Overview

Weeks 1 & 2 - **September 18th - November 3rd**

- Brainstorm ideas about what it means for something to be compatible. Can be gross assumptions.
- Search for literary articles, research papers, and applicable databases to get more understanding on previous research on recommender systems.
- Understand the difference between previous research on recommender systems and how our research to create a new recommender system based on compatibility is different.
- Discuss with Dr. James Caverlee about presenting work publicly in Infolab in the following weeks after we have finished reading research papers on our topic. We will be giving a presentation about previous research papers that we read and what we have learned. This will allow other students in Infolab to understand our current research topic and how our research topic is different from previous research papers.
- Discuss with Dr. James Caverlee about research compliance approval and whether or not we need research compliance approval with our current research project and future plans. If the answer is yes, we will need to attend a Research Compliance Informational and contact the office of Research Compliance & Biosafety and complete training/fill out forms to continue.
- Think more in depth about research question and describe even further what is means to be compatible based on literary articles.
- Start thesis statement for thesis, begin creating a thesis template, and download the thesis template that we want to use throughout the program.
- Schedule project meetings with Dr. James Caverlee (faculty advisor) and Yin Zhang (graduate student advisor) weekly to discuss thesis template, installments, progress reports, thesis statement, and research progress.

Schedule

Tuesday, September 19th, 2017 Notes:

- Had weekly meeting with Dr. James Caverlee discussing research progress (getting datasets).
- Scheduled with Dr. James Caverlee for a research paper/project presentation on October 18th.
- We will be presenting our top choice research article and how to improve the implementation as well as tying in our relation to these articles (introducing our project).
- We will include positive and negative aspects of the research article.
- Discussed with Dr. James Caverlee and agreed that we DO need research compliance approval. We understand that we will have to attend a Research Compliance Informational and contact the office of Research Compliance & Biosafety and complete training/fill out forms to continue.

Wednesday, September 20th, 2017 Notes:

- We are looking at the following link for applicable data about products featured on Amazon to get more understanding on previous research on recommender systems: http://jmcauley.ucsd.edu/data/amazon/

- Inside the data set provided by Julian McAuley, we will be analyzing the “small” subsets for experimentation data provided as well as the metadata datasets.
- In this particular case, we began analyzing the electronics dataset as mentioned in our thesis.
- The goal now would be to utilize the small subsets of data for information of the datasets and further apply our research to the complete sets of data.
- The electronic dataset provides useful information about reviews for electronics and the metadata provides information on products with their “also bought” and “also viewed” information.
- Unfortunately we are not able to access the metadata information without requesting it from Julian McAuley directly. We are going to draft an email and ask for the data.
- Furthermore, the description of the product is not mentioned in the sample JSON provided but is mentioned in the description of the metadata. If the description of the item is not provided, our solution is to potentially use Amazon Product API and request that information manually.
- We hope to get the metadata and image features data by the end of next week, September 29th, 2017 so that we can analyze it.
- In the meantime, we will create scripts to assume the format of the JSON to read in the file into memory from mock data to prepare for when we receive the data and commit them to source control (Github) in Python.
- 2:59 PM, sent out email to Julian McAuley about the request access mentioned below.

Julian McAuley (julian.mcauley@gmail.com)
Requesting for Metadata and image features data

Email (Access to Amazon review and product data):
Hello Professor McAuley,

My name is Kevin J Nguyen, and I’m an undergraduate research student at Texas A&M University. I have found your research on the Amazon recommendation products fascinating. Currently with Victoria Wei (CC’ed on the email), we are researching under Dr. James Caverlee (also CC’ed on the email) on specific use cases in recommendation systems where compatibility might provide more relevant results for recommendation systems.

After taking a look at your research papers and the dataset you have provided online, we would like to request access to the per-category information of electronics on metadata and image features. We believe that access to this data will provide a tremendous asset to our undergraduate research. Please let me know if this is possible.

Thank you,

Kevin J Nguyen.

Thursday, September 21st, 2017 Notes:

- Dr. Julian McAuley kindly provided us with the working links in minutes!

- The working links are located at: http://jmcauley.ucsd.edu/data/amazon/links.html.
- We are in the process of downloading the image features and electronics metadata into our laptops. Because the dataset is big, it will take ~1 day.

Monday, October 2nd, 2017 Notes:

- Define traditional recommendation methods specifically for item-based recommendations and list advantages and disadvantages of each:
    - Collaborative filtering - Collaborative filtering is the process at which social filtering occurs by using item-based recommendations. For example, if someone recommends an item to their friends (if they all share similar interests in items) then this recommendation would be proven more useful than other non-social recommendations. Collaborative filtering is based on the fact that if a person likes an item in the past, then similar items will be recommended to them assuming that this person still continues to like that particular item. Instead of the traditional neighborhood-based approach where a select number of people that share similar interests to the active user can predict the active user's interests, we will be using the item-to-item based approach where we measure the correlations between different items to see which items belong the closest together to make our recommendations for users.
        - Advantages
            - More accurate recommendations based on social filtering/environment (bandwagoning is a popular activity in our generation).
            - Less user action needed to find items based on previously-bought items.
        - Disadvantages
            - The sparsity problem is a big disadvantage in collaborative filtering. This is an issue where as the number of items increases, the number of items users have rated decreases, which decreases the correlation coefficient, making it very small and less reliable. Therefore, when we compare between two users, it may be portrayed that the number of items that they have in common is very small compared to the total number of items.
            - Some users may start to dislike the items that they previously bought, making the recommendation not as reliable. Users may also no longer need items from similar categories because they have already bought what they needed.
    - Matrix factorization - Matrix factorization is the mathematical way of factorizing a matrix. For example, if we have a matrix, we can find two matrices such that when we multiply these two matrices together, we will end up with the original matrix. This can be used in collaborative filtering methods because we can use matrix factorization to predict which items users would want to use based on other users and their ratings of what they like and what they don't like.
        - Advantages
            - Ability to predict items for users who have similar interests.

            - Different types of matrix factorization for different recommendation methods.
            - Ability to detect latent features and variables during decomposition of recommendation models. These features underlie the interactions between two different entities for further discussion.
            - Reduces dimensions when working with matrices.
        - Disadvantages
            - Doesn't allow for predictions based upon item compatibility.
            - Takes time to complete all variations of factorizations for item compatibility.

- Downloaded the following files from the URL provided:
    - Metadata for all categories: **metadata.json.gz**
        - The metadata includes the following information: descriptions, price, sales-rank, brand info, co-purchasing links, image url, titles, and categories. We believe that we can utilize this information after we have analyzed the data for electronics.
    - Metadata for Electronics: **meta_Electronics.json.gz**
        - Like the overall metadata, the filtered categories have the same fields of data. Our goal is to analyze and graph out how this data is related to one another so that we can effectively define our definition of compatibility and our assumptions about the data.
    - Metadata for Cell Phones and Accessories: **meta_Cell_Phones_and_Accessories.json.gz**
        - The metadata for cell phones and accessories could provide insight in the electronic definition of compatibility. By utilizing this dataset, we go under the assumption that cell phones and accessories encompasses our research of electronics. The benefit here is that usually for electronics, all the products may be physical extensions or actual electronics. By utilizing cell phone and accessories data, we can provide more data on compatibility. Cross referencing metadata for cell phones and electronics may provide useful information on compatibility as well.
    - Question\Answer data for Electronics: **qa_Electronics.json.gz**
        - The question and answer data can contain valuable information about the usability and compatibility of each product. For example, if I am purchasing a USB-C dongle for my 4K TV, I would like to know if the dongle can support 4K, 60 FPS, etc. In this case, question and answers would be better.
    - Question\Answer data for Cell Phones and Accessories: **qa_Cell_Phones_and_Accessories.json.gz**
        - Here is where I am expecting to find a good qualitative match on the concept of compatibility for cell phones and accessories so that we can utilize this to define our definition of compatibility. Users generally will ask whether or not a product is compatible with their current model before purchasing.
    - Review data for Electronics: **reviews_Electronics_5.json.gz**

        - The data collected for the reviews are known as 5-core which means the user has at least 5 reviews which defines their core review and removes the review data from user that could potentially have been made to inflate/deflate the quality of a product. The review of electronic products can have information pertaining to their use with other electronics. For example, a DVD player must be paired with a TV in order to review so in this case we can analyze the compatibility of the devices.
    - Review data from Cell Phones and Accessories: **reviews_Cell_Phones_and_Accessories.json.gz**
        - The review data is also 5 core and the reviews of cell phones and accessories will provide information about certain products with devices which may prove helpful in the expression of compatibility.
- Code for reading these files:
    - The files are written in a gzip compressed format and as a result reading the files is difficult and we must use plugins to extract this data.
    - Specifically McAuley states that the JSON from the files are not proper JSON files so other programing languages besides Python will have issues reading the file in.
    - Luckily, we are both moderately experienced with Python so writing the scripts to read in these files so that we can manipulate the data should not be hard.
    - Kevin has began writing the scripts and has written generic scripts to read in any gzip file from Professor McAuley and print out each unique item.
    - What is next is that each object be its own custom class and then the Python scripts will read them into a database or an in memory store. After reading, it may be more efficient to store this information in a SQL or NoSQL database to make querying for the data faster.
    - Currently, all code is being stored in a private repo at: https://github.com/KevinJ97/CSCE491-Research request access first.

Tuesday, October 3rd, 2017 Notes:

- At 9:30 AM had general meeting with Caverlee and discussed overall performance throughout the past two weeks.
- Informed Caverlee that we will be attending Grace Hopper Conference in Florida and will have limited productivity.
- Summarized our work over the preparation of files, scripts, revisions, and research.
- Revision of the research proposal from the comments provided by the URS.
- Collaborative filtering essentially is a technique to make automatic predictions about a user's interests by utilizing a collection of preferences or tastes from many other users. This is the collaborative part of collaborative filtering, in that each user will contribute in making the overall results better.
- Matrix factorization is a method of calculating preferences of other items not yet rated by utilizing the ratings from previous items. By multiplying the matrices together we can calculate an original matrix which can be used to determine the rating system in preferences such as how Netflix will recommend you a movie with a 95% like rate.

- Compatibility between objects is defined in their systematic similarities in some ways but also their systematic differences.
    - Electronics
        - Similarities can be in brands, substitutes, etc.
        - Differences can be in their components, power consumption, ports, etc.

Friday, October 13th, 2017 Notes:

- Victoria attended the research compliance drop-in session and completed the IRB application for our research project.
- We need PI signoff from Dr. James Caverlee, and we will discuss this in our weekly meeting next week.

Monday, October 16th, 2017 Notes:

- We found a research paper that we will be presenting for our research presentation to Infolab in the coming weeks: http://cseweb.ucsd.edu/~jmcauley/pdfs/icdm16b.pdf. We will be starting to read this research paper and annotating our findings in the coming weeks.
- Setup project for ShareLatex that will utilize the STEM thesis template.
- Added Victoria to the Git Repository.

Thursday, October 19th, 2017 Notes:

- Attended the mandatory undergraduate research scholars orientation at the MSC.

Friday, October 27th, 2017 Notes:

- Finished first progress report and first installment submissions.

Monday, October 30th, 2017 Notes:

- We are reading http://cseweb.ucsd.edu/~jmcauley/pdfs/icdm16b.pdf and annotating our findings with presentation notes. We understand the difference between previous research on recommender systems and how our research to create a new recommender system based on compatibility is different by noting it in our findings in our presentation notes.
- The presentation can be accessed using the link: https://docs.google.com/presentation/d/1ZXcR4gFm3UGeiMOJDd3R2EOrhm-CFaJkML9ghKNjJhY/edit?usp=sharing.

Wednesday, November 1st, 2017 Notes:

- Create presentation for research talk.
- Gave a research presentation on “Learning Compatibility Across Categories for Heterogeneous Item Recommendation” by Ruining He, Charles Packer, and Julian McAuley.
- Dr. Caverlee completed the PI signoff.
- IRB set our IRB application to “Not Human Subjects Research”. We don’t need research compliance approval. Victoria made a comment to the URS to remove our “Pending Research Compliance Approval” status and set it to “Approved”.
- Had a meeting with Dr. Caverlee about:

- Submitting research to Explorations: We will be submitting research to Explorations next year.
- Told him about research progress.
- Thesis statement will be done by next year. We will be starting on our thesis statement soon.