ASSIGNMENT 1

Elements of Data Processing

**Introduction**

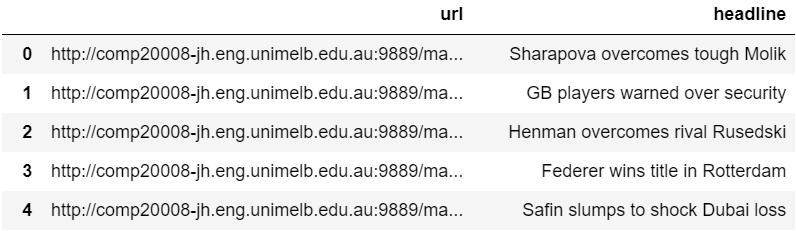
This report attempts to communicate the processing and exploratory analysis techniques used throughout the given tasks in assignment 1. The analysis focused on understanding player performance from a number of media reports on tennis matches between 2004 and 2005. This report contains a summary of activities undertaken, analysis used in each task, visual analysis where possible, limitations and suggestions to better understand the performance.

**Task 1**

The first task required us to set up a crawler (a program to scour the given domain in methodical manner) and collect all articles linked. The link given in question served as the base URL, the link from where the crawler can start working to retrieve information. The crawler then followed the steps of

* Scanning for new links (<a> tags in HTML)
* Storing every link in a list of pages to visit
* Adding the current page to the list of pages visited
* Extracting the headline from the <h1> tag

It was important to ensure the crawler would not back-track into an infinite loop and so a list of pages visited was used. And every headline was stored into a list with its corresponding URL visited and were merged onto a pandas data frame. 100 articles were found and stored as shown below. This was done using the urllib and BeautifulSoup4 library.

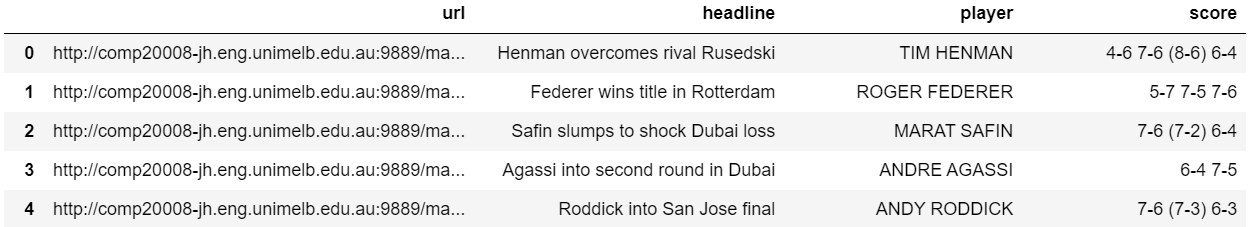


**Task 2**

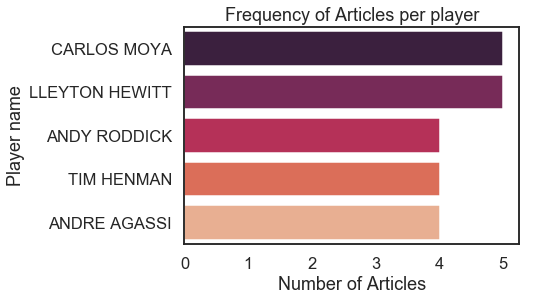
The task was split into two parts – to extract names from every article, then identifiable scores from the respective names. And to help identify names, a .json file was provided containing a list of players and their data. The scraper (piece of code that extracts data from a given website) loaded the entire HTML code as text and searched for a matching name. The assumption for the first part was that the article was related to the first player mentioned and was about that player only. Therefore, the logic was to loop through the entire article – for every URL in task 1 – and find a first and last name(since players can have similar last or first names) that matches any player in the .json file.

And for every name found in Task 2A, we were to find a valid score. That was done using the RE library and the expression – ‘\d-\d ((| |\()\d+(-|\/)\d+( |\)|.))+’.

The method used was similar to 2A where we had a list of URLs to go through, until it matched the regular expression in the text. Our count came down to 41 articles that contain a valid score. The expression used searched for two or more scores where the second score may or may not contain brackets, and end with a space or period. After that I visually inspected for outliers or abnormal scores such as articles where players had retired – which required removal as mentioned in the FAQs. Once that was done, I cleaned the score from punctuation and linked the articles to their respective player names in the .json file and stored in a pandas Data Frame as seen below. The output contains 4 required field – URL, headline, player (names), and scores.

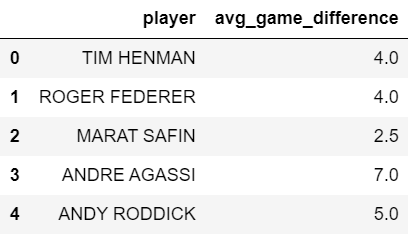


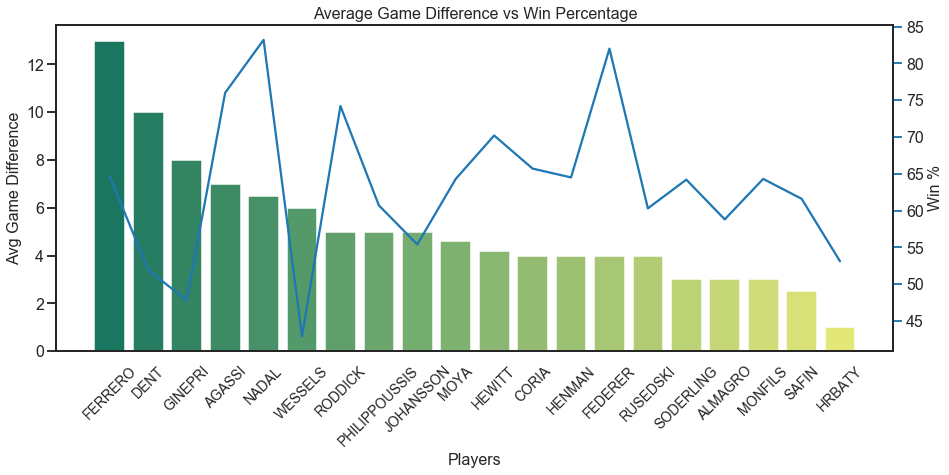
**Task 4**

Task 4 required us to plot a graph of the top 5 players about whom articles were most frequently written about. The data used for graphing was obtained from task 2 which refined all the articles to only those that have a relevant score along with names mentioned. A dictionary was used to count the number of occurrences of players.

The graph indicates the Carlos Moya and Lleyton Hewitt were popular in several articles (5 each) around 2004-05. This matches the events we know have occurred during that time period, i.e. Moya won the Davis Cup in 2004, and Hewitt won the same in 2003. Along with them were Andre Agassi (Rank no. 8 – 2004), Andy Roddick (Rank no. 2 – 2004), and Tim Henman (Rank no. 6 – 2004). The data provided seems to be consistent with the timeline in the past. From our limited sample size, these players took the spotlight with most mentions, but does not indicate whether it was for a victory, loss, or any other event (injury, forfeit, etc.).

**Task 5**

The plot used data from the .json file for win percentages and average game difference from task 3. In the third task we computed the difference based off the found scores in task 2 and averaged multiple occurrences of a specific player’s match as seen below. This was done by simply adding the scores on either side of the ‘-‘token to two players and computing the absolute value. We also assumed that the score found was matched to the player linked with that article. With two streams of data sharing a similar X axis, I could make two graphs and combine them into one. Both the plots above used seaborn since it is more visually appealing and easier to use while retaining concepts from matplotlib.



The average game difference indicates the margin by which the player wins the match and his win percentage tells us his consistency of winning. As seen in the image, there is no definite pattern between the two plots however, there are certain outliers like Agassi, Federer, and Nadal with a high win percentage who have truly are this era’s tennis stars. While Ferrero, Dent, and Ginepri win big or lose big. Also, most of the players tend to have a game difference of 3-6 points. What is rather unusual is that Hrbaty had a low win percentage and low game difference but came back to win the Davis cup in 2005.

**DISCUSSION**

**Q.** Discuss the appropriateness of associating the first name in the article with the first matched score.

The first name presented in the article can be a fair assumption to use since articles generally start by bringing the main point in and then talk about the events at the scene, which include stating the match score. Some examples of this would be – “Rafael Nadal brings the championship home…”, “Andy Roddick injured while …” and then the article moves ahead with more content in relation to the opening lines. However, this would only work when we have an updated list of names to compare with. In this case, women tennis players are not updated in the .json and so any male mentioned in a women’s article receives their mention and score. This would also fail when there are many names in the article, as it would be tough to link the score to the many players mentioned. Therefore, with our data, the accuracy is around half or more since there are a greater number of articles with male tennis players than female tennis players, but the accuracy could decrease with every name being matched to the first score which shows that it is not the most appropriate method to use.

**Q.** Suggest at least one method on how you could figure out whether the first named player won or lost the match being reported.

Player names are often accompanied by keywords with respect to their victories or losses. Some examples of victory keywords that follow or surround the player are – “won, defeats, claims victory, beat, takes home, etc.” while some loss keywords that surround the player are – “Surrenders, gives in, loses, is defeated, etc.” Looking for these keywords that accompany the player name would be a good indicator to whether he/she won or lost. And similar to finding the name, comparing it to a list of these victory/loss keywords would be helpful to figure out the outcome of the match.

This is homogenous to the core idea of NLP, where stop words are removed, and key features are extracted from the text.

**Q.** What other information could be extracted? And how could this be done?

Other Information such as whom the first player played against would be vital to building a database and updating matches played by every player involved. And surrounding information such as the tournament played (Davis Cup, Australian Open, etc.), location, and date would also be useful.

This could be done similar to how we find names, where we match words in the article to certain keywords we expect. Having an updated list of tournaments during that time period and list of locations would be useful to compare and collect data.