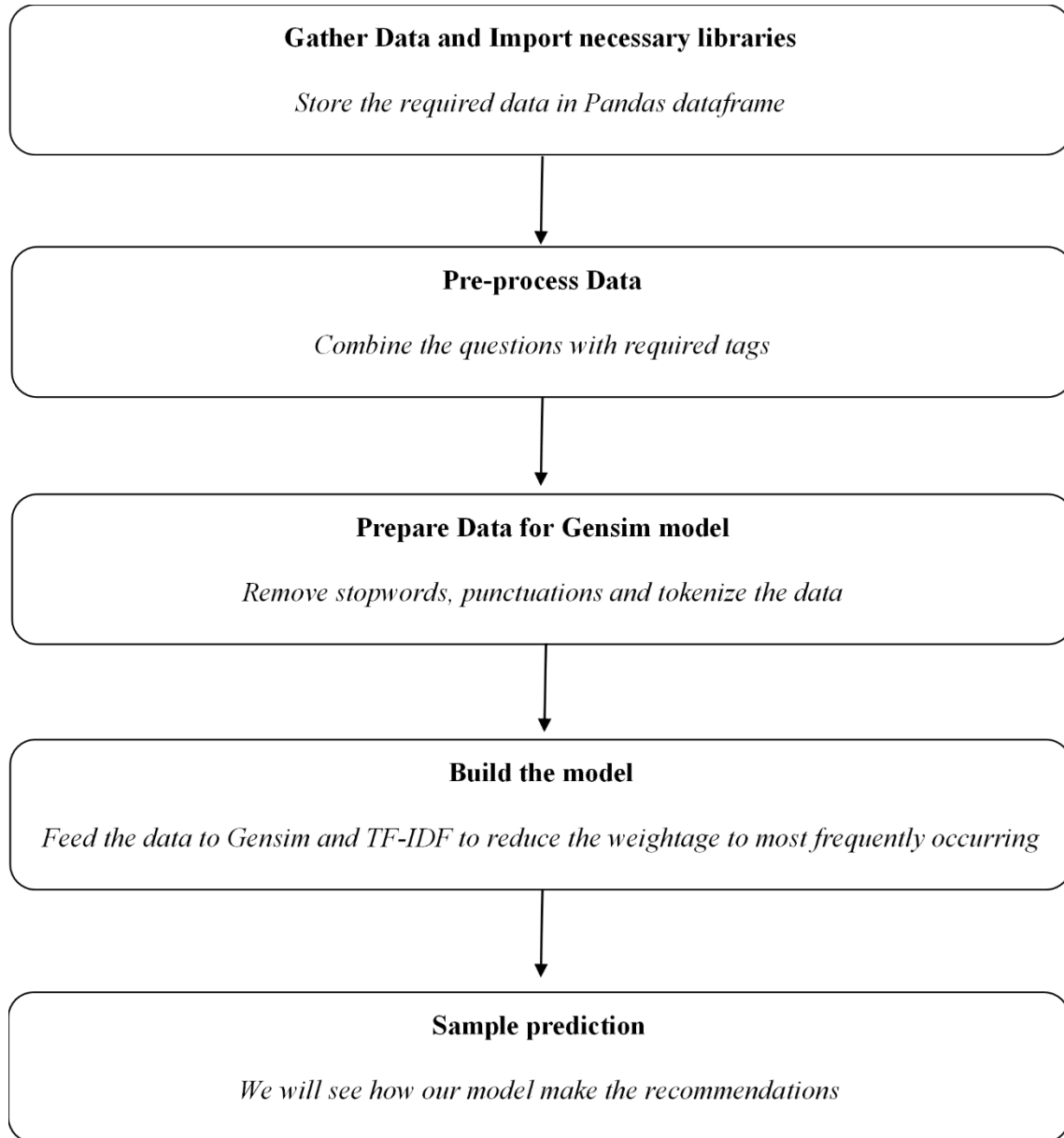
for performing NLP tasks. Gensim also has ways for removing stopwords during pre-processing. The remove stopwords method from the class gensim.parsing.preprocessing can be readily imported. While using Gensim for removing stopwords, we can directly use it on the raw text. There's no need to perform tokenization before removing stopwords. This can save us a lot of time. We don't need to generate the DTM matrix from scratch while using Gensim for Document Term Matrix (DTM). The DTM is generated via an internal process in the Gensim library. The sole stipulation for using the Gensim package is that the cleaned data be sent as tokenized words.

- The next step is to use the dictionary we created earlier to turn the corpus (list of documents) into a document-term Matrix.
- We implemented the LDA model by creating the object and passing the required arguments.
- Each of the unique words is given weights based on the topics. In other words, it implies which of the words dominate the topics. we find the topics from Corpus.
- We assigned these resultant topics to the documents.

Benefits of Gensim

- Gensim is an open-source vector space and topic modelling toolkit.
- Gensim used Numpy and SciPy.
- Designed to handle large text collections
- Generates similarity score and above a threshold value, we can identify similar questions in the dataset

Kernel Overview:

Gather Data and Import necessary libraries

Store the required data in Pandas dataframe

Pre-process Data

Combine the questions with required tags

Prepare Data for Gensim model

Remove stopwords, punctuations and tokenize the data

Build the model

Feed the data to Gensim and TF-IDF to reduce the weightage to most frequently occurring

Sample prediction

We will see how our model make the recommendations

Results and Conclusion

Remember this image (to the right)? we have used it in the start of the paper while defining the problem statement. There, we have defined that our goal is to increase the number of answers for the questions that are not answered, i.e., to promote the questions that are less popular.

Now let's take 2 examples (real profiles of Professionals) from the Career Village data and see how our LightFM Hybrid Recommendation System will recommend questions to the Professionals based on their profile, the tags they liked and the questions they have answered previously.

Professional Id (19897) – Professional with only 1 previously answered question

Professional Id (19897): Previous Answered Questions

	questions_title	question_features
22784	Do companies truly focus on your college major when applying for jobs?	(22784, [major])

	professionals_id_num	professionals_tag_name
19897	19897	illustration,graphic-design,adobe-creative-suite,comic-books

Professional Id (19897): Recommended Questions:

	questions_title	question_features
19407	How can you be a successful photographer? What...	(19407, [graphic-design, art, photography])
6058	How should you start in the Graphic Design ind...	(6058, [design, graphic-design, art])
2310	what is one of best things about being an anim...	(2310, [art, animation, design, artist])
13484	Would a Graphic Design degree be a feesible op...	(13484, [graphic-design, art])
1416	Is Competition In The Animation Field Low or H...	(1416, [animation, art])
17325	what are the required fields forgraphic design?	(17325, [graphic-design, art])
19275	how do i get my content (animations) out in th...	(19275, [animation, art])
19471	Graphic Design - job outlook for the next 10 y...	(19471, [graphic-design, art])

In the above image, we can see that the Professional with ID 19897, has answered one question previously which has one tag "Major". The tags that the Professional follow are illustration, graphic design, comic books. With the help of these two data points combined when feeded to the LightFM Model, the model will recommend questions with similar tags that are mentioned in the above image. We can see clearly that the recommended questions are very closely related to the tags the user follows. Since the Professional has previously answered only one question, the cold start could have been a problem if we have only gone with Collaborative filtering approach.

Professional Id (3) – Professional with 3 previously answered questions

Professional Id (3): Previous Answered Questions

	questions_title	question_features
11339	What are the different jobs a person can do in Forensic Science?	(11339, [justice, forensic, criminal, science])
14818	What does a typical work day for a forensic scientist look like?	(14818, [No Tag])
19077	Is most of your day spent working when being a detective?	(19077, [detective])

Professional Id (3): Recommended Questions:

	questions_title	question_features
2423	How long does it take to become a Detective?	(2423, [law, police, lawyer, law-enforcement, ...
17184	What types of Detectives are there?	(17184, [law, police, lawyer, law-enforcement,...
8863	What qualifications are needed to be promoted ...	(8863, [police, criminal-justice, law-enforcem...
9778	I want to be a police officer or a police disp...	(9778, [police-officer, law, police, law-enfor...
20003	how many criminal psychologist jobs are our th...	(20003, [criminal-justice, psychology, law-enf...
11514	What does an aspiring cop have to look forward...	(11514, [police, criminal-justice, law-enforce...
21030	How can I get into Law Enforcement whle in hig...	(21030, [criminal-justice, law-enforcement])
16214	Do you go to college, then B.L.E.T(Basic Law ...	(16214, [police, law-enforcement, law])

The Professional with ID 3 has answered 3 questions which has tags "justice", "forensic", "criminal", "detective". When this data is fed to the model, we can see the questions that are recommended to the Professional have similar tags and tags like law, criminal, detective which come under the same category. Since the model is recommending questions, tags that aligns with the professional's interests and also as we have weighted the questions (questions with less answers have more weights) there is a high chance that the number of unanswered questions decreases.

pinnacle

There is a high chance of possibility for different students to post same questions/ similar questions that are answered to the Career Village website. This in return increases the count of duplicate questions. This duplicity of questions can damage our model's accuracy as these duplicate questions can be recommended to Professional multiple times. So, inorder to avoid this we have added this feature of "Query Suggestions"

```
1  Query='Can I become data scientist without studying at university?#technology #data-science'
2  Query
```

```
1  q_sim=similar_qs[query_doc_tf_idf]
2  sim_threshold=0.10
3  qs_with_tags['Similarity']=q_sim
4  ques=qs_with_tags[qs_with_tags['Similarity']>=sim_threshold]
5  ques=ques.sort_values('Similarity',ascending=False)
6  ques.head()
```

Out[240]:

questions_author_id	questions_date_added	questions_title	questions_body	questions_id_num
d969b3d11f2b4317b8391bc51434954a	2014-03-24 03:23:55 UTC+0000	What do data scientists do?	What kind of companies do data scientists work...	9478
aef198c7c57d43fda9a6f64b5fad12f6	2018-03-31 03:59:09 UTC+0000	How will the field of education be affected by...	I am asking this because I am interested in be...	13432
a527b1588d724854a8dbafaf39421f13	2018-03-26 02:14:50 UTC+0000	What is the difference between data science an...	I've been looking into data science careers, a...	181
0dd8594cab784294b9b7069086df7cb3	2018-11-28 23:32:30 UTC+0000	what should I learn to be a data scientist?	I want to be a data scientist, what online co...	15175

These Query Suggestions matches the similarity of the current query of student with exisiting questions present in the database and shares the similar questions and their answers with the student as shown in the above image. For example the question " Can I become data scientist without studying at university" is similar to "What should I learn to be a data scientist "

pinnacle

Challenges

- Choosing which recommendation system to implement and understanding the concept
- Learning the implementation of AWS S3 storage, pulling the data into Databricks for processing
- Merging 15 separate csv files into a single data frame, includes extensive research regarding each data file and different columns associated with it

References:

- *https://canvas.uw.edu/courses/1176574/pages/reading-assignment-who-are-the-underserved*
- *https://salud-america.org/47-states-dont-meet-the-recommended-student-to-counselor-ratio/*
- *http://pzs.dstu.dp.ua/DataMining/recom/bibl/1aggarwal_c_c_recommender_systems_the_textbook.pdf*
- *https://docs.databricks.com/data/data-sources/aws/amazon-s3.html*
- *https://making.lyst.com/lightfm/docs/examples/dataset.html*
- *https://www.kaggle.com/c/data-science-for-good-careervillage*
- *https://www.tableau.com/*
- *https://medium.com/product-design-data/intro-to-recommendation-engines-how-they-work-practical-guide-part-i-fdd94078af53*
- *https://www.programmingcube.com/how-do-you-parse-text-in-python/*