CSC 466 Analytical Project

Jeff McGovern and Nicole Martin
CSC 466, Section 03
December 11, 2015

Abstract

Contents

1 Introduction

Overall the approach taken by this team was to follow our interests and pick a data set that we were interested in applying knowledge and discovery from data (KDD) techniques to. For both the Untapped and Archive of Our Own (AO3) data sets, the format of the data lent itself well to the analysis of frequent item sets and association rules mining.

1.1 Untappd

Untappd is a social media platform for rating beer. Users can share their beer experiences with friends and the nearby public, adding commentary, scores, location, and beer metadata describing the ber they're drinking.

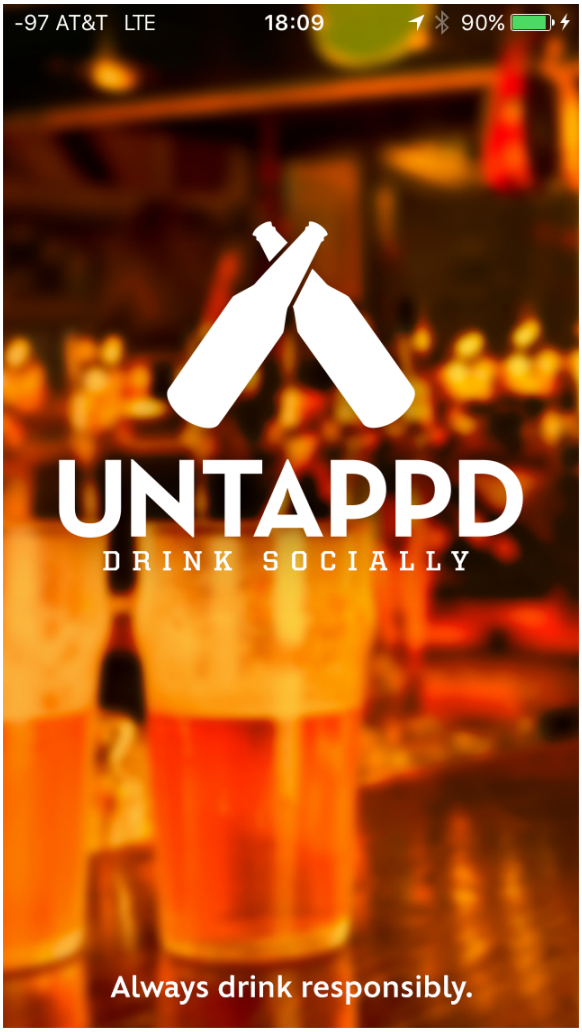

Figure 1: Untappd splash screen

For the Untappd dataset, we had one primary question:

- What beers do users consider equivalent?

This question is relevant for many reasons. For many avid beer enthusiasts, beer discovery is a large part of the exploration process. Much of it occurs through a social network of other beer

enthusiasts, which is part of why there is such a heavy social media aspect to Untappd, but it can be difficult to take into account someone else's relative tastes when receiving a recommendation. Having the ability to recommend a beer to a user by comparing their rating habits to the rating habits of other users can bring tremendous value to the beer enthusiast and will encourage them to use the app more often.

Breweries can also benefit from such an investigation. Knowing what beers users are rating the same can be an indicator of success or failure. For example, some brewers strive to accurately reproduce annual anniversary beers from popular micro breweries, like Firestone Brewing Company, which tend to be a blend of that year's barrel aged works. Witnessing associations between user ratings may affirm some goals that brewers or breweries have.

1.2 Archive Of Our Own

For the AO3 data set, the key questions of interest were related to the use of optional relationship, character, and freeform tags used to categorized the fanworks. These tags are used within the indexing and searching systems on AO3, but are not mandatory. Key questions were:

- Do creator's always tag the characters involved in a relationship tag? - This was relevant because searching for a character by their tag does not also search through all the relationship tags that include the character's name, resulting in relevant fan works being excluded from search results.

- Do characters who are not in the target fandom occur frequently enough to have association rules? - This question is relevant in cases where multiple fandoms exist in the same universe or for crossovers. For example it is possible that works tagged as "Thor (Movies)" may include a character tag for "Steve Rogers" despite him not being in those movies. Or in the case of crossovers, enough content creators may decide to create fanworks with Tenth Doctor from "Doctor Who" that that character tag would occur frequently despite being from a completely unrelated series.

- What association rules exist involving freeform tags? Freeform tags often are used to designate character tropes, fanfic tropes (such as "AU" to designate works that occur in alternate universes from canon), or author commentary. Whether certain character or relationship tags had associations, or if freeform tags had associations with other freeform tags, was of interest for finding writing trends within the fandom.

2 Data Sets

2.1 Untapped

The Untappd dataset is a popular social media platform for checking in, rating, and sharing beers. Much in the way sites like Twitter, Facebook, or Google+ enable users to create a network of followers, post updates, and check into venues, Untappd offers a social media experience focused on the consumption of beer and cider.

Fortunately, there exists an API for app makers who are interested in tapping into Untappd's vast databases of checkins and beers. Unfortunately, the use cases allowed primarily fall under remaking the Untappd app — research is expressly unsupported and storing the data is not allowed. That being said, while this project may be considered research, our hope is to be able to turn this into a beer discovery app. After a few attempts and many weeks of waiting, we were finally given an API key for our `BeerPredictor` app.

2.1.1 Retrieval

The API uses HTTP GET requests to offer data. Essentially, this looks like a URL with the method name and API key encoded into the format. For example, a request for user rectangleboy would look like the following, where "CLIENTID" and "CLIENTSECRET" are the aforementioned API keys:

https://api.untappd.com/v4/user/checkins/rectangleboy?client_id=ID&client_secret=SECRET

2.1.2 Format

The format of the data is JSON. Listing 1 shows an example output from the above call. JSON dictionaries made parsing this dataset very manageable.

2.1.3 Relevant Data

Relevant to this project are three areas of data:

- User Checkins
- Beers
- The Pub

2.1.3.1 User Checkins User checkins contain a wealth of information.
When a user checks in a beer, they can do the following things:

- Rate the beer on a $0, 0.25, 0.5, \ldots, 5$ scale
- Comment on the beer up to 140 characters
- Upload a photo
- Checkin to a venue
- Cross post to Twitter or Facebook

From checkins, we chose to focus primarily on the beer checked in and its rating.

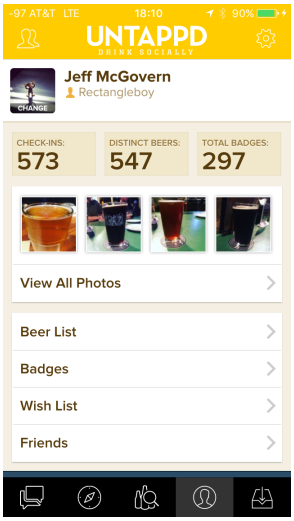

Figure 2: User information for user rectangleboy.

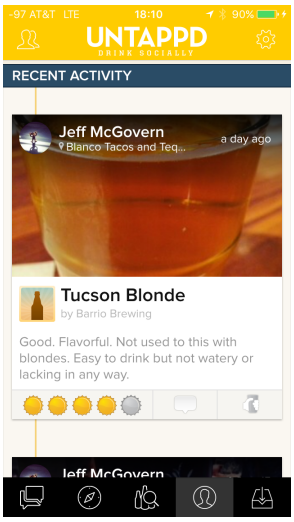

Figure 3: An example of user activity, showing already checked in beers.

2.1.3.2 Beers For the purposes of this investigation, beers are uniquely defined by the brewery name and the beer name. While there may be some corner cases where obscure, micro, or home breweries may have the same name, since there is such a vast number of them, we felt that this was a trivial issue.

2.1.3.3 The Pub "The Pub" is the name Untappd gives to the local feed of public beer check-ins. Using latitude-longitude coordinates, the API can give a list of the most recent checkins within a 25 miles radius, allowing users to see checkins near their location.

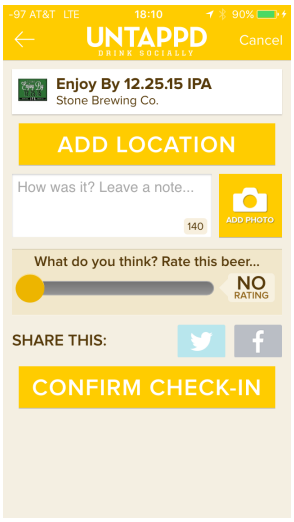

Figure 4: A user checking in a new beer.

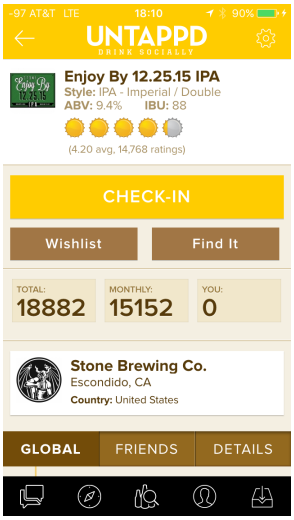

Figure 5: The information page for a beer, showing the brewery it came from and the beer's name.

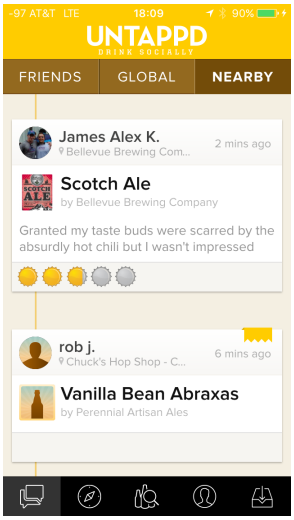

Figure 6: Nearby checkins referred to in the API as "The Pub."

2.2 Archive Of Our Own

This data set was selected because one of the authors volunteers at fan conventions, finds the dynamics of fan communities fascinating, and assists with editing and proofreading their author friends' writing, both for original fiction and fanfic.

2.2.1 Context

For as long as people have produced creative works, people who are fans of those creative works have created their own pieces that celebrate, explore, extend, subvert, or otherwise play with the canon of the original piece. Often the line between what considered "art" vs "fanart" or a "story" vs a "fanfic" and the like can be hard to determine. For example, all church-commissioned artwork of biblical stories and saints could be considered bible fanart since the people who commissioned the works were not the authors of the bible. For literary works, the argument can also be made that Dante Alighieri's *Divine Comedy*, the first part *Inferno* being the most well known, is self-insert fanfic based on the Bible's canon. For a more recent example of this phenomenon, it is well known that the notorious *Fifty Shades of Grey* by E.L. James started out as a fanfic of the also notorious *Twilight* novel series by Stephenie Meyer.

With the internet and technological improvements, the ability to create, distribute, and access fanworks has increased greatly in the last decade. Prior to the internet, fanworks would be available either at fan conventions or in fan-published magazines (zines) which were not widely printed or distributed. However, given the wide variety of fan communities (fandoms) there was no central organization to publishing or archiving these fanworks. Many authors created their own websites to archive their work or used blogging and journal websites such as LiveJournal and Dreamwidth. Others would create their own individual archives to store other fanworks within a given fandom. All of these websites were subject to being abandoned or deleted by their owners. In response to this, more centralized, standardized, multi-fandom archiving websites emerged such as FanFiction.net and ArchiveOfOurOwn.org.

The Organization for Transformative Works (OTW) is a nonprofit organization that provides access to and preserves the history of fanworks and fan cultures. Archive of Our Own (commonly referred to as AO3 and available at archiveofourown.org) is a fanwork preservation project under OTW. AO3 describes itself as "a fan-created, fan-run, non-profit, non-commercial archive for transformative fanworks, like fanfiction, fanart, fan videos, and podfic", and the website is well known in online fandom communities. As of December 7th, AO3 contained content from 19,973 fandoms, had 718,524 registered users, and 1,976,945 works available on the website.

2.2.2 Scope of Data

The scope of this analysis was limited to fan works that were tagged as a Marvel TV show because they provide an interesting case of interrelated series based on an extended comic book canon while still limiting the scope of the data set. Since Iron Fist and Luke Cage did not have any tagged fanworks in the archive when data was pulled, the data set was limited to:

- Agent Carter (TV) - 2,434 works
- Agents of S.H.I.E.L.D. (TV) - 13,757 works
- Daredevil (TV) - 2,895 works
- Jessica Jones - 193 works

- The Defenders (Marvel TV) - 4 works

It is worth noting that fan works can be tagged with multiple series, so though it appears that there are a total of 19,283 distinct fan works in this data set, when all of the data was processed together there where only 19,019 distinct works present when de-duplicated based on the unique id for each fan work. For tag analysis, only the tags within these works that occurred in more than 0.5% of this collection were collected and used. This is because AO3 encourages creators to include custom tags and because AO3 does not index tags for searches until after they reach a certain usage threshold. This means that, similarly to Tumblr tags, it is not uncommon for the creators to essentially talk to themselves within the tags as they please after including the indexing tags, which leads to a lot of tags that only occur once in the data set and are not relevant for association rules mining.

2.2.3 Retrieval

The htm pages from Archive of Our Own were scraped from archiveofourown.org via Python's urllib library and the data for each fan work was extracted using the BeautifulSoup (bs4) library.

2.2.4 Format

Within these data pools, originally only the tags that occurred in more than 0.5% of the works were used for this analysis. Of these tags, the type breakdown was:

- Relationship tags - consist of 2 character names separated by either a "/" or a "&" mark to indicate romantic and platonic relationships respectively.
- Character tags - consist of the name of a character who is in the story. Typically used to indicate characters that are integral to the plot. Are sometimes omitted by the creator if the character is included in a relationship tag.
- Freeform tags - tags used to indicate other aspects of the story, such as content, warnings, tropes, and assorted creator notes. These tags do not have restrictions on their content.

There are additional required tags used by AO3, such as authors, content warnings, rating, language, word count, chapter count, comments, number hits, and number of "kudos" which is the AO3 equivalent of giving a fanwork a thumbs up. These tags were not used for association rule mining. The `ao3parser_collection_tags.py` script parsed all htm files in the current directory for these selected tag types. The selected tags were then organized into a sparse vector format with a generated `alltags.csv` file storing the tag information and an `alldata-out1.csv` file to store the sparse vector of the tags occurring in each fanwork identified by their unique id scraped from the htm pages. Tags that occurred in fewer that 0.5% of the works were treated as outliers and excluded from the vector representation.

3 Methods & Results

3.1 Untapped

To gather data for this investigation, we chose to inspect the local feed within 25 miles of Spike's — The American Pub, a popular local venue in downtown San Luis Obispo, in order to find users to pull data from. Its (latitude,longitude) coordinates are (35.278205, -120.666584). From there, we traversed into each user's checkin history and pulled everything we could before the API limit restricted us.

With our original question in mind:

- What beers do users consider equivalent?

we sought to use Association Rule Mining to answer it. In order to do so, we need a notion of a "transaction." For this dataset, we felt that the most interesting way to encapsulate that was by creating a transaction from each user's beers that have been given a certain rating. For example, if `rectangleboy` gave Sculpin IPA, Black Butte Porter, and Lizard's Mouth a 5 star rating, then those three beers would constitute a transaction.

Thus, we created two `.csv` files — a beer database in the SQL table format and a transactions list like the `*out1.csv` files. This allowed us to turn our code from the last lab towards this data with minimal change.

The data at hand is thusly distributed:

- Users: 97
- Transactions: 922
- Distinct Beers: 9,800
- Checkins: 280,779

With this dataset, it was possible to find frequent itemsets only after lowering the min_sup value tremendously. With min_sup = 0.005, frequent itemsets showed up in around 5 or more transactions. Shown in Figure 1 is a select view of some 1, 2, and 3-itemsets. At this min_sup value, 4-itemsets were found, but cannot fit on the width of this page and are thus omitted. Since there are 922 transactions, a minimum support of 0.005 means that the itemset showed up in around 5 user-score transactions. We will discuss these numbers more in the Analysis section, but while it is a very low support number, it is still interesting and meaningful that an itemset showed up in 5 transactions.

Table 1: Select frequent item sets in the Untappd user-transaction database, min_sup 0.005

size	min_sup 0.005 support	itemset
1	0.0477	[Firestone=Easy Jack]
1	0.0271	[Firestone=Opal]
1	0.0260	[Firestone=Pivo Pils]
1	0.0412	[Dogfish Head Craft=90 Minute IPA]
1	0.0477	[Lagunitas=IPA]
1	0.0314	[Bear Republic=Racer 5 IPA®]
1	0.0336	[21st Amendment=Hell or High Watermelon]
1	0.0488	[Figueroa Mountain=Lizard's Mouth]
1	0.0639	[Firestone=805 Blonde]
1	0.0357	[Stone=Stone IPA]
1	0.0292	[New Belgium=Fat Tire]
1	0.0347	[Arrogant Bastard=Arrogant Bastard Ale]
1	0.0488	[Firestone=DBA (Double Barrel Ale)]
1	0.0368	[Sierra Nevada=Pale Ale]
1	0.0379	[Russian River=Pliny the Elder]
1	0.0314	[Firestone=Union Jack IPA]

Table 1: Select frequent item sets in the Untappd user-transaction database, min_sup 0.005

size	min_sup 0.005 support	itemset
1	0.0260	[Modern Times Beer=Black House]
1	0.0260	[Goose Island Beer Co.=Goose IPA]
1	0.0292	[Ballast Point=Victory at Sea]
1	0.0292	[Deschutes=Fresh Squeezed IPA]
1	0.0260	[Grupo Modelo S.A. de C.V.=Negra Modelo]
1	0.0422	[Belching Beaver=Peanut Butter Milk Stout]
1	0.0455	[Ballast Point=Grapefruit Sculpin]
1	0.0303	[Firestone=Velvet Merlin]
1	0.0271	[Guinness=Guinness Draught]
1	0.0422	[Ballast Point=Sculpin IPA]
1	0.0325	[Boston Beer Company (Samuel Adams)=Samuel Adams OctoberFest]
1	0.0271	[Stone=Stone Cali-Belgique IPA]
1	0.0303	[Sapporo Breweries=Sapporo Premium Beer]
1	0.0271	[21st Amendment=Back in Black]
1	0.0271	[Lagunitas=A Little Sumpin' Sumpin' Ale]
1	0.0271	[Stella Artois=Stella Artois]
1	0.0260	[Big Sky=Moose Drool Brown Ale]
1	0.0357	[Deschutes=Black Butte Porter]
1	0.0260	[Figueroa Mountain=Hoppy Poppy IPA]
1	0.0292	[Firestone=Pale 31]
1	0.0281	[Deschutes=Mirror Pond Pale Ale]
1	0.0314	[Anheuser-Busch=Bud Light]
1	0.0336	[Coors=Coors Light]
1	0.0303	[Firestone=Wookey Jack]
1	0.0260	[Grupo Modelo S.A. de C.V.=Modelo Especial]
1	0.0271	[Shock Top=Belgian White]
1	0.0260	[North Coast=Old Rasputin]
1	0.0260	[BarrelHouse Co=Sunny Daze - Clementine Blonde Ale]
2	0.0054	[North Coast=Old Rasputin]-[Creekside=Double Dark Stout]
2	0.0054	[Boston Beer Company (Samuel Adams)=Samuel Adams Summer Ale]-[Widmer Brothers=Hefeweizen]
2	0.0054	[Lost Coast=Downtown Brown]-[New Belgium=1554]
2	0.0054	[Sierra Nevada=Yonder Bock]-[Sierra Nevada=Myron's Walk]
2	0.0054	[Golden Road=Heal The Bay IPA]-[BarrelHouse Co=Sunny Daze - Clementine Blonde Ale]
3	0.0054	[Firestone=Pivo Pils]-[Firestone=Easy Jack]-[Stella Artois=Stella Artois]
3	0.0054	[Lagunitas=IPA]-[Firestone=Easy Jack]-[New Belgium=Fat Tire]
3	0.0054	[Sierra Nevada=Pale Ale]-[Lagunitas=IPA]-[Firestone=Easy Jack]
3	0.0054	[Lagunitas=IPA]-[Firestone=Union Jack IPA]-[Firestone=Easy Jack]

Figure 2 shows the highest scoring 1-antecedent-1-consequent rules found. These results were discovered using a minimum support of 0.005 and minimum confidence of 0.2. Many more were found, but for brevity only the top 40 are shown.

Table 2: Best scoring (highest support and confidence) association rules for 1-itemset antecedent to 1-itemset consequents

sup	min_sup 0.005 conf	min_conf 0.2 antecedent	consequent
0.012	0.244	[Firestone=DBA (Double Barrel Ale)]	[Sierra Nevada=Pale Ale]
0.012	0.250	[Firestone=Easy Jack]	[Lagunitas=IPA]
0.012	0.250	[Lagunitas=IPA]	[Firestone=Easy Jack]
0.012	0.250	[Lagunitas=IPA]	[Firestone=Velvet Merlin]
0.012	0.250	[Lagunitas=IPA]	[Sierra Nevada=Pale Ale]
0.012	0.262	[Ballast Point=Grapefruit Sculpin]	[Firestone=Double Jack]
0.012	0.282	[Ballast Point=Sculpin IPA]	[Stone=Stone IPA]
0.012	0.314	[Russian River=Pliny the Elder]	[Ballast Point=Victory at Sea]
0.012	0.324	[Sierra Nevada=Pale Ale]	[Arrogant Bastard=Arrogant Bastard Ale]
0.012	0.324	[Sierra Nevada=Pale Ale]	[Firestone=DBA (Double Barrel Ale)]
0.012	0.324	[Sierra Nevada=Pale Ale]	[Lagunitas=IPA]
0.012	0.333	[Stone=Stone IPA]	[Ballast Point=Sculpin IPA]
0.012	0.333	[Stone=Stone IPA]	[Figueroa Mountain=Hoppy Poppy IPA]
0.012	0.344	[Arrogant Bastard=Arrogant Bastard Ale]	[Sierra Nevada=Pale Ale]
0.012	0.355	[Coors=Coors Light]	[Anheuser-Busch=Bud Light]
0.012	0.379	[Anheuser-Busch=Bud Light]	[Coors=Coors Light]
0.012	0.393	[Firestone=Velvet Merlin]	[Lagunitas=IPA]
0.012	0.393	[Sapporo Breweries=Sapporo Premium Beer]	[Firestone=805 Blonde]
0.012	0.407	[Ballast Point=Victory at Sea]	[Russian River=Pliny the Elder]
0.012	0.440	[Stella Artois=Stella Artois]	[Firestone=805 Blonde]
0.012	0.458	[Figueroa Mountain=Hoppy Poppy IPA]	[Stone=Stone IPA]
0.012	0.478	[Firestone=Double Jack]	[Ballast Point=Grapefruit Sculpin]
0.013	0.203	[Firestone=805 Blonde]	[Lagunitas=IPA]
0.013	0.273	[Lagunitas=IPA]	[Firestone=805 Blonde]
0.013	0.273	[Lagunitas=IPA]	[Stone=Stone IPA]
0.013	0.364	[Stone=Stone IPA]	[Firestone=Union Jack IPA]
0.013	0.364	[Stone=Stone IPA]	[Lagunitas=IPA]
0.013	0.414	[Firestone=Union Jack IPA]	[Stone=Stone IPA]
0.013	0.429	[Sapporo Breweries=Sapporo Premium Beer]	[Grupo Modelo S.A. de C.V.=Modelo Especial]
0.013	0.500	[Grupo Modelo S.A. de C.V.=Modelo Especial]	[Sapporo Breweries=Sapporo Premium Beer]
0.014	0.220	[Firestone=805 Blonde]	[Firestone=DBA (Double Barrel Ale)]
0.014	0.220	[Firestone=805 Blonde]	[Firestone=Easy Jack]
0.014	0.289	[Firestone=DBA (Double Barrel Ale)]	[Firestone=805 Blonde]
0.014	0.295	[Firestone=Easy Jack]	[Firestone=805 Blonde]
0.014	0.295	[Firestone=Easy Jack]	[Firestone=Union Jack IPA]
0.014	0.295	[Lagunitas=IPA]	[Bear Republic=Racer 5 IPA®]

Table 2: Best scoring (highest support and confidence) association rules for 1-itemset antecedent to 1-itemset consequents

	min_sup	min_conf	
	0.005	0.2	
sup	conf	antecedent	consequent
0.014	0.310	[Ballast Point=Grapefruit Sculpin]	[Ballast Point=Sculpin IPA]
0.014	0.333	[Ballast Point=Sculpin IPA]	[Ballast Point=Grapefruit Sculpin]
0.014	0.448	[Bear Republic=Racer 5 IPA®]	[Lagunitas=IPA]
0.014	0.448	[Firestone=Union Jack IPA]	[Firestone=Easy Jack]
0.015	0.318	[Lagunitas=IPA]	[Ballast Point=Sculpin IPA]
0.015	0.359	[Ballast Point=Sculpin IPA]	[Lagunitas=IPA]

Figure 3 contains select mult-itemset antecedent-consequent rules. The results were discovered using a minimum support of 0.005 and minimum conidence of 0.2. These are much lower in scoring (support, confidence) overall compared to the 1-itemset rules.

Table 3: Select multi-itemset antecedent and consequents

	min_sup	min_conf	
	0.005	0.2	
sup	conf	antecedent	consequent
0.007	0.600	[Paulaner Gruppe=Paulaner Hefe-Weizen]	[Deschutes=Mirror Pond Pale Ale]-[Stella=Stella Artois]
0.007	0.750	[Firestone=805 Blonde]-[Golden Road=Point The Way IPA]	[Firestone=Easy Jack]
0.007	0.857	[Deschutes=Mirror Pond Pale Ale]-[Firestone=Easy Jack]	[Karl Strauss=Tower 10 IPA]
0.007	0.857	[Deschutes=Mirror Pond Pale Ale]-[Sapporo Breweries=Sapporo Premium Beer]	[Grupo Modelo S.A. de C.V.=Modelo Especial]
0.007	0.857	[Deschutes=Mirror Pond Pale Ale]-[Stella Artois=Stella Artois]	[Paulaner Gruppe=Paulaner Hefe-Weizen]
0.007	0.857	[Firestone=805 Blonde]-[Grupo Modelo S.A. de C.V.=Modelo Especial]	[Sapporo Breweries=Sapporo Premium Beer]
0.007	0.857	[Firestone=805 Blonde]-[Oskar Blues=Mama's Little Yella Pils]	[Firestone=DBA (Double Barrel Ale)]
0.007	0.857	[Firestone=Agrestic]-[Ballast Point=Victory at Sea]	[Ballast Point=Grapefruit Sculpin]
0.007	0.857	[Firestone=Agrestic]-[Firestone=Double Jack]	[Ballast Point=Grapefruit Sculpin]
0.007	0.857	[Firestone=DBA (Double Barrel Ale)]-[Oskar Blues=Mama's Little Yella Pils]	[Firestone=805 Blonde]
0.007	0.857	[Firestone=Pivo Pils]-[Stella Artois=Stella Artois]	[Firestone=805 Blonde]
0.008	0.636	[Sierra Nevada=Pale Ale]-[Arrogant Bastard=Arrogant Bastard Ale]	[Lagunitas=IPA]
0.009	0.242	[Stone=Stone IPA]	[Ballast Point=Sculpin IPA]-[Firestone=Union Jack IPA]

3.2 Archive of Our Own

After de-duplication of the data for fan works that were tagged with multiple shows, there were a total of 19,019 works tagged under "Agent Carter (TV)", "Agents of S.H.I.E.L.D. (TV)", "Daredevil (TV)", "Jessica Jones", and/or "The Defenders". Within those works, there were 1,923 tags that occurred in more than 0.5% of the works that were subsequently used for analysis. Their type breakdown was:

- Relationship tags - 319
- Character tags - 325
- Freeform tags - 1,279

Frequent item sets with a minimum support of 0.05 within this group are shown in Table 4. Given that there are 19,283 distinct fanworks in this data set, essentially that minimum support requires the tag set to occur in at least 964 works. Overall this was entertaining and interesting. The majority of the frequently occurring tags were for characters and relationships. Characters that occurred in the Marvel Cinematic Universe but not the TV shows (such as Clint Barton, Steve Rogers, Natasha Romanov, and Tony Stark) occurred frequently. This analysis also uncovered issues with name consistency within the data set, which influences the frequent item sets artificially lowering the support since the item is falsely split. The character Skye in "Agents of S.H.I.E.L.D. (TV)", for example, had 2 different character tags: "Skye (Agents of SHIELD)" and "Skye | Daisy Johnson". If is possible that these issues incorrectly excluded characters and concepts from association rules mining due to this issue.

Table 4: Frequent item sets derived from the all Marvel TV shows, min_sup 0.05

size	supoort	tag
1	0.12624217887375783	("Leo Fitz/Jemma Simmons")

Table 4: Frequent item sets derived from the all Marvel TV shows, min_sup 0.05

size	supoort	tag
1	0.06688048793311951	("Skye/Grant Ward")
1	0.06288448393711552	("Phil Coulson/Melinda May")
1	0.06845785793154215	("Peggy Carter/Angie Martinelli")
1	0.06766917293233082	("Leo Fitz & Jemma Simmons")
1	0.09343288290656712	("Franklin Foggy Nelson")
1	0.28313791471686206	("Leo Fitz")
1	0.19838056680161945	("Skye (Agents of SHIELD)")
1	0.08002523791997476	("Bobbi Morse")
1	0.23791997476208002	("Grant Ward")
1	0.12356064987643935	("Peggy Carter")
1	0.1306588148693412	("Matt Murdock")
1	0.3608496766391503	("Jemma Simmons")
1	0.18055628581944372	("Skye | Daisy Johnson")
1	0.10053104789946896	("Steve Rogers")
1	0.11251905988748094	("Clint Barton")
1	0.09490509490509491	("Natasha Romanov")
1	0.25074925074925075	("Melinda May")
1	0.05184289394815711	("Karen Page")
1	0.05941426994058573	("James Bucky Barnes")
1	0.07403123192596876	("Tony Stark")
1	0.06961459593038541	("Lance Hunter")
1	0.08023555391976445	("Antoine Triplett")
1	0.051632577948367425	("Alphonso Mack Mackenzie")
1	0.053252421946474576	("Nick Fury")
1	0.07718597192281403	("Angie Martinelli")
1	0.3431831326568169	("Phil Coulson")
1	0.11288711288711288	("Fluff")
1	0.0886481939113518	("Angst")
2	0.11825017088174983	("Leo Fitz")-("Leo Fitz/Jemma Simmons")
2	0.11898627688101372	("Jemma Simmons")-("Leo Fitz/Jemma Simmons")
2	0.05773174194226826	("Skye (Agents of SHIELD)")-("Skye/Grant Ward")
2	0.06440927493559072	("Grant Ward")-("Skye/Grant Ward")
2	0.06099163993900836	("Melinda May")-("Phil Coulson/Melinda May")
2	0.05978232294021768	("Phil Coulson/Melinda May")-("Phil Coulson")
2	0.06477732793522267	("Peggy Carter/Angie Martinelli")-("Peggy Carter")
2	0.0646195909353804	("Peggy Carter/Angie Martinelli")-("Angie Martinelli")
2	0.0643566959356433	("Leo Fitz")-("Leo Fitz & Jemma Simmons")
2	0.06519795993480204	("Leo Fitz & Jemma Simmons")-("Jemma Simmons")
2	0.09054103790945896	("Franklin Foggy Nelson")-("Matt Murdock")
2	0.0916451969083548	("Skye (Agents of SHIELD)")-("Leo Fitz")
2	0.11499027288500972	("Grant Ward")-("Leo Fitz")
2	0.25548136074451866	("Leo Fitz")-("Jemma Simmons")
2	0.0781323939218676	("Skye | Daisy Johnson")-("Leo Fitz")
2	0.12708344287291656	("Melinda May")-("Leo Fitz")
2	0.057521425942478575	("Antoine Triplett")-("Leo Fitz")
2	0.14238393185761608	("Leo Fitz")-("Phil Coulson")
2	0.10847047689152953	("Skye (Agents of SHIELD)")-("Grant Ward")
2	0.1179346968820653	("Skye (Agents of SHIELD)")-("Jemma Simmons")
2	0.09085651190914348	("Melinda May")-("Skye (Agents of SHIELD)")
2	0.11309742888690257	("Skye (Agents of SHIELD)")-("Phil Coulson")
2	0.05731110994268889	("Bobbi Morse")-("Jemma Simmons")
2	0.13617960986382038	("Grant Ward")-("Jemma Simmons")
2	0.06362058993637941	("Skye | Daisy Johnson")-("Grant Ward")
2	0.10163520689836479	("Melinda May")-("Grant Ward")
2	0.11588411588411589	("Grant Ward")-("Phil Coulson")
2	0.07529312792470687	("Peggy Carter")-("Angie Martinelli")
2	0.09758662390241338	("Skye | Daisy Johnson")-("Jemma Simmons")
2	0.1421736158578264	("Melinda May")-("Jemma Simmons")
2	0.06351543193648457	("Antoine Triplett")-("Jemma Simmons")
2	0.15652768284347232	("Phil Coulson")-("Jemma Simmons")
2	0.0703507019296493	("Melinda May")-("Skye | Daisy Johnson")
2	0.09842788790157211	("Skye | Daisy Johnson")-("Phil Coulson")
2	0.052421262947578734	("Steve Rogers")-("Clint Barton")
2	0.053946053946053944	("Steve Rogers")-("Natasha Romanov")
2	0.053472842946527155	("Steve Rogers")-("Tony Stark")
2	0.05094905094905095	("Steve Rogers")-("Phil Coulson")
2	0.06614438193385562	("Clint Barton")-("Natasha Romanov")
2	0.05126452494873548	("Clint Barton")-("Tony Stark")
2	0.08570376991429624	("Clint Barton")-("Phil Coulson")
2	0.05873074294126926	("Phil Coulson")-("Natasha Romanov")
2	0.053052210946947786	("Melinda May")-("Antoine Triplett")
2	0.20553131079446868	("Melinda May")-("Phil Coulson")
2	0.05610179294389821	("Antoine Triplett")-("Phil Coulson")
3	0.11651506388348494	("Leo Fitz")-("Jemma Simmons")-("Leo Fitz/Jemma Simmons")
3	0.0568378989431621	("Skye (Agents of SHIELD)")-("Grant Ward")-("Skye/Grant Ward")
3	0.05930911194069089	("Melinda May")-("Phil Coulson/Melinda May")-("Phil Coulson")
3	0.06414637993585362	("Peggy Carter/Angie Martinelli")-("Peggy Carter")-("Angie Martinelli")
3	0.06372574793627425	("Leo Fitz")-("Leo Fitz & Jemma Simmons")-("Jemma Simmons")
3	0.061307113938692884	("Skye (Agents of SHIELD)")-("Grant Ward")-("Leo Fitz")
3	0.08728113991271887	("Skye (Agents of SHIELD)")-("Leo Fitz")-("Jemma Simmons")
3	0.06304222093695779	("Melinda May")-("Skye (Agents of SHIELD)")-("Leo Fitz")
3	0.06656501393343499	("Skye (Agents of SHIELD)")-("Leo Fitz")-("Phil Coulson")
3	0.10631473789368526	("Grant Ward")-("Leo Fitz")-("Jemma Simmons")
3	0.07902623692097377	("Melinda May")-("Grant Ward")-("Leo Fitz")
3	0.08523055891476944	("Grant Ward")-("Leo Fitz")-("Phil Coulson")
3	0.07129712392870288	("Skye | Daisy Johnson")-("Leo Fitz")-("Jemma Simmons")
3	0.1209316998790683	("Melinda May")-("Leo Fitz")-("Jemma Simmons")
3	0.05536568694463431	("Antoine Triplett")-("Leo Fitz")-("Jemma Simmons")
3	0.1324465008675535	("Leo Fitz")-("Phil Coulson")-("Jemma Simmons")
3	0.05184289394815711	("Skye | Daisy Johnson")-("Leo Fitz")-("Phil Coulson")

Table 4: Frequent item sets derived from the all Marvel TV shows, min_sup 0.05

size	supoort	tag
3	0.11446448288553551	("Melinda May")-("Leo Fitz")-("Phil Coulson")
3	0.06561859193438141	("Skye (Agents of SHIELD)")-("Grant Ward")-("Jemma Simmons")
3	0.05636468794363531	("Melinda May")-("Skye (Agents of SHIELD)")-("Grant Ward")
3	0.06246385193753615	("Skye (Agents of SHIELD)")-("Grant Ward")-("Phil Coulson")
3	0.0693517009306483	("Melinda May")-("Skye (Agents of SHIELD)")-("Jemma Simmons")
3	0.07271675692728324	("Skye (Agents of SHIELD)")-("Phil Coulson")-("Jemma Simmons")
3	0.08034071191965929	("Melinda May")-("Skye (Agents of SHIELD)")-("Phil Coulson")
3	0.08155002891844997	("Melinda May")-("Grant Ward")-("Jemma Simmons")
3	0.08749145591250855	("Grant Ward")-("Phil Coulson")-("Jemma Simmons")
3	0.09080393290919607	("Melinda May")-("Grant Ward")-("Phil Coulson")
3	0.05094905094905095	("Melinda May")-("Skye | Daisy Johnson")-("Jemma Simmons")
3	0.05610179294389821	("Skye | Daisy Johnson")-("Phil Coulson")-("Jemma Simmons")
3	0.12350807087649193	("Melinda May")-("Phil Coulson")-("Jemma Simmons")
3	0.06004521793995478	("Melinda May")-("Skye | Daisy Johnson")-("Phil Coulson")
4	0.05962458594037542	("Skye (Agents of SHIELD)")-("Leo Fitz")-("Jemma Simmons")-("Grant Ward")
4	0.06146485093853515	("Skye (Agents of SHIELD)")-("Leo Fitz")-("Jemma Simmons")-("Melinda May")
4	0.06414637993585362	("Skye (Agents of SHIELD)")-("Leo Fitz")-("Jemma Simmons")-("Phil Coulson")
4	0.05941426994058573	("Skye (Agents of SHIELD)")-("Leo Fitz")-("Melinda May")-("Phil Coulson")
4	0.0766076029233924	("Leo Fitz")-("Jemma Simmons")-("Melinda May")-("Grant Ward")
4	0.08149744991850255	("Leo Fitz")-("Jemma Simmons")-("Grant Ward")-("Phil Coulson")
4	0.07471475892528524	("Leo Fitz")-("Melinda May")-("Grant Ward")-("Phil Coulson")
4	0.10957463589042536	("Leo Fitz")-("Jemma Simmons")-("Melinda May")-("Phil Coulson")
4	0.05000262894999737	("Skye (Agents of SHIELD)")-("Jemma Simmons")-("Grant Ward")-("Phil Coulson")
4	0.05236868394763131	("Skye (Agents of SHIELD)")-("Melinda May")-("Grant Ward")-("Phil Coulson")
4	0.063568010936432	("Skye (Agents of SHIELD)")-("Jemma Simmons")-("Melinda May")-("Phil Coulson")
4	0.0761343919238656	("Jemma Simmons")-("Melinda May")-("Grant Ward")-("Phil Coulson")
5	0.057942057942057944	("Skye (Agents of SHIELD)")-("Leo Fitz")-("Jemma Simmons")-("Melinda May")-("Phil Coulson")
5	0.07266417792733582	("Leo Fitz")-("Jemma Simmons")-("Grant Ward")-("Melinda May")-("Phil Coulson")

The association rules found with minimum support of 0.05 and a minimum confidence of 0.80 are shown in Table 5. No association rules involving freeform tags were found.

Table 5: Association Rules derived from the Marvel TV shows, min_sup 0.05 min_conf 0.80

sup	conf	antecedent	consequent
0.064	0.993	("Peggy Carter/Angie Martinelli")-("Angie Martinelli")	("Peggy Carter")
0.059	0.992	("Phil Coulson/Melinda May")-("Phil Coulson")	("Melinda May")
0.064	0.990	("Leo Fitz")-("Leo Fitz & Jemma Simmons")	("Jemma Simmons")
0.064	0.990	("Peggy Carter/Angie Martinelli")-("Peggy Carter")	("Angie Martinelli")
0.117	0.985	("Leo Fitz")-("Leo Fitz/Jemma Simmons")	("Jemma Simmons")
0.057	0.985	("Skye (Agents of SHIELD)")-("Skye/Grant Ward")	("Grant Ward")
0.117	0.979	("Jemma Simmons")-("Leo Fitz/Jemma Simmons")	("Leo Fitz")
0.064	0.977	("Leo Fitz & Jemma Simmons")-("Jemma Simmons")	("Leo Fitz")
0.075	0.975	("Angie Martinelli")	("Peggy Carter")
0.061	0.975	("Melinda May")-("Skye (Agents of SHIELD)")-("Leo Fitz")	("Jemma Simmons")
0.058	0.975	("Melinda May")-("Skye (Agents of SHIELD)")-("Leo Fitz")-("Phil Coulson")	("Jemma Simmons")
0.073	0.973	("Melinda May")-("Grant Ward")-("Leo Fitz")-("Phil Coulson")	("Jemma Simmons")
0.060	0.973	("Skye (Agents of SHIELD)")-("Grant Ward")-("Leo Fitz")	("Jemma Simmons")
0.059	0.972	("Melinda May")-("Phil Coulson/Melinda May")	("Phil Coulson")
0.061	0.970	("Phil Coulson/Melinda May")	("Melinda May")
0.091	0.969	("Franklin Foggy Nelson")	("Matt Murdock")
0.077	0.969	("Melinda May")-("Grant Ward")-("Leo Fitz")	("Jemma Simmons")
0.064	0.964	("Skye (Agents of SHIELD)")-("Leo Fitz")-("Phil Coulson")	("Jemma Simmons")
0.055	0.963	("Antoine Triplett")-("Leo Fitz")	("Jemma Simmons")
0.065	0.963	("Leo Fitz & Jemma Simmons")	("Jemma Simmons")
0.064	0.963	("Skye/Grant Ward")	("Grant Ward")
0.110	0.957	("Melinda May")-("Leo Fitz")-("Phil Coulson")	("Jemma Simmons")
0.081	0.956	("Grant Ward")-("Leo Fitz")-("Phil Coulson")	("Jemma Simmons")
0.073	0.954	("Melinda May")-("Grant Ward")-("Phil Coulson")-("Jemma Simmons")	("Leo Fitz")
0.121	0.952	("Melinda May")-("Leo Fitz")	("Jemma Simmons")
0.087	0.952	("Skye (Agents of SHIELD)")-("Leo Fitz")	("Jemma Simmons")
0.064	0.951	("Leo Fitz & Jemma Simmons")	("Leo Fitz")
0.060	0.951	("Phil Coulson/Melinda May")	("Phil Coulson")
0.073	0.949	("Melinda May")-("Grant Ward")-("Leo Fitz")-("Jemma Simmons")	("Phil Coulson")
0.065	0.946	("Peggy Carter/Angie Martinelli")	("Peggy Carter")
0.075	0.945	("Melinda May")-("Grant Ward")-("Leo Fitz")	("Phil Coulson")
0.065	0.944	("Peggy Carter/Angie Martinelli")	("Angie Martinelli")
0.119	0.943	("Leo Fitz/Jemma Simmons")	("Jemma Simmons")
0.058	0.943	("Melinda May")-("Skye (Agents of SHIELD)")-("Leo Fitz")-("Jemma Simmons")	("Phil Coulson")
0.059	0.943	("Phil Coulson/Melinda May")	("Melinda May")-("Phil Coulson")
0.064	0.942	("Leo Fitz & Jemma Simmons")	("Leo Fitz")-("Jemma Simmons")
0.059	0.942	("Melinda May")-("Skye (Agents of SHIELD)")-("Leo Fitz")	("Phil Coulson")
0.077	0.939	("Melinda May")-("Grant Ward")-("Jemma Simmons")	("Leo Fitz")
0.118	0.937	("Leo Fitz/Jemma Simmons")	("Leo Fitz")
0.064	0.937	("Peggy Carter/Angie Martinelli")	("Peggy Carter")-("Angie Martinelli")
0.076	0.934	("Melinda May")-("Grant Ward")-("Jemma Simmons")	("Phil Coulson")
0.081	0.931	("Grant Ward")-("Phil Coulson")-("Jemma Simmons")	("Leo Fitz")
0.132	0.930	("Leo Fitz")-("Phil Coulson")	("Jemma Simmons")
0.052	0.929	("Melinda May")-("Skye (Agents of SHIELD)")-("Grant Ward")	("Phil Coulson")
0.106	0.925	("Grant Ward")-("Leo Fitz")	("Jemma Simmons")
0.117	0.923	("Leo Fitz/Jemma Simmons")	("Leo Fitz")-("Jemma Simmons")
0.073	0.919	("Melinda May")-("Grant Ward")-("Leo Fitz")	("Phil Coulson")-("Jemma Simmons")
0.058	0.919	("Melinda May")-("Skye (Agents of SHIELD)")-("Leo Fitz")	("Phil Coulson")-("Jemma Simmons")
0.064	0.917	("Melinda May")-("Skye (Agents of SHIELD)")-("Jemma Simmons")	("Phil Coulson")
0.071	0.913	("Skye | Daisy Johnson")-("Leo Fitz")	("Jemma Simmons")
0.058	0.911	("Melinda May")-("Skye (Agents of SHIELD)")-("Phil Coulson")-("Jemma Simmons")	("Leo Fitz")

Table 5: Association Rules derived from the Marvel TV shows, min_sup 0.05 min_conf 0.80

sup	conf	antecedent	consequent
0.060	0.909	("Skye (Agents of SHIELD")-("Grant Ward")-("Jemma Simmons")	("Leo Fitz")
0.110	0.906	("Melinda May")-("Leo Fitz")-("Jemma Simmons")	("Phil Coulson")
0.058	0.903	("Skye (Agents of SHIELD")-("Leo Fitz")-("Phil Coulson")-("Jemma Simmons")	("Melinda May")
0.255	0.902	("Leo Fitz")	("Jemma Simmons")
0.114	0.901	("Melinda May")-("Leo Fitz")	("Phil Coulson")
0.091	0.893	("Melinda May")-("Grant Ward")	("Phil Coulson")
0.059	0.893	("Skye (Agents of SHIELD")-("Leo Fitz")-("Phil Coulson")	("Melinda May")
0.073	0.892	("Grant Ward")-("Leo Fitz")-("Phil Coulson")-("Jemma Simmons")	("Melinda May")
0.073	0.891	("Melinda May")-("Grant Ward")-("Jemma Simmons")	("Leo Fitz")-("Phil Coulson")
0.110	0.887	("Melinda May")-("Phil Coulson")-("Jemma Simmons")	("Leo Fitz")
0.061	0.886	("Melinda May")-("Skye (Agents of SHIELD")-("Jemma Simmons")	("Leo Fitz")
0.080	0.884	("Melinda May")-("Skye (Agents of SHIELD")	("Phil Coulson")
0.057	0.882	("Grant Ward")-("Skye/Grant Ward")	("Skye (Agents of SHIELD)")
0.064	0.882	("Skye (Agents of SHIELD")-("Phil Coulson")-("Jemma Simmons")	("Leo Fitz")
0.075	0.877	("Grant Ward")-("Leo Fitz")-("Phil Coulson")	("Melinda May")
0.064	0.874	("Skye (Agents of SHIELD")-("Phil Coulson")-("Jemma Simmons")	("Melinda May")
0.055	0.872	("Antoine Triplett")-("Jemma Simmons")	("Leo Fitz")
0.076	0.870	("Grant Ward")-("Phil Coulson")-("Jemma Simmons")	("Melinda May")
0.058	0.870	("Skye (Agents of SHIELD")-("Leo Fitz")-("Phil Coulson")	("Melinda May")-("Jemma Simmons")
0.124	0.869	("Melinda May")-("Jemma Simmons")	("Phil Coulson")
0.058	0.863	("Skye/Grant Ward")	("Skye (Agents of SHIELD)")
0.110	0.862	("Melinda May")-("Leo Fitz")	("Phil Coulson")-("Jemma Simmons")
0.060	0.854	("Melinda May")-("Skye | Daisy Johnson")	("Phil Coulson")
0.073	0.853	("Grant Ward")-("Leo Fitz")-("Phil Coulson")	("Melinda May")-("Jemma Simmons")
0.064	0.852	("Peggy Carter")-("Angie Martinelli")	("Peggy Carter/Angie Martinelli")
0.121	0.851	("Melinda May")-("Jemma Simmons")	("Leo Fitz")
0.057	0.850	("Skye/Grant Ward")	("Skye (Agents of SHIELD)")-("Grant Ward")
0.132	0.846	("Phil Coulson")-("Jemma Simmons")	("Leo Fitz")
0.076	0.838	("Melinda May")-("Grant Ward")-("Phil Coulson")	("Jemma Simmons")
0.052	0.838	("Skye (Agents of SHIELD")-("Grant Ward")-("Phil Coulson")	("Melinda May")
0.065	0.837	("Angie Martinelli")	("Peggy Carter/Angie Martinelli")
0.058	0.835	("Melinda May")-("Skye (Agents of SHIELD")-("Jemma Simmons")	("Leo Fitz")-("Phil Coulson")
0.064	0.831	("Angie Martinelli")	("Peggy Carter/Angie Martinelli")-("Peggy Carter")
0.073	0.831	("Grant Ward")-("Phil Coulson")-("Jemma Simmons")	("Melinda May")-("Leo Fitz")
0.110	0.827	("Leo Fitz")-("Phil Coulson")-("Jemma Simmons")	("Melinda May")
0.075	0.823	("Melinda May")-("Grant Ward")-("Phil Coulson")	("Leo Fitz")
0.206	0.820	("Melinda May")	("Phil Coulson")
0.114	0.804	("Leo Fitz")-("Phil Coulson")	("Melinda May")
0.082	0.802	("Melinda May")-("Grant Ward")	("Jemma Simmons")
0.050	0.801	("Skye (Agents of SHIELD")-("Grant Ward")-("Phil Coulson")	("Jemma Simmons")
0.073	0.800	("Melinda May")-("Grant Ward")-("Phil Coulson")	("Leo Fitz")-("Jemma Simmons")

4 Analysis

4.1 Untapped

4.1.0.1 Data Size Data size seemed to be an important factor in interpreting the association rule mining results. In the dataset are 922 transactions, but they're derived from 97 users. Thinking back to the question at hand — What beers do users consider equivalent? — we have to understand what such low support numbers mean for many of the results.

When we look at support values of 0.005 or above, we're looking at itemsets that are present in about 5 of the transactions. Since these transactions are made up of (user,rating) pairs, this means that 5 users have the beers in the itemset rated the same[1]. That is something we consider important, especially considering that the 922 transactions are made from 97 users. In other words, 10% of users consider the beers in the itemset to be equivalent.

Further issues in the sort of unbalanced data size collected comes in when considering that we have 9,800 beers and 280,779 checkins. This means that we have a lot of places that a beer can show up in many many transactions, skewing confidence as well. For our purposes, we were okay with low confidence values, especially because most associations made sense.

4.1.0.2 Associations Associations discovered made a lot of sense. In Table 2, we see that the strongest single itemset association rules are Sculpin IPA → Lagunatas IPA and Lagunatas IPA → Sculpin IPA, followed closely by Firestone's Union Jack IPA → Firestone's Easy Jack and many other IPA → Other IPA associations.

Such IPA associations show up even in the multi-itemset associations, such as Stone IPA → {Ballast Point Sculpin IPA,Firestone's Union Jack IPA} and {Sierra Nevada Pale Ale, Arrogant Bastard Ale} → {Lagunitas=IPA}. It's important to now that even though Sierra Nevada Pale Ale is called a Pale Ale, it is particularly hoppy, similar to Arrogant Bastard and Lagunitas IPA.

Moving onto blondes, Firestone's 805 Blonde is particularly frequent in the dataset, most likely because the feed comes from users who checked in beers within 25 miles of Spike's Pub in San Luis Obispo. From personal experience, we know that 805 is somewhat pervasive in the Central Coast — more so than even in other places in California. Some flavor-related associations include {Firestone's Pivo Pils, Stella Artois} → {Firestone's 805 Blonde}, where Pivo Pils is a pilsner and Stella Artois is a European Pale Lager. These beers are smooth, light, and easy to drink, making it a highly relevant association to discover.

One particular Pale Ale came up with some interesting associations: Deschutes Mirror Pond Pale Ale. Mirror Pond is a lighter beer with a subtle hops profile. IPA fans find it enjoyable, but for palates that balk at the stronger IPA bitterness, Mirror Pond offers a fresh reprieve from the palate-wrecking hops.That being said, three interesting associations appeared:

- {Deschutes Mirror Pond Pale Ale,Firestone Easy Jack} → {Karl Strauss Tower 10 IPA}
- {Deschutes Mirror Pond Pale Ale,Sapporo Premium Beer} → {Modelo Especial}
- {Deschutes Mirror Pond Pale Ale,Stella Artois} → {Paulaner Gruppe Paulaner Hefe-Weizen}

These three associations are wonderful. Consider each 2-itemset antecedent and notice that we have Mirror Pond — a beer that's halfway between a pale ale and an IPA — and another beer. Consider the itemset with Mirror Pond and Firestone Easy Jack — an IPA which is, of course, hoppy — and see that the association is Karl Strauss's Tower 10 IPA — yet another hoppy IPA.

[1] assuming users rated the beer with one score — which is a reasonable assumption to make with this data

Now, inspect Mirror Pond and Sapporo — a Japanese Rice Lager — and see that the association is with Modelo Especial — an American Adjunct Lager. Furthermore, the Mirror Pond and Stella Artois — a European Pale Lager — and see that it associates with Paulaner Hefeweizen — a style of German Wheat Beer. As an avid beer enthusiast, I find that these associations are absolutely on point.

4.2 Archive of Our Own

Overall the association rules found in the data set skewed to character and relationship tags. The disproportionate number of Agents of S.H.I.E.L.D. fanworks (13,757 compared with the next highest count of 2,895 for Daredevil) also skewed the data to characters commonly found in those works. Crossovers with Marvel Cinematic Universe characters were popular given that frequent items included the character tags for Clint Barton, Steve Rogers, Natasha Romanov, and Tony Stark, but none of the association rules at our threshold included these crossover characters. Only 2 freeform tags were frequent items sets when the minimum support was 0.05: "Fluff" at 0.11289 support and "Angst" at 0.08865 support. Neither of these freefrom tags resulted in association rules at 0.80 confidence.

It was very interesting to see that none of relationship tags had 100% confidence when there was an association rule between the relationship tag and a component character tag. The highest confidence for a antecedent of a single relationship tag occurred with the "Phil Coulson/Melinda May" tag at 0.970 confidence. The lowest given the minimum confidence of 0.80 was Skye/Grant Ward at 0.850, though that case did have a consequent of both relevant character tags. Presumably if a relationship is relevant enough to the work to be included in the tags the characters in the relationship would also be relevant, but this did not hold true 100% of the time.

There are also issues with name consistency within the data set, which influences the frequent item sets artificially lowering the support since the item is falsely split. The character Skye in "Agents of S.H.I.E.L.D. (TV)", for example, had 2 different character tags: "Skye (Agents of SHIELD)" and "Skye | Daisy Johnson". Both were frequent items, but this is more of a lucky happenstance since she is the protagonist of Agents of S.H.I.E.L.D. (TV). In contrast, tags involving Patricia "Trish" Walker in Jessica Jones ran into this issue significantly more:

- Platonic relationship tags:
  - Jessica Jones & Patricia Walker - 11 occurrences
- Romantic relationship tags:
  - Jessica Jones/Patricia Trish Walker - 24 occurrences
  - Jessica Jones/Trish Walker - 17 occurrences
  - Jessica Jones/Patricia Walker - 43 occurrences
- Character tags:
  - Trish Walker - 25 occurrences
  - Patricia Walker - 57 occurrences
  - Patricia Trish Walker - 26 occurrences

If these duplicated tags did not co-occur on works, it is possible that Trish's character tag actually occurred 108 times, which almost twice as often as her most used character tag. It is also very likely that this occurred with a variety of "Alternate Universe" tags used on AO3.

5 Conclusions

The Untappd dataset was such a small subset of the Untappd universe. Not only was the collection limited, but methods that may have achieved a more representative sample were not employed. This created a very strangley unbalanced dataset that required possibly unorthodox minimum support and minimum confidence values to get meaningful association rules.

That being said, the associations discovered made sense overall. IPAs linked with IPAs and multi-itemset associations gve some interesting insight into the beers people enjoy at similar levels. Not covered in this analysis were stout/porter/dark-flavored beer associations, associations between rare beers from different breweries[2], and regional associations[3].

One inherent limitation of association rule mining in this sense is that we have no way of comparing ratings to each other, which is still useful. Right now, we cannot ask what beers users are rating similarly, just equivalently. Clustering beers by ratings using a similarity metric like Pearson Correlation can account for the variation in users' score can possibly produce meaningful clusters towards a similar end to the one in this project. In order to do such an analysis, either beers with ratings by nearly every user involved are necessary, or some notion of only measuring the distance using dimensions that have been scored by every user involved must be formulated.

All the tag data from Archive of Our Own (AO3) could have benefited immensely from master data management techniques to de-duplicate the redundant tags. It is likely association rules missed in this analysis due to this data issue. It may be worth the time for AO3 to analyze the works that include a relationship tag but do not include all component character tags to see if an automated script to add those character tags would be relevant or appropriate. Overall learning how to scrape web pages with a python script was an interesting and I was surprised to find that none of the freeform tags produced association rules.

[2] that may get rated similarly high because of their rarity

[3] the associations between Firestone beers were pervasive throughout the dataset

Appendix A

A.1 Untappd

Listing 1: Example Untappd API JSON Output


```
{
  "meta": {
    "code": 200,
    "response_time": {
      "time": 0,
      "measure": "seconds"
    }
  },
  "notifications": {},
  "response":
   {
   "checkins": {
     "count": 25,
     "items": {
       "checkin_id": 137117722,
       "created_at": "Sat, 13 Dec 2014 19:15:38 +0000",
       "checkin_comment": "When in Rome..",
       "rating_score": 3,
       "user": {
         "uid": 1,
         "user_name": "gregavola",
         "first_name": "Greg",
         "last_name": "Avola",
         "location": "New York, NY",
         "is_supporter": 1,
         "url": "http://gregavola.com",
         "bio": "Co-Founder and CTO of Untappd, Web Developer, Beer Drinker & Community
             Guy",
         "relationship": "self",
         "user_avatar": "https://gravatar.com/avatar/0c6922e238dae5cccce96a32889fc911?size
             =100&d=htt44.cloudfront.net%2Fsite%2Fassets%2Fimages%2Fdefault_avatar_v2.jpg
             %3Fv%3D1",
         "is_private": 0,
         "contact": {
           "foursquare": 195741,
           "twitter": "gregavola",
           "facebook": 18603076
         }
       },
       "beer": {
         "bid": 7481,
         "beer_name": "Brooklyn Bowl Pale Ale",
         "beer_label": "https://d1c8v1qci5en44.cloudfront.net/site/assets/images/temp/
             badge-beer-default.png",
         "beer_style": "American Pale Ale",
         "beer_abv": 0,
         "auth_rating": 0,
         "wish_list": false,
```

```
  "beer_active": 1
},
"brewery": {
  "brewery_id": 1954,
  "brewery_name": "Kelso of Brooklyn",
  "brewery_slug": "kelso-of-brooklyn",
  "brewery_label": "https://d1c8v1qci5en44.cloudfront.net/site/brewery_logos/
      brewery-KelsoofBrooklyn_1954.jpeg",
  "country_name": "United States",
  "contact": {
    "twitter": "KelsoBeer",
    "facebook": "",
    "instagram": "",
    "url": "http://www.kelsoofbrooklyn.com/"
  },
  "location": {
    "brewery_city": "Brooklyn",
    "brewery_state": "NY",
    "lat": 40.6823,
    "lng": -73.9656
  },
  "brewery_active": 1
},
"venue": {
  "venue_id": 2141,
  "venue_name": "Brooklyn Bowl",
  "primary_category": "Arts & Entertainment",
  "parent_category_id": "4d4b7104d754a06370d81259",
  "categories": {
    "count": 3,
    "items": [
      {
        "category_name": "Bowling Alley",
        "category_id": "4bf58dd8d48988d1e4931735",
        "is_primary": true
      },
      {
        "category_name": "Music Venue",
        "category_id": "4bf58dd8d48988d1e5931735",
        "is_primary": false
      },
      {
        "category_name": "Bar",
        "category_id": "4bf58dd8d48988d116941735",
        "is_primary": false
      }
    ]
  },
  "location": {
    "venue_address": "61 Wythe Ave",
    "venue_city": "Brooklyn",
    "venue_state": "NY",
    "venue_country": "United States",
    "lat": 40.7219,
```

```
    "lng": -73.9575
  },
  "contact": {
    "twitter": "@brooklynbowl",
    "venue_url": "http://www.brooklynbowl.com"
  },
  "public_venue": true,
  "foursquare": {
    "foursquare_id": "4a1afeb7f964a520b77a1fe3",
    "foursquare_url": "http://4sq.com/3fjtlA"
  },
  "venue_icon": {
    "sm": "https://ss3.4sqi.net/img/categories_v2/arts_entertainment/bowling_bg_64.
        png",
    "md": "https://ss3.4sqi.net/img/categories_v2/arts_entertainment/bowling_bg_88.
        png",
    "lg": "https://ss3.4sqi.net/img/categories_v2/arts_entertainment/bowling_bg_88.
        png"
  }
},
"comments": {
  "total_count": 0,
  "count": 0,
  "items": []
},
"toasts": {
  "total_count": 0,
  "count": 0,
  "auth_toast": false,
  "items": []
},
"media": {
  "count": 0,
  "items": []
},
"source": {
  "app_name": "Untappd for iPhone - (V2)",
  "app_website": "http://untpd.it/iphoneapp"
},
"badges": {
  "count": 1,
  "items": [
    {
      "badge_id": 189,
      "user_badge_id": 39410316,
      "badge_name": "Taste the Music",
      "badge_description": "Badge Description Here",
      "created_at": "Sat, 13 Dec 2014 19:15:41 +0000",
      "badge_image": {
        "sm": "https://d1c8v1qci5en44.cloudfront.net/badges/bdg_ConcertVenue_sm.jpg
            ",
        "md": "https://d1c8v1qci5en44.cloudfront.net/badges/bdg_ConcertVenue_md.jpg
            ",
        "lg": "https://d1c8v1qci5en44.cloudfront.net/badges/bdg_ConcertVenue_lg.jpg
```

```
                "
            }
          }
        ]
      }
    }
  }
 }
}
```