AIT580 – Data Analysis Individual Project

Deliverable 1 – Dataset Selection:

I. Who (company, agency, organization) collected the data?

The data was collected by OpenDota (formerly YASP) – "a volunteer-developed, open source platform providing Dota 2 (an eSports battle arena game) data. It provides a web interface for casual users to browse through the collected data, as well as an API to allow developers to build their own applications with it" (Cui, 2014).

II. Need

Why collect the data?
The contributors started collecting the data out of curiosity – they wondered what data they could extract from game replays. They were interested in building fun tools for themselves and then sharing those tools with others in the community. The project has stayed true to its open-source roots and to this day releases their parsed data through their OpenDota API.

Big data problem, data sources
Volume and velocity make this a big data problem. At the end of 2015, they released a large parsed data dump of 3.5 million matches played during 2015 that amounted to a gzip JSON file of 100GB. As of 2016, people are playing about 1.1 million matches every day and 1.4 million on the weekends with most users playing around 3 matches a day, though OpenDota only parses a fraction of all these matches. "Data is collected through the Steam WebAPI, as well as replay parsing of [Dota 2 replay] files" (Cui, 2014). Steam community profiles are included as well.

Data types
The dataset includes an array of match objects stored in JSON format. This means there is a long list of objects each with similar structure containing information about an individual match that occurred. This includes a match_id, radiant_win (which team won), start_time, duration, human_players (number of players in game), chat, skill, players (list of players in game with data about each player's game outcome), etc. It is quite detailed, enough for a large range of gamers from amateur to professional to find it very useful in diagramming past games and learning from their mistakes and successes to improve their skills and strategies.

Data owners – access, privacy, quality
This is a public dataset that was released with a CC BY-SA 4.0 (Creative Commons license) on AcademicTorrents.com. All of the data collected comes from public sources and Valve (Dota2's parent company) allows each user to opt-out of sharing their identifiable game data by unchecking a box in the game. This will not remove existing data from their databases, but will make the user appear anonymous in future matches. I don't have a good idea about the quality of the data, but the project seems to be well funded and has a solid community of developers behind it along with a large community of users who enjoy its services.

III. What potential questions could be answered by studying this data?

- What is the average number of kills in a Dota 2 game stratified by game duration?
- What is the distribution of game duration?
- How does kill to death ratio correlate with experience per minute (xpm)?
- Does the team that drew first blood (the first game kill) win more often?
- Does gold per minute (gpm) or xpm have a higher correlation with a game win?
- Do players with 500 or more gpm win games more often than those with less than 500 gpm?
- What features lead to an edge in winning (> 50% accuracy)?

IV. What software and hardware resources will you need to study this data?

To analyze the full dataset (700 GB compressed gzip files), it will require a rather large and high performance hard disk at least 10 TB in size, preferably NVMe SSDs in a RAID configuration to ensure data protection in case of disk failure. Fast DDR4 ram with at least 16 GB size and a mid to high tier GPU would also be necessary to perform more complex analyses like AI and deep learning. For software resources, a computer with Windows 10 or a Linux distribution would be required with Python or R installed with their many analysis packages. A relational database management system like PostgreSQL / MySQL or a NoSQL database like MongoDB / Cassandra are also required to facilitate the large data management needs.

V. References

Cui, A. (2014). FAQ. Retrieved 18 February 2021, from https://blog.opendota.com/2014/08/01/faq/

Cui, A., Chung, H. YASP December 2015 500k Data Dump. OpenDota.com. 18 December 2015. Retrieved 18 February 2021, from https://academictorrents.com/details/384a08fd7918cd59b23fb0c3cf3cf1aea3ea4d42. Formatted as a gzip JSON file.