World of Warcraft Character Dataset

Author: Jinyuan Qiu

Overview

For the final project, I will work alone to analyze a set of data from an online multiplayer game called "World of Warcraft". I will look for relationships as well as create graphs to evaluate the relationships between the levels of the players and the class that they play as well as the relationships between the race and the class that they play. I will also calculate the probabilities of each attribute of a player for each set of players in a given zone of the game to find the 'representative' player of that zone.

Motivation

The results of this program will allow players as well as companies to see trends in the World of Warcraft population as a whole and make choices more intelligently whether if it is playing a character or class that is less played or looking for trends in the general World of Warcraft. This project will also allow Blizzard, the company that runs World of Warcraft, to see the' typical' characteristics of players who are in a given zone as well as to find race, class combinations that are less popular.

Dataset

The dataset is given in the form of a .txt file located within folders with dates as its name. I will be taking the sample from year 2009 January because the size of the dataset would be appropriate for this assignment. Judging by the .txt files, the data is given in a SQL statement format, in two arrays called Persistent_Storage and RoundInfo. All the information collected about the avatars' history is stored in the Persistant_Storage array. Each element of the Persistant_Storage is a string containing 11 fields separated by commas in the following order: dummy, query time, query sequence number, avatar ID, guild, level, race, class, zone, dummy, dummy.

For Example:

"0, 01/01/09 01:44:52, 1,78649,478, 55, Orc, Hunter, Azshara, HUNTER, 0", -- [1]

I can extract the relevant lines by counting the amount of commas in each line. The fields which are relevant to my analysis are the level, race, class and zone.

The dataset can be downloaded from:
http://mmnet.iis.sinica.edu.tw/dl/wowah/

Methodology

To find the relationship between the levels of the players and the class that they play, I will create a data structure where I can obtain the counts for different classes based on a level. I will then output the top 10 levels and the bottom 10 levels followed by the respective player counts of each class. I will also create graphs for sums of the top 20 levels and the bottom 20 levels where x is the different classes and y is the sums of the players for each class. The analysis will be significant if there is a high amount of players in one class in the higher levels but no significant differences in the lower levels. Therefore the p-value r is calculated afterwards to show the correlation between the two sets of results; the count of players in each class for upper 10 levels verses the count of players for each class in the bottom 10 levels. A high value of |r| (0.5 to 1.0) will mean that there is a strong correlation between the two different datasets, in this case the counts of players in each class of the lower levels versus those of the higher levels. In other words, it will suggest that the classes that players play in the lower levels are indicative of the classes that upper levels play. This will be expected because in order to reach higher levels the player must stay in the class that they chose and level up. If the p-value is low (0 to 0.5), we will know that the classes that players play in the lower levels are not indicative of the classes that upper levels play, so we know that most of the players would most likely be switching classes to a different one. The class “death knight” would not be included in the p-value calculation because this class starts at level 60.

To find the relationship between the race and the class I will do the same thing as above but instead of levels, it will be the different races in the game. This time instead, I will attempt to predict the class that the player will choose after picking a race. This can be determined by finding the class with the highest percentage of players within the same race.

To find the trends of in a given zone, I will this time create a data structure where the set of most common traits can be obtained through a given zone. The set of most common traits will be defined to be each individual trait with the highest counts in each category in the given zone (i.e. the most common class is Hunter). This will then be outputted as 229 strings (one for each zone).

Results

After completing and running the program, I discovered that from levels 70 to 80, the most popular classes are death knight, hunter, and priest in the that order, whereas, in levels 0 to 10, the most popular classes are hunter, paladin, and warrior in the that order. The two tailed p-value of the two sets came out to be -0.27 and 0.49, which suggests that players typically swaps classes in between levels 0-80. This suggests that players tend to switch toward classes that are “stronger” than others and those classes would be death knight, hunter, and priest respectively. This trend is also noticeable in the result because the counts of the ‘priest’ class, which placed 2nd to last in levels 0 to 10 placed 3rd from levels 70 to 80.

As for the graphs of the character counts of each class between the top 20 levels and the bottom 20 levels, it is also significant to note that the top 3 played class in the top 20 levels is still Death Knight, Hunter, and Priest. The top 2 character counts of the bottom 20 levels also remained to be Hunter and Paladin.

As for the predictions for each race, if a random player is chosen from the Troll race, he or she would most likely be a Hunter. An Undead would most likely be a Mage. A Tauren would most likely be a Druid. An Orc would most likely be a Hunter. Lastly, a Blood Elf would most likely be a Death Knight.

Out of the 229 zone-character predictions, traits that were common based on the results printed from the program were Blood Elves and Death Knights. This also is a reasonable result because, as we could tell form the level-class relationships earlier, the most popular class in the higher levels would be Death Knights. Based on the outputs of the program, the results from the zone-character predictions also suggest that most zones in World of Warcraft contain level 80 players because the most common level for the players in each zone 80.

Reproducing the Results

In order to reproduce the results of this analysis, one can download the dataset from

http://mmnet.iis.sinica.edu.tw/dl/wowah/

After extracting the .rar file there will be two folders “img” and “WoWAH”. After opening the “WoWAH” folder there will be a set of folders with dates as its names. In my analysis, I used the “2009_01” folder, representing the data collected from January, 2009. Extract 2009_01 folder to the same directory as my python program will allow it to be read by the program.

To execute the program in the command line client of the computer, one needs to navigate to the directory where the program is located along with “2009_01” folder. The command to type in the command line client to reproduce my results is “python wowdata.py 2009_01”.

Collaboration

I worked alone on the assignment.

Reflection

This assignment takes a lot of thinking and it is the very definition of working on real world problems. In the future I would suggest students work in groups if possible because it would lead to a lot of discussion about what data do they want to obtain and how to obtain and calculate them. The due date of this assignment is also kind of rushed because, from the time that the students gets back their comments for part 1, the students only have approximately 4 days (not including late days) to do part 2.