The Problem:

Careervillage.org is a website where people with career related questions can go to get advice and answers to those questions from professionals in the field. The challenge that they’re trying to address here is to increase response rates from the volunteers. They are looking for a way to match questions to volunteers in the related profession and to try to drive questions to the volunteers most likely to answer that question. The prompt mentioned that select questions might be sent to professionals in an email form to solicit an answer from them in digest form on a periodic basis. This means that of the many questions asked, only a few most relevant questions should be sent to each professional. There are no doubt many reasons why we should want to send these people only a few of the most relevant questions, but CareerVillage does not elaborate and I will not speculate.

The Data:

Careervillage.org has provided several large files that provide the bulk of the data used in this project. Included in the data are questions that were asked, answers to questions, anonymized lists of users including students and professionals, who asked and answered those questions, and tags of topics. There are several other files as well. The provided data is relatively well organized, but turned out to be in dire need of preprocessing and cleaning up, which I will address in more detail later. Each question and answer had associated data such as the question body, title, tags, and userID of the question author and answer author. Below is an example of one question element.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| questions\_id | questions\_author\_id | questions\_date\_added | questions\_title | questions\_body |
| 332a511f1569444485cf7a7a556a5e54 | 8f6f374ffd834d258ab69d376dd998f5 | 2016-04-26 11:14:26 UTC+0000 | Teacher career question | What is a maths teacher? what is a maths teacher useful? #college #professor #lecture |

Other files included a list of tags and lists of users by type (student or professional). I discovered that the tag data did not add very much value because much of the time users provided only provided one or two keywords if they provided any at all, and those keywords already existed in the message body. I neglected to include this data, but I did merge the question title and body into a single feature vector.

The answers data was useful because it included lists of who asked questions and who answered them, which allowed me to correlate answering users with the questions they’ve answered. The actual answer text was of little interest for me. A sample answer is included below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| answers\_id | answers\_author\_id | answers\_question\_id | answers\_date\_added | answers\_body |
| 4e5f01128cae4f6d8fd697cec5dca60c | 36ff3b3666df400f956f8335cf53e09e | 332a511f1569444485cf7a7a556a5e54 | 2016-04-29 19:40:14 UTC+0000 | <p>Hi!</p> <p>You are asking a very interesting question. (…) |

The Approach:

I evaluated several possible methods for creating a match between question and answer. I first considered using a perceptron-based approach or an SVM, but I encountered memory issues when trying to standardize the vectors between all questions. English contains a lot of words, so this represented an enormously high-dimensional space that my modest laptop was ill equipped to handle. Specifically, the dictionary I used to map to included 143,000 words.

I also faced an issue with categorization. Obviously, a multi-classifier is possible, but usually this is done with a pre-defined low number of categories c < 10. In trying to ‘classify’ for each user who answers questions, I have introduced two major divergences. First, the number of categories is not explicitly defined, and second, the number of categories is enormously high (thousands). I explored using user profile data such as profession or tag data to reduce the number of categories. Unfortunately, the quality of the data set in this area was very inconsistent. It turned out to be unfeasible to restrict the category space to those with matching categories because too many users were missing tag data. This would have failed to provide useful recommendations for a large segment of users.

I also considered grouping questions together based on the users who answered them and then comparing new questions using a similarity or distance measurement with each cluster of questions. The goal would be to return the users associated with the closest clusters. This would represent a sort of k-means clustering approach, but still with a high-k value. I opted not to do this because I felt that it might consistently direct users to answer too-similar questions.

I decided to use an item-item comparison using the questions in the dataset. In fact, the dataset provided consisted mostly of answered questions. I used a cosine similarity method to compare normalized feature vectors between pairs of comments and kept the top five comments returned from a comparison against the database. Then, I looked at users who answered those questions and provided a truncated list (for brevity and consistency) of users who answered the most similar questions. I chose to return the top five users from those top five questions. Given that the dataset consisted mostly of answered questions, nearly all queries returned five users likely to answer a given question.

This approach does not funnel users toward the same type of questions consistently, but it also does not guarantee that every user will have suggested questions each round. I personally feel that avoiding a narrow recommendation channel is worth the trade-off, but I can understand that CareerVillage might want a more predictable recommendation engine that can shoot a set number of questions to each user each x period.

Challenges:

I have discussed a few of the challenges that dicated my approach to solving this problem above, but I did have some more mechanical challenges.

Much of the data included punctuation, common words, html tags, urls, and other junk that did not help with a keyword based approach. I spent quite some time learning how to clean up the data before processing it. I also had to massage the data as I loaded it from the files so that important details could be accessed quickly. Largely, this entailed loading up files into relevant dictionaries.

An initial problem solving approach used sparse feature vectors out of a pre-defined dictionary, but this used far too much memory, as the dictionary contained roughly 143,000 words even after removing common words. I changed my approach to compare the feature vectors normalized between two questions at runtime rather than comparing input against pre-computed vectors. The normalized vectors are almost always fewer than 100 features. This is much faster and more memory efficient.

Examples of the solution:

The solution comes in the form of the questionID followed by a list of most eligible users who should be notified of the question in the next email blast.

One example of this follows:

49c8360cbe854647a989f92f138ab1c5 :['5fffce2c016940ad96766dffbcecfe37', '819f5f5dbb664501ab3e300d96e37e24', '2a71e2b2f66943749cad7dcc36c1202d', '89ffc7c3266344b68c2dbce832d102f6', '239f56f5ad5546bb98e32dfd79b934c8']

This corresponds to question:

|  |  |
| --- | --- |
| What is the role and responsibilities of a project manager at a nonprofit organization? | I want to work for a nonprofit organization and I am trying to figure out what position would be best suited for me. #project-management #nonprofits #non-profit #non-profits #project-manager |

The users list corresponds as follows:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 5fffce2c016940ad96766dffbcecfe37 | | Menlo Park, California | | Financial Services and Non-Profit Management | | ▶▶ Program & Project Management ▶▶ |
| 819f5f5dbb664501ab3e300d96e37e24 | Washington | | Financial Services | | Manager of Portfolio Management | | |
| 2a71e2b2f66943749cad7dcc36c1202d | Seattle, Washington | | Management Consulting | | Associate - Management Consulting at PwC | | | |
| 89ffc7c3266344b68c2dbce832d102f6 | Allen, Texas | | E-learning | | Aspiring adviser to STEM teens. Explains clearly, exhorts best practices. Raspberry Pi zealot. | | | |
| 239f56f5ad5546bb98e32dfd79b934c8 | Los Angeles, California | | Computer Games | | Developer at FunWall inc. | | | |

The above list is presented in order of relevance according to previously answered questions and this example was picked entirely at random. I would say that the algorithm has successfully matched the question to users who would be suited to answering the question, particularly the top three.

Analysis and deeper dive:

The above section provides an example of the solution, but one answer is hardly enough evidence that this approach provides consistent high-value recommendations. In fact, analysis of recommendation algorithms turns out to be a relatively challenging task. One cannot precisely compare labeled data as one can with perceptrons and the like, since there is not a finite number of pre-defined categories. A live test might be a better way to measure the effectiveness. That is, check user engagement versus another algorithm or “no” algorithm (most recent/random). Certainly, a computational analysis of the effectiveness of the recommender does not seem forthcoming. In fact, that is essentially what the recommender is trying to do. Given my lack of access to running a live test, a hand labeled comparison, then, is the best remaining method to evaluate the effectiveness. Essentially, it’s a judgment call.

To see if my recommendation system is working, I compared 275 of my “top 5 recommendations” against the set of users who answered a question. That is, I tested to see if any of my top 5 recommendations matched a user who answered the question. Necessarily, the set providing samples for recommendation is partitioned from the set that is being compared against. Therefore, an exact match was not possible. Unfortunately, not one of my recommendation sets overlapped with a set of users who did answer any of those 275 questions. At first, this appears to be a failure of my algorithm. However, we should consider the fact that CareerVillage is asking for help with a better recommendation system and that the most relevant questions may not be currently surfaced to the most relevant users.

It is not feasible for me to manually compare 275\*5 users against the input question, but I will defend my recommendation algorithm by providing the user profile of the top recommended user for 10 recommendations and the user profile of the user who actually answered those questions. This is by no means a perfect metric. The user profiles that CareerVillage stores is usually extremely terse and focuses mostly on someone’s stated professional credentials and less on their life experience. This comparison is attached at the end of the essay.

In summary, of the 10 manually inspected question and answer sets, my algorithm came out ahead of reality twice, lost three times, and tied five times.

Conclusion:

This was a challenging task, given that I am still new to the concept of data mining and that this is a project with little pre-defined structure. I feel that I picked a task with an appropriate difficulty and executed it successfully, for the most part. I do feel that further tweaks and refinements to this algorithm could still improve the accuracy of the results, but as far as the code is concerned, I am satisfied with the run-time given a dataset of roughly 24,000 questions and 51,000 answers. In this dataset most but not all questions have answers already. I have selected to only process the first 10,0000 questions and select the input questions from the remainder. This simulates a real-world application, where we will not be able to know who has answered a similar question already.

Lessons:

If I were to start this project over again from scratch today, I feel that I would end up in roughly the same place I am now. However, I would probably take a more direct route and explore fewer dead-ends. I think I learned something about the nature of the firehose, that is the vast quantity of data requires an efficient and well-thought-out approach to processing it. A brute-force approach will not work (see using an English dictionary as a feature vector template). I think given more time I would add additional checks to ensure relevance to further filter the relevant results rather than simply truncating them. In my current recommendations, I got a few that seemed not exactly relevant. The example where my recommender suggested a college counselor instead of a computer science person (last example) stands out. Seeing the comment, there are more words about school and education than there are about computer science, so I understand the mistake, but it didn’t quite get the message exactly. Probably filtering results by matching tags when possible would help to improve reliability.

Below, you will find my hand-labeled questions matched with recommended answerer and actual answerer. Determination of the winner was made by my judgement of which profile credentials seemed most relevant to the question asked.

Q: Question  
UR: User recommended

UA: Actual user who answered

Algorithm Wins  
Reality Wins

Tie

Q1: Where can I find discounted textbooks? Textbooks are so expensive and I am wondering how else I can access used textbooks... #textbooks

UR1: Automotive Mechanical Engineer I Automotive

UA1: Legal Services Employment Counselor | Open Records Specialist

Q2: How did you get a job as a social worker? I'm going to college to get a degree in social work so I'm curious as to what would happen after college. #work #social

UR2: Telecommunications Sr. Project Manager

UA2: Individual and Family Services LMSW

Q3: Is there a personality difference between PA's and DNP's and if so, should I consider this when making a career choice? I understand that PA's tend to be task oriented, check it off the list, get it done and move on type of personalities vs nurses who might have a similar job description, but tend to be more compassionate and nurturing. I definitely tend to lean towards the PA type but my college of choice does not have a PA school. Therefore, should I consider nursing school instead? #nursing #physician #hospital-and-health-care #hospital #medicine #college

UR3: Mental Health Care Assist with Recognizing and Developing Potential

UA3: Hospital and Health Care Healthcare Recruiter

Q4: Will I need to observe all grade levels if I only plan to teach band in high school or above? I plan on becoming a band director in a high school and possibly a private school or boarding school later on in my career. I’m curious when I am getting my teaching degree if I will have to observe younger classes like elementary and middle school if I don’t plan on teaching in those environments. #music #teaching

UR4: Higher Education Higher Education Professional

UA4: President/CEO

Q5: What is the engineering field in the largest demand today? Perspective Engineering student #engineering

UR5: Computer Hardware Product Technologist - Dell Solutions Excellence Center

UA5: Information Technology and Services Corporate Development & Strategy Director at Dell

Q6: Would a Biomedical Engineer recommend getting a Biomedical Eng. degree or get a Mechanical engineer degree and move laterally to the biomedical field? Considering Both options as I get apply to engineering programs. #engineering #mechanical-engineering #biomedical-engineering #engineers

UR6: Higher Education Professor at Ivy Tech Community College

UA6: Defense and Space Retired Aerospace Engineer and Rocket Scientist

Q7: Is it better to be college-ready or career ready? We are talking about this but which is more important? Should one be more required than the other? It's confusing me because we are constantly told that college must be a priority but yet there are places you can get a full-time job by just passing the Work Keys test and get paid a pretty decent salary. #college #career #job #career-paths #priorities

UR7: Entertainment Music Producer-Composer-Arranger-Saxophone Recording Artist at Robert-O Productions

UA7: Consumer Electronics Retail Online Coach

Q8: Is age a factor for hiring entry level computer engineers? I am in my mid thirties with a Bachelor's degree in computer engineering but no professional experience. After graduation I worked in other fields just to gain a living. I am looking for a job as programmer, but my age and lack of hands on experience seem to block me. I had couple of internships, I have worked on some projects in my free time just to keep up with market demands, but it doesn't seem to satisfy employers. How can I break through? #computer-science #programming #computer-engineering #java

UR8: Higher Education Higher Education Professional

UA8: Computer Software Senior Software Developer - AVP at Barclaycard

Q9: What kind of internships are helpful, and available for English majors? I've heard that it is a bit tough to get a job with an English major; so, i was wondering if an internship may help and what those may look like. #literature #english #career #higher-education #humanities

UR9: Sales Strategy

UA9: Accounting Director, National Deals Risk & Quality at PwC Canada

Q10: After trying all you can do and not getting the desired job in you career field what do you do? I don't want to put all of my eggs in one basket, but I also don't want to give up. My career field is competitive and I don't know if its reachable. I don't want to give up but I also need advise. Everyone seems to think I'm over my head and shooting way beyond the stars but I don't have a backup and I really want to reach for this one dream. #acting #film #theatre #actor

UR10: Information Technology and Services, Franchise Owner, Commercial Cleaning Business, Recruiter/Career Counselor Career Guru

UA10: Internet Training Program Manager