Data Science for Good: Recommendation System

Destiny Adams
Hannah Roe
Harrison Ratcliffe
Olivia Lyons
Sebastian Cortes

Introduction

- Followed a Data Science for Good challenge from Kaggle, posted by CareerVillage.org
- CareerVillage is a forum website like stack overflow, but for career questions.
    - Professionals can volunteer their time to answer questions from students
    - Wanted to improve their method for recommending questions to professionals for their daily/weekly/monthly emails
- We'll discuss our: Problem, Data, Process, Review and Results

“Can we develop a way to accurately and effectively recommend students’ questions to qualified professionals who are most likely to answer them?”

Why This Question Is Important

- Many students cannot get the career advice they need
  - Current student to guidance counselor ratio in the U.S. is about 500:1
  - Too busy to search out career resources
  - Limited access to role models in desired career
  - Lack of resources in community
- Since CareerVillage provides a way for students to get advice and ask questions, it's important for their matching of questions to professionals to be as accurate as possible
  - This benefits Student, Professionals, and CareerVillage.org as a platform

Why This Question Is Important, Cont.

By developing a way to accurately and effectively recommend students' questions to qualified professionals who are most likely to answer them:

- Students can receive better advice, may be encouraged to ask more questions, recommend CareerVillage to others
- Professionals get questions directly related to them, allowing them to provide better answers and encouraging them to answer more
- CareerVillage.org becomes a more reliable platform for students to receive career advice, bringing in more users

Why We Chose This Question

- As students ourselves, we understand the pressure and confusion of picking a career path
- We would like to help students receive valuable career advice and achieve their goals
- Since this question was hosted on Kaggle, gathering the data was simple and easy, and we could view other peoples approach to the problem

Various approaches

- Content-based filters
  - Suggestions based on user activity

- Collaborative filters
  - Suggestions based on activity of similar users

- We chose a content-based filter

Goal

- The goal here is to be able to specify a professional’s ID and recommend ten questions to the professional that they are likely to be able to help with based on questions they’ve already answered.

- How do we do this?

Methodology

- Data Cleaning
  - Professionals.csv - removed inactive professionals
  - Questions.csv - removed author ID and date

- Keyword Extraction
  - Rake()
  - New column of keywords

Methodology, cont.

- CountVectorizer()
  - Counted frequency of all question keywords

- Cosine Matrix
  - Shows similarity of questions based on keyword frequency

- Recommender
  - Pick ten most similar questions

Data

- Kaggle & CareerVillage provided our dataset

- 15 .CSV files

Our dataset

- Answers
- Answers_scores
- Comments
- Emails
- Group_memberships
- Groups
- Matches
- Professionals
- Questions
- Questions_scores
- School_memberships
- Students
- Tag_questions
- Tag_users
- Tags

Our dataset

- **Answers**
- Answers_scores
- Comments
- Emails
- Group_memberships
- Groups
- Matches
- **Professionals**
- **Questions**
- Questions_scores
- School_memberships
- Students
- Tag_questions
- Tag_users
- Tags

Why those files?

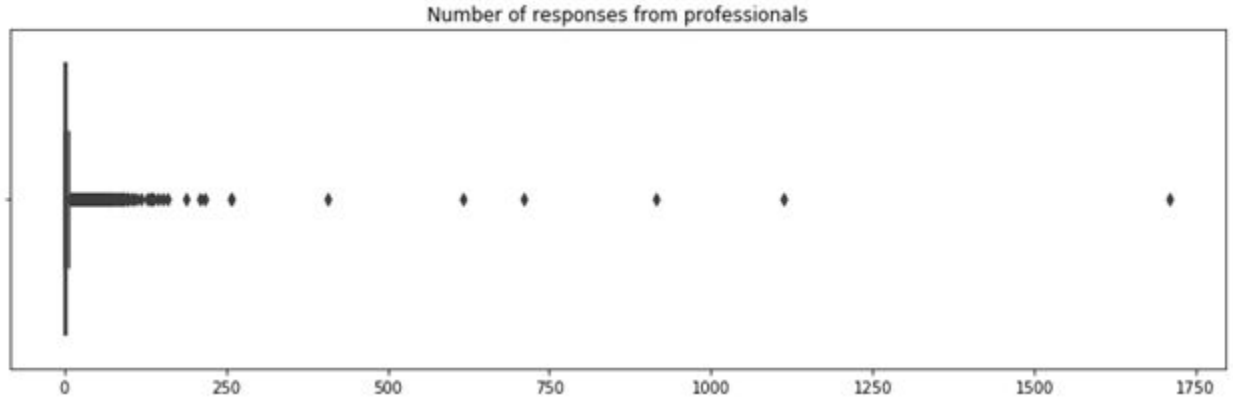
- Other files were deemed unnecessary
  - Could be used for future recommendation updates
- We can cross-reference data between those files
- Professionals: ID
- Answers: Author (Professional) ID, Question ID
- Questions: Author (Student) ID, Question ID
- By looking at a professionals' ID, we can discover answers they've provided to retrieve keywords

Data cleaning

- 12 CSV files dropped
- Professionals
  - Remove professionals who did answer questions
    - About 20,000 of 30,000
  - Scanned through answers.csv and check author IDs
- Questions
  - Questions' title and body were used to detect keywords using rake()
  - Resulting keywords had special characters
  - Regex function removed these special characters

Data cleaning

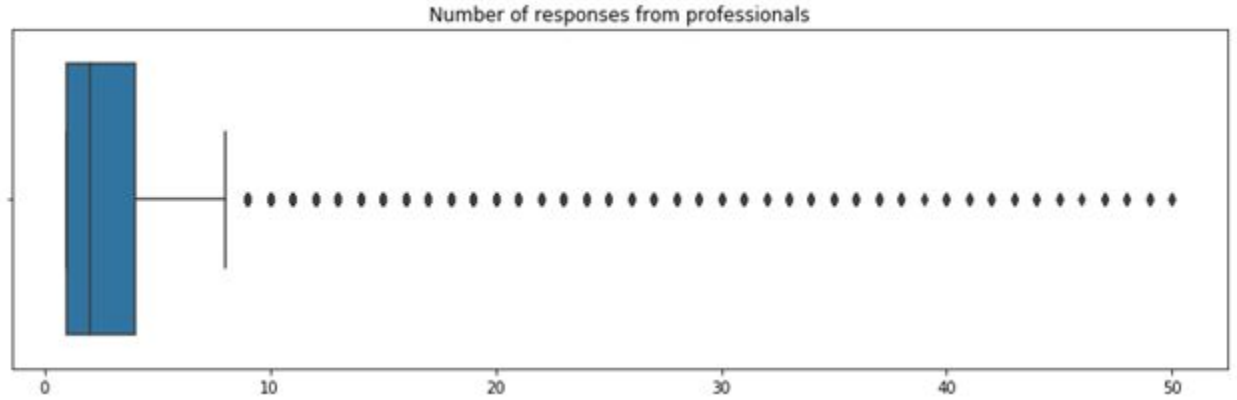
- While looking into answers.csv, a boxplot was created to see how many questions each individual answered.
  - 51123 Answers
  - 10169 Authors



Data cleaning

- Large outliers
  - Removed individuals who answered more than 50 times.

```
Outliers removed: 95
Top 10 number of answers from professionals [1710, 1112, 915, 711, 616, 406, 259, 257, 217, 210]
```



```
Importing......done!
Performing data cleaning...(this might take a few minutes)......pruning extra professionals....done!
generating keywords......done!
Beginning count vectorizer......done!
Now creating count matrix......done!
Now creating cosine similarity matrix......done!
enter a professional id# (just press return for a default ID for testing):

10 questions to recommend to this professional:
What are some things that interest you about Auto Mechanics?
How long have you studied being a mechanic
How to get in to auto body?
What colleges are good for auto mechanic/ engineering?
What advice could someone offer an 8th grader, who wants to be a auto mechanic?
What are auto Mechanic colleges to start looking at?
Where is a good place to start training for auto mechanics?
What's a good head start for going in to auto mechanics
Anybody been through the automotive mechanics program at tcc
Best internships for auto mechanic?
```

Results

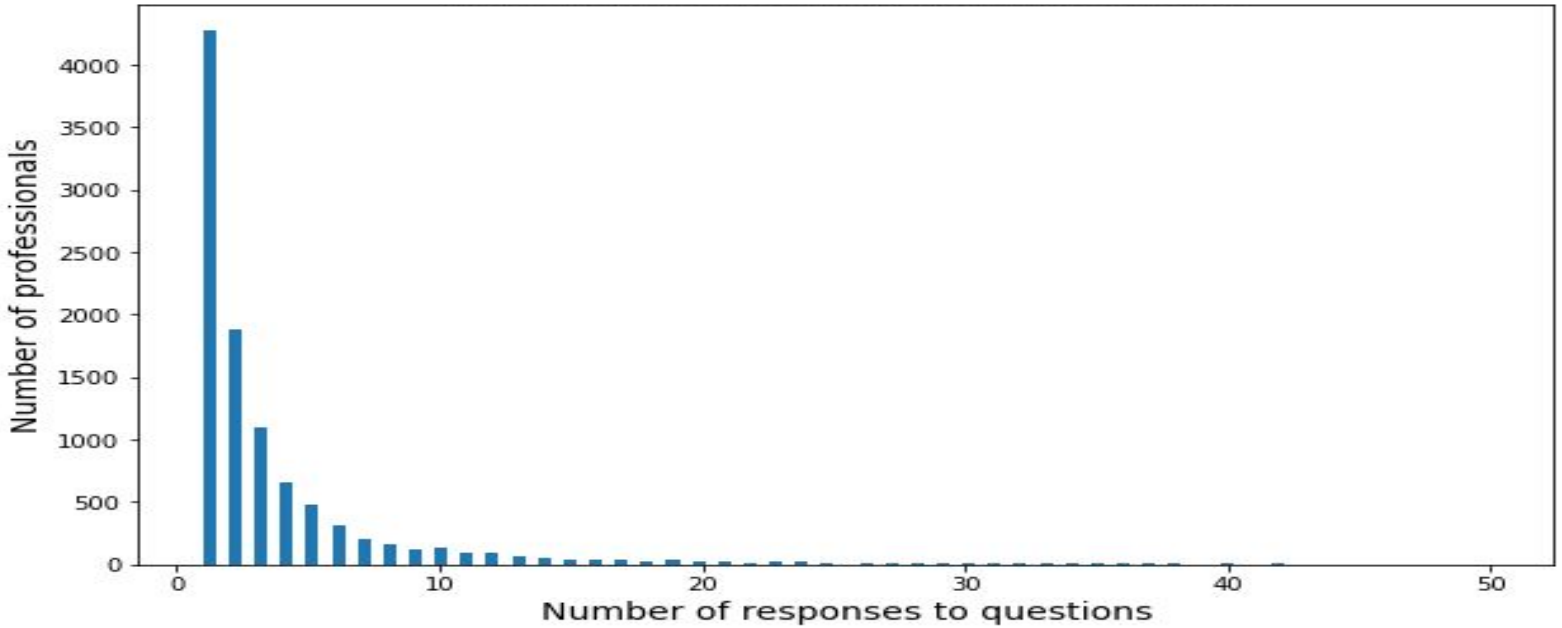
- Successful in getting related question recommendations!
  - Given a professional id#, our program spits out the top 10 recommended questions
- Not recommended for industry-level, but worked great as an insight into how modern recommendation systems work
  - Problems with content based systems (seen previously)
- A couple possible mismatches, possibly due to the way the keyword rake was performed

```
Importing......done!
Performing data cleaning...(this might take a few minutes)......pruning extra professionals....done!
generating keywords......done!
Beginning count vectorizer......done!
Now creating count matrix......done!
Now creating cosine similarity matrix......done!
enter a professional id# (just press return for a default ID for testing): 01b3e033848b41f6b7a55cefc59ba61a

10 questions to recommend to this professional:
How many years of school do you have to go through to be an vet?
How many years of school to I have to go through in total if i want to become a Pediatrician
I want to work with animals when I grow up.What can I do?
What is the hardest or most difficult thing to learn while studying for a Veterinarian?
Do veterinarians work with all animals or can they work with a certain group/species?
how many years do i have to be in school to be a lawyer?
what can i do to work with animals
What are some requirements to become a veterinarian?
what type of vet would you recommend the most?
Do you have to work with all kinds of animals when becoming a veterinarian?
```

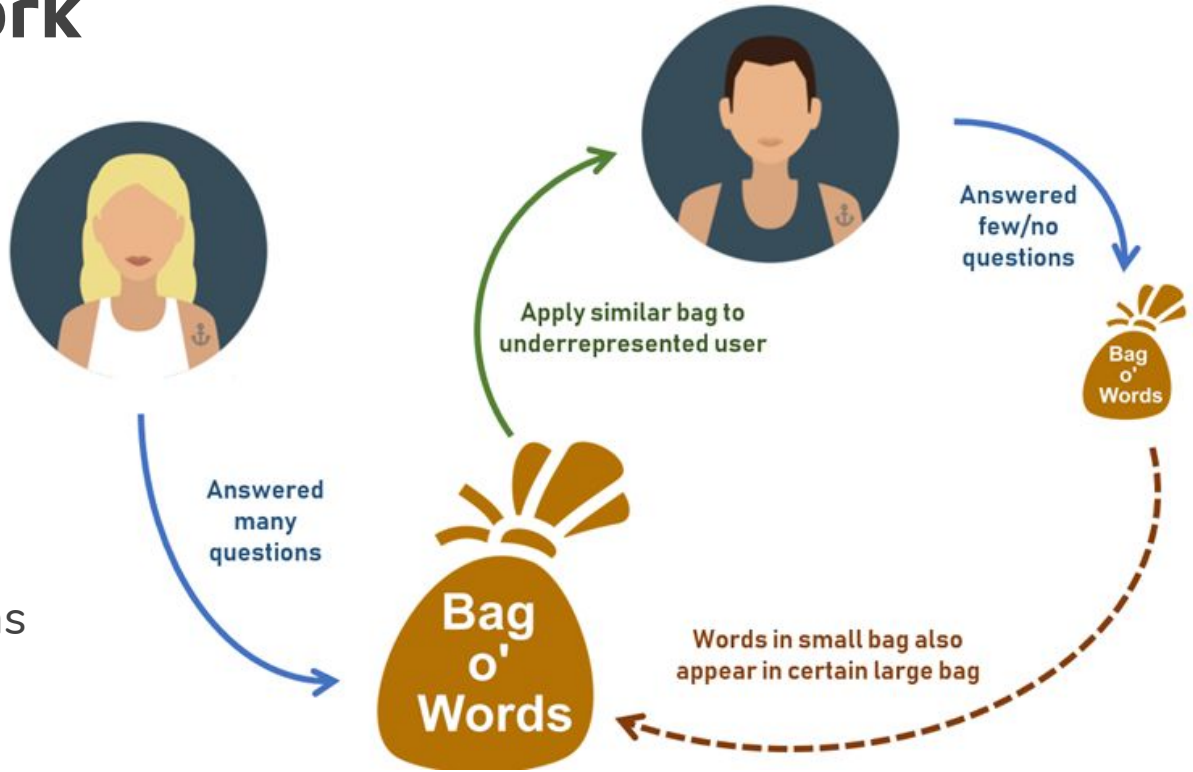
Further Work

Problem: Our current solution doesn't account for 64% of professionals in the data provided by CareerVillage.



Further Work

- Small bag-of-words
- Collaborative filtering
  - By user information
  - By group
- Machine learning for Location-based questions
- Email frequency preferences



Yeehaw?

(that’s cowboy for ‘any questions?’)

References

Herlocker, Jonathan L., Joseph A. Konstan, Loren G. Terveen, and John T. Reidl. “Evaluating Collaborative Filtering Recommender Systems.” *ACM Transactions on Information Systems, Volume 22 Issue 1*, January 2004, Pages 5-53

Nakkula, M., Danylchuk, L., Miller, K., & Tamerler, K. “Promoting Career Development with Low-Income Students of Color.” *Compelling counseling interventions: Celebrating VISTAS' fifth anniversary*, 2008, pp. 115-124, https://www.counseling.org/resources/library/vistas/2008-V-Print-complete-PDFs-for-ACA/Nakkula_Article_12.pdf. Accessed 21 April 2019.

National Association for College Admission Counseling (NACAC), and American School Counselor Association (ASCA). *NACAC and ASCA State-by-State Student-to-Counselor Ratio Report*. National Association for College Admission Counseling (NACAC) and American School Counselor Association (ASCA). https://www.schoolcounselor.org/asca/media/asca/Publications/ratioreport.pdf. Accessed 23 April 2019.

Van den Broeck, J., Cunningham, S., Eeckels, R., Herbst, K. “Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities”. *PLos Med*. https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020267. Accessed 23 April 2019.

Rose et al, “Automatic Keyword Extraction from Individual Documents”. *Text Mining: Applications and Theory*. Wiley, 2010.

Kwak, S. K., Kim, J. H. “Statistical data preparation: management of missing values and outliers”. *Korean Journal of Anesthesiology*. 2017 Aug., pp. 407-411.

Gong, SongJie. “A Collaborative Filtering Recommendation Algorithm Based on User Clustering and Item Clustering.” *Jsoftware.us*, Journal of Software, July 2010, www.jsoftware.us/vol5/jsw0507-9.pdf

McDonald, David W, and Mark S Ackerman. “Expertise Recommender: A Flexible Recommendation System and Architecture.” *ACM Digital Library*, University of California, Irvine, 1 Dec. 2000.

Ricci, Francesco, Rokach, Lior, and Shapira, Bracha. “Introductions to Recommender Systems Handbook”. *Recommender Systems Handbook*. Boston: *Springer*, 2011. pp. 1-35.