Movie Recommendations & Relationships Between Greek Gods

CSC 466
Analytics Project
Jorge Guevara
Arkadiy Kraminsky

1. Abstract

Data mining and understanding data has huge application across many domains of interest ranging from sports statistics to movie recommendations/categorizations. Along with many different datasets that can be generated, many algorithms and methods have been developed to extract information. This joint analytics project attempts to analyze and answer questions relating to discovering movie recommendations by using collaborative filtering, and also what Greek Gods have the most relationships with other gods using the PageRank algorithm. This analysis documents how the datasets were obtained and processed, as well as the results of the findings.

2. Introduction

The Greek Gods, or Deities, played an important religious and political role in the day to day lives of Greeks. So much emphasis was put on the Gods and their domains that there are 370 gods in the Greek Pagan religion. A good question to ask of the list of Gods, is which God has the most relations with other Gods? This question can be easily answered by using the PageRank algorithm, which was luckily implemented for lab 3 of the Knowledge in Discovery of Data class. The database used for the PageRank algorithm was generated by web scraping the data from an online website on the Greek Gods.

Movie recommendations is not just a fun academic exercise, but a multi-billion dollar business. Arguably, the biggest player is Netflix. Netflix is essentially a subscription-based business; a customer can cancel their subscription at any time. If enough people cancel their subscription, Netflix will go out of business (or, as is fashionable within the Silicon Valley culture, Netflix will pivot into some other business). So how does Netflix keep its customer base? By recommending movies to watch. Now that the secret is out, we can all go and build a Netflix clone and become filthy rich. If you are the intrepid reader that we think you are, please continue reading to discover the secrets of recommendation engines.

3. Datasets

3.1 Greek Gods Dataset

The Greek Gods dataset was acquired by web scraping a website[1] containing a database of Greek Gods. Each Greek God has it's own web page with a box of quick facts about the God including the God's' mother, father, and any consorts they had. The web scraper was written in python, and uses the lxml and requests libraries to connect to the database of Greek Gods and parse the html contents. As each God's web page is scraped, a bidirectional relationship is created between the Gods' father, mother, or consort if they have any. The relationship is then written to a csv filed very much like the datasets used for the PageRank algorithm from the PageRank analysis lab. The

PageRank algorithm developed from the lab is then applied to the greek gods csv and creates a list of items with the highest and lowest prestige.

3.2 Movie Recommendations Dataset

Having ran out of land to discover, the modern age conquistador must turn to the wilderness of the information age; raw, unadulterated data. The data that we used for the movie recommendation comes from a website called movielens.org, which provides non-commercial, personalized movie recommendations. Ideally, the user of movielens rates movies and the website recommends new movies to the user. We couldn't find a better dataset. The dataset consists of 100,000 ratings, 943 users and 1682 movies. Each user has rated at least 20 movies. We also have access to simple demographic info for the user such as age, gender, occupation and zip. The dataset was collected from September 19th, 1997 to April 22nd, 1998. The dataset is complete in the sense that only users that had demographic information and at least 20 movie ratings are included. The file named u.data contains the most important information for our investigation. Each line in u.data contains the user id, movie id, rating and timestamp, separated by space characters. The movielens dataset that we used was the smallest dataset available. The other datasets contain 1 million, 10 million and 20 million ratings. While playing with the bigger datasets would have been cool, we opted for the smaller dataset so that we could concentrate on implementing and analyzing collaborative filtering.

4. Questions

4.1 Greek Gods Dataset

There were two questions of interest the PageRank algorithm could answer about the Greek Gods dataset. The first question was which Gods are more related to the other Gods. The second question about the dataset was if there is a way to determine which Gods were born earlier and which ones were born later.

4.2 Movie Recommendations Dataset

For the movielens dataset, the main question we had was, how good are the predictions given by a collaborative filtering algorithm. Specifically, we were interested in the accuracy of the k-nearest neighbor algorithm.

5. Methods Deployed

4.1 Greek Gods Dataset

In addition to the python script developed to scrape the dataset from the online database, the method deployed for the Greek Gods database was the PageRank algorithm. The PageRank algorithm was ideal in producing a positive result to both questions asked about the Greek Gods.

4.2 Movie Recommendations Dataset

We rolled our own implementation of the k-nearest neighbor algorithm to test it on the movielens dataset. We opted to do all calculations in-memory. Since we used the small dataset, this decision was sound. We decided to do an item-based implementation. While we could have precomputed the entire similarity matrix, we find the k nearest neighbors for the item of interest at run time. We could have done a user-based implementation as well, since we don't add any of our ratings to the dataset and we don't have a need to do live analysis, but we wanted to experiment with the item-based approach. For the similarity formula, we made use of the adjusted cosine similarity. Some of the research papers we found on k-nearest neighbor suggested that a good value for k was $\sqrt{number\ of\ features}$. We settled on k = 20.

6. Results

6.1 Greek Gods Dataset

Applying the PageRank algorithm on the Greek Gods dataset revealed some interesting features about the database. Nyx, the Goddess of Darkness, resulted as having the highest PageRank prestige followed by Gaia, Uranus, and Erebus. Zeus, surprisingly, was not to be found anywhere in the top 20. A quick inspection of the website revealed that even Gods who are well known have incomplete data about their relationships with other Gods inside their facts box. Interested with the result, we used a handmade bidirectional csv file from a family tree found online[2] to obtain the PageRanks from a more accurate dataset. Indeed the results were different, with Zeus coming out on top and Nyx coming in second.

PageRank Web Scraped Data- Ordered by Highest Prestige:
Read time: 0.007716
Processing time: 0.009249
Iterations until convergence: 30

pageRanks:
1 obj: NYX with PageRank: 0.043460 indegree: 45 outdegree: 45 total: 90
2 obj: GAIA with PageRank: 0.041973 indegree: 54 outdegree: 54 total: 108
3 obj: URANUS with PageRank: 0.037378 indegree: 48 outdegree: 48 total: 96
4 obj: EREBUS with PageRank: 0.034289 indegree: 37 outdegree: 37 total: 74
5 obj: ATLAS with PageRank: 0.024149 indegree: 20 outdegree: 20 total: 40
6 obj: OCEANUS with PageRank: 0.022447 indegree: 26 outdegree: 26 total: 52
7 obj: TETHYS with PageRank: 0.021159 indegree: 24 outdegree: 24 total: 48
8 obj: CRONUS with PageRank: 0.020933 indegree: 28 outdegree: 28 total: 56
9 obj: PLEIONE with PageRank: 0.019930 indegree: 19 outdegree: 19 total: 38
10 obj: HELIOS with PageRank: 0.019218 indegree: 18 outdegree: 18 total: 36
11 obj: RHEA with PageRank: 0.018360 indegree: 24 outdegree: 24 total: 48
12 obj: ZEPHYRUS with PageRank: 0.017643 indegree: 16 outdegree: 16 total: 32

13 obj: ELECTRA with PageRank: 0.017103 indegree: 19 outdegree: 19 total: 38
14 obj: THAUMAS with PageRank: 0.016739 indegree: 19 outdegree: 19 total: 38
15 obj: ASTRAEUS with PageRank: 0.014647 indegree: 15 outdegree: 15 total: 30
16 obj: CLYMENE with PageRank: 0.013461 indegree: 9 outdegree: 9 total: 18
17 obj: IAPETUS with PageRank: 0.013312 indegree: 11 outdegree: 11 total: 22
18 obj: EOS with PageRank: 0.013120 indegree: 13 outdegree: 13 total: 26
19 obj: PALLAS with PageRank: 0.012975 indegree: 10 outdegree: 10 total: 20
20 obj: PHOEBE with PageRank: 0.012739 indegree: 14 outdegree: 14 total: 28
21 obj: COEUS with PageRank: 0.012739 indegree: 14 outdegree: 14 total: 28
22 obj: BOREAS with PageRank: 0.012528 indegree: 14 outdegree: 14 total: 28
23 obj: HYPNOS with PageRank: 0.012143 indegree: 8 outdegree: 8 total: 16
24 obj: PERSE with PageRank: 0.011656 indegree: 12 outdegree: 12 total: 24
25 obj: ASTERIA with PageRank: 0.011589 indegree: 12 outdegree: 12 total: 24
26 obj: STYX with PageRank: 0.011273 indegree: 8 outdegree: 8 total: 16
27 obj: OREITHYIA with PageRank: 0.010889 indegree: 12 outdegree: 12 total: 24
28 obj: THEA with PageRank: 0.010663 indegree: 11 outdegree: 11 total: 22
29 obj: ZEUS with PageRank: 0.010577 indegree: 12 outdegree: 12 total: 24
30 obj: NEAERA with PageRank: 0.010481 indegree: 10 outdegree: 10 total: 20
31 obj: PASITHEA with PageRank: 0.010422 indegree: 6 outdegree: 6 total: 12
32 obj: POSEIDON with PageRank: 0.010244 indegree: 10 outdegree: 10 total: 20
33 obj: CRIUS with PageRank: 0.010118 indegree: 11 outdegree: 11 total: 22
34 obj: ALCYONE with PageRank: 0.009692 indegree: 9 outdegree: 9 total: 18
35 obj: PHOBETOR with PageRank: 0.009630 indegree: 6 outdegree: 6 total: 12
36 obj: HYPERION with PageRank: 0.009272 indegree: 9 outdegree: 9 total: 18
37 obj: PENIA with PageRank: 0.009266 indegree: 8 outdegree: 8 total: 16
38 obj: BOREADS with PageRank: 0.009027 indegree: 10 outdegree: 10 total: 20
39 obj: EURYBIA with PageRank: 0.008866 indegree: 9 outdegree: 9 total: 18
40 obj: HADES with PageRank: 0.008834 indegree: 9 outdegree: 9 total: 18

These are the PageRanks obtained from the web scraped dataset of the Greek Gods. The results reveal that the Gods born earlier have a high PageRank prestige.

PageRank Handmade Data - Ordered by Highest Prestige:
Read time: 0.007884
Processing time: 0.006354
Iterations until convergence: 25

pageRanks:
1 obj: Zeus with PageRank: 0.032928 indegree: 20 outdegree: 20 total: 40
2 obj: Nyx with PageRank: 0.023841 indegree: 19 outdegree: 19 total: 38
3 obj: Cronus with PageRank: 0.022431 indegree: 18 outdegree: 18 total: 36
4 obj: Gaia with PageRank: 0.021783 indegree: 13 outdegree: 13 total: 26
5 obj: Titans with PageRank: 0.021533 indegree: 17 outdegree: 17 total: 34

6 obj: Aphrodite with PageRank: 0.021398 indegree: 13 outdegree: 13 total: 26
7 obj: Lapetus with PageRank: 0.020102 indegree: 15 outdegree: 15 total: 30
8 obj: Uranus with PageRank: 0.019563 indegree: 11 outdegree: 11 total: 22
9 obj: Ares with PageRank: 0.019180 indegree: 12 outdegree: 12 total: 24
10 obj: Hyperion with PageRank: 0.018934 indegree: 15 outdegree: 15 total: 30
11 obj: Theia with PageRank: 0.018934 indegree: 15 outdegree: 15 total: 30
12 obj: Eros with PageRank: 0.018354 indegree: 12 outdegree: 12 total: 24
13 obj: Coeus with PageRank: 0.017461 indegree: 14 outdegree: 14 total: 28
14 obj: Phoebe with PageRank: 0.017461 indegree: 14 outdegree: 14 total: 28
15 obj: Oceanus with PageRank: 0.015959 indegree: 13 outdegree: 13 total: 26
16 obj: Tethys with PageRank: 0.015959 indegree: 13 outdegree: 13 total: 26
17 obj: Oneiroi with PageRank: 0.015256 indegree: 12 outdegree: 12 total: 24
18 obj: Rhea with PageRank: 0.014414 indegree: 12 outdegree: 12 total: 24
19 obj: Mnemosyne with PageRank: 0.014414 indegree: 12 outdegree: 12 total: 24
20 obj: Crius with PageRank: 0.014414 indegree: 12 outdegree: 12 total: 24
21 obj: Themis with PageRank: 0.014414 indegree: 12 outdegree: 12 total: 24
22 obj: Demeter with PageRank: 0.013539 indegree: 8 outdegree: 8 total: 16
23 obj: Nemesis with PageRank: 0.013492 indegree: 12 outdegree: 12 total: 24
24 obj: Momus with PageRank: 0.013492 indegree: 12 outdegree: 12 total: 24
25 obj: Hypnos with PageRank: 0.013492 indegree: 12 outdegree: 12 total: 24
26 obj: Moros with PageRank: 0.013492 indegree: 12 outdegree: 12 total: 24
27 obj: Moirai & Keres with PageRank: 0.013492 indegree: 12 outdegree: 12 total: 24
28 obj: Oizys with PageRank: 0.013492 indegree: 12 outdegree: 12 total: 24
29 obj: Apate with PageRank: 0.013492 indegree: 12 outdegree: 12 total: 24
30 obj: Geras with PageRank: 0.013492 indegree: 12 outdegree: 12 total: 24
31 obj: Philotes with PageRank: 0.013492 indegree: 12 outdegree: 12 total: 24
32 obj: Eris with PageRank: 0.013492 indegree: 12 outdegree: 12 total: 24
33 obj: Clymene with PageRank: 0.013434 indegree: 6 outdegree: 6 total: 12
34 obj: Erebus with PageRank: 0.012498 indegree: 7 outdegree: 7 total: 14
35 obj: Thanatos with PageRank: 0.012467 indegree: 11 outdegree: 11 total: 22
36 obj: Oceanids with PageRank: 0.011749 indegree: 6 outdegree: 6 total: 12
37 obj: Helios with PageRank: 0.011623 indegree: 6 outdegree: 6 total: 12
38 obj: Deimos with PageRank: 0.011584 indegree: 7 outdegree: 7 total: 14
39 obj: Himeros with PageRank: 0.011584 indegree: 7 outdegree: 7 total: 14
40 obj: Anteros with PageRank: 0.011584 indegree: 7 outdegree: 7 total: 14

These are the PageRanks obtained from the handmade dataset of the Greek Gods. Here we see that Zeus is correctly placed as having the highest PageRank.

PageRank Web Scraped Data - Ordered by Lowest Prestige:
Read time: 0.006050
Processing time: 0.005491
Iterations until convergence: 25

pageRanks:
1 obj: Athene with PageRank: 0.002469 indegree: 0 outdegree: 1 total: 1
2 obj: Thantos with PageRank: 0.003486 indegree: 1 outdegree: 1 total: 2
3 obj: Athena with PageRank: 0.003786 indegree: 1 outdegree: 0 total: 1
4 obj: Persephone with PageRank: 0.005140 indegree: 2 outdegree: 2 total: 4
5 obj: Typhon with PageRank: 0.005277 indegree: 2 outdegree: 2 total: 4
6 obj: Heliades with PageRank: 0.005810 indegree: 2 outdegree: 2 total: 4
7 obj: Asteria with PageRank: 0.005960 indegree: 3 outdegree: 3 total: 6
8 obj: Hemera with PageRank: 0.006684 indegree: 3 outdegree: 3 total: 6
9 obj: Aether with PageRank: 0.006684 indegree: 3 outdegree: 3 total: 6
10 obj: Dione with PageRank: 0.006970 indegree: 3 outdegree: 3 total: 6
11 obj: Ourea with PageRank: 0.007135 indegree: 3 outdegree: 3 total: 6
12 obj: Pontus with PageRank: 0.007135 indegree: 3 outdegree: 3 total: 6
13 obj: Apollo with PageRank: 0.007202 indegree: 3 outdegree: 3 total: 6
14 obj: Artemis with PageRank: 0.007202 indegree: 3 outdegree: 3 total: 6
15 obj: Atlas with PageRank: 0.007374 indegree: 3 outdegree: 3 total: 6
16 obj: Prometheus with PageRank: 0.007374 indegree: 3 outdegree: 3 total: 6
17 obj: Eos with PageRank: 0.007548 indegree: 4 outdegree: 4 total: 8
18 obj: Selene with PageRank: 0.007548 indegree: 4 outdegree: 4 total: 8
19 obj: Erinyes with PageRank: 0.008681 indegree: 4 outdegree: 4 total: 8
20 obj: Meliae with PageRank: 0.008681 indegree: 4 outdegree: 4 total: 8
21 obj: Gigantes with PageRank: 0.008681 indegree: 4 outdegree: 4 total: 8
22 obj: Chaos with PageRank: 0.008933 indegree: 5 outdegree: 5 total: 10
23 obj: Hecatonchires with PageRank: 0.009185 indegree: 5 outdegree: 5 total: 10
24 obj: Echidna with PageRank: 0.009185 indegree: 5 outdegree: 5 total: 10
25 obj: Cyclopes with PageRank: 0.009185 indegree: 5 outdegree: 5 total: 10
26 obj: Epimetheus with PageRank: 0.009333 indegree: 4 outdegree: 4 total: 8
27 obj: Pleione with PageRank: 0.009459 indegree: 4 outdegree: 4 total: 8
28 obj: Eileithyia with PageRank: 0.009740 indegree: 5 outdegree: 5 total: 10
29 obj: Hebe with PageRank: 0.009740 indegree: 5 outdegree: 5 total: 10
30 obj: Enyo with PageRank: 0.009740 indegree: 5 outdegree: 5 total: 10
31 obj: Hephaestus with PageRank: 0.009740 indegree: 5 outdegree: 5 total: 10
32 obj: Hestia with PageRank: 0.010164 indegree: 6 outdegree: 6 total: 12
33 obj: Poseidon with PageRank: 0.010164 indegree: 6 outdegree: 6 total: 12
34 obj: Hades with PageRank: 0.010164 indegree: 6 outdegree: 6 total: 12
35 obj: Tartarus with PageRank: 0.011006 indegree: 6 outdegree: 6 total: 12
36 obj: Leto with PageRank: 0.011212 indegree: 6 outdegree: 6 total: 12
37 obj: Inachus with PageRank: 0.011351 indegree: 5 outdegree: 5 total: 10
38 obj: Melia with PageRank: 0.011351 indegree: 5 outdegree: 5 total: 10
39 obj: Hera with PageRank: 0.011520 indegree: 7 outdegree: 7 total: 14
40 obj: Deimos with PageRank: 0.011584 indegree: 7 outdegree: 7 total: 14

These are the PageRanks obtained from the web scraped data of the Greek Gods. These results reveal that the Gods born later or those that never had little baby Gods had the lowest prestige.

6.2 Movie Recommendations Dataset

Here is a sample run of the recommendation program:

```
 $ python recommend.py 20
362 313 4 0.498441468033 3.50155853197
776 496 3 0.49943031644 2.50056968356
271 87 3 0.0 3.0
170 292 5 0.0 5.0
739 96 5 0.619936149637 4.38006385036
795 719 2 0.0 2.0
922 227 4 2.61134364731 1.38865635269
621 540 3 0.0 3.0
896 423 3 3.00228024823 0.00228024822585
837 275 4 2.01416432542 1.98583567458
729 346 1 0.0 1.0
680 273 3 0.0 3.0
539 170 5 0.0 5.0
62 138 1 0.0 1.0
90 196 4 0.607998216156 3.39200178384
600 172 4 3.11084811171 0.889151888293
159 1012 5 0.0 5.0
366 854 5 0.0 5.0
343 786 4 0.0 4.0
597 275 4 0.523835800669 3.47616419933
Mean absolute error 2.92581411064
```

recommend.py takes a parameter that represents the number of ratings to try. In this example, recommend.py tries to predict the rating of 20 random samples. Each line represents the user id, movie id, actual rating, predicted rating, and delta in ratings. At the very end of the output, the mean absolute error is given. We were not very happy with our error since it alternated between 2.5 and 3.3 for the sample runs that we did. Ideally, we would have liked to see the mean absolute error between 1 and 2.

7. Conclusion

In conclusion, applying the PageRank algorithm on the Greek Gods dataset was successful in obtaining the correct rankings. Choosing to web scrape from an incomplete database, however, resulted in rankings being skewed and some Gods like Zeus being underrepresented. After applying the PageRank algorithm to a handmade

dataset, which was obtained from an online image of a family tree, the results correctly ranked Zeus at number one with Nyx, the Goddess of Darkness being placed in second place. Both questions which were asked initially about the Greek Gods were answered by the PageRank algorithm. The first question of which Gods were most related to other Gods was answered by obtaining the highest ranked prestiges; Nyx, Gaia, and Uranus were ranked first, second, and third respectively. By analyzing the results, the PageRank algorithm also seemed to rank the Gods in higher positions if they were born earlier and ranked them lower if they were born later. These results answered the second question about this dataset - which Gods were born earlier and which gods were born later?

While we did not achieve a low mean absolute error, we enjoyed this project very much. We learned about the intricacies of recommender systems and we gained an appreciation for the customizations that are needed in recommender systems. From choosing the number of neighbors to the similarity formulas. While we used item-based collaboration, we had to learn about user-based collaboration to choose between one or the other. We wish we had more time to improve our recommender system.

[1] Dataset obtained from http://www.godchecker.com/pantheon/greek-mythology.php?list-gods-names

[2] Family Tree obtained from https://hesiodstheogony.files.wordpress.com/2010/12/family-tree.jpg