Task B2: Visualization of Characters

Shubham Gupta
A0225160U

April 14, 2022

1 Introduction

- For this task, we aim to analyze datasets focussed on characters, both fictional and real-world based.
- We first analyze the dataset containing information about Messi and Ronaldo, two of the most legendary football players of all time. The dataset contains statistics about their goals scored, the number of minutes played, etc. We use this dataset to create **Statistical Visualizations**
- Next, we analyze the Harry Potter Movie dataset, and aim to create **Statistical and Text based** visualization.
- We aim to answer a few question with each visualization.

2 Datasets

2.1 Messi-Ronaldo Dataset

- The Messi-Ronaldo dataset was obtained from a previous MakeoverMonday challenge, and it is available here.
- The dataset has the following fields:
  - **Season**: String variable, specifying the season for which the statistics are given.
  - **Player**: Categorical variable, specifying the player. It can only be "Messi" or "Ronaldo".
  - **Liga_Goals**: Ordinal variable, specifying the number of goals scored in La Liga by the given player.
  - **Liga_Asts**: Ordinal variable, specifying the number of assists given in La Liga by the given player.
  - **Liga_Aps**: Ordinal variable, specifying the number of appearences in La Liga by the given player.
  - **Liga_Mins**: Ordinal variable, specifying the number of minutes in La liga by the given player.
  - **CL_***: Four columns that are ordinal variables. These columns specify the above stats for the Champions League contest.

2.2 Harry Potter Dataset

- The Harry Potter dataset consists of dialog by each character in each movie, along with metadata about the movies such as duration, total budget, etc.
- The main fields in the **movies.csv** file are:
  - **Movie**: String/Categorical Variable, containing the movie name.
  - **Released year**: Ordinal variable, containing the year of release of the movie.
  - **Running Time**: Quantitative variable, containing the running time of each movie in minutes.
  - **Budget**: Quantitative variable, containing the total budget for the movie.
  - **Box Office**: Quantitative variable, containing the total box office collection for the given movie.
- There are also individual files for each movie, titled hp1.csv, hp2.csv, etc. Each of these files containing the following main fields:
  - **Movie**: String/Categorical Variable, containing the movie name.
  - **Chapter**: String variable, containing the name of the chapter in the movie.
  - **Character**: String/Categorical variable, containing the name of the character speaking the current dialog.
  - **Dialog**: String variable, containing the actual dialog spoken by the character.

3 Preprocessing

- For the Messi-Ronaldo dataset, we only perform basic preprocessing such as creating seperate dataframes for Champions League and La Liga statistics, and renaming the columns to be more human-readable. For example, changing Liga_Asts to Assists in the La Liga dataframe.
- For the Harry Potter dataset, while we do not do any preprocessing initially, we do some preprocessing for each visualization, which will be detailed later in the document.

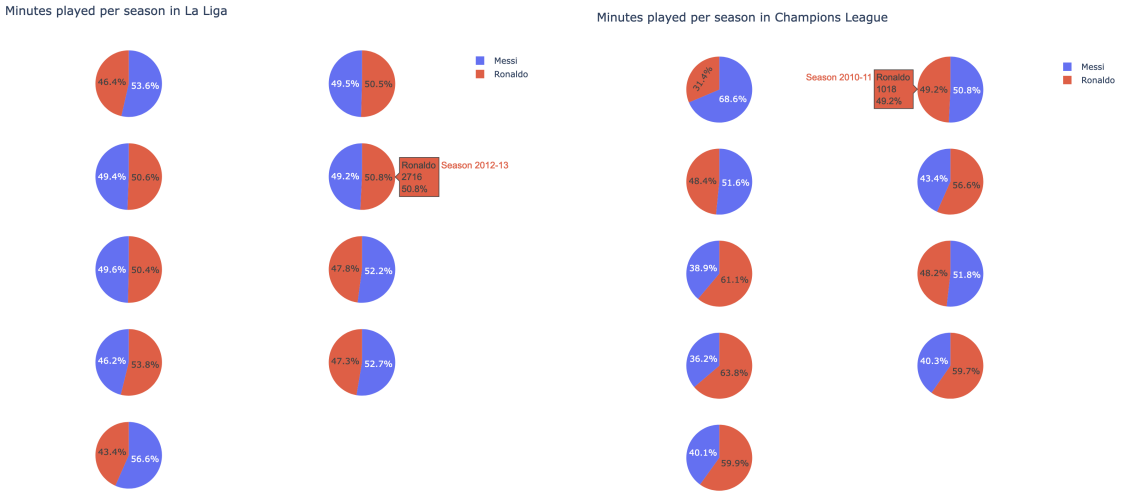
4 Implementation

4.1 Messi-Ronaldo Dataset: Statistical Visualizations

- For this dataset, we focus on creating **Statistical** visualizations.

4.1.1 How many minutes per season do the players play?

- For this question, we aim to visualize the number of minutes played by each player in both the competetions, La Liga and Champions League.



(a) Minutes played per season in La Liga

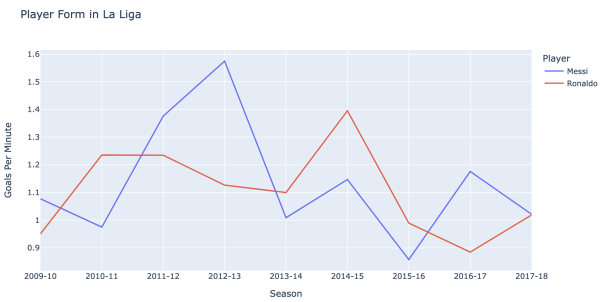
(b) Minutes played per season in Champions League

Figure 1: Minutes played per season

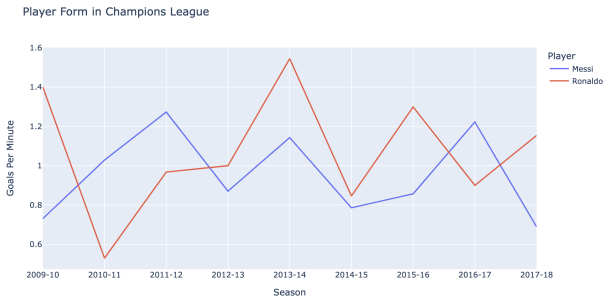
- From the visualization, it's clear that the focus of Ronaldo shifted from la Liga to the Champions League towards the last few seasons, whereas Messi aimed to balance his time between both competetions more evenly.
- For the visual encoding, we use the following:
  - Chart type used is a **Pie Chart**
  - We use colors as a **Channel**, with Red signifying statistics for Ronaldo and Blue for Messi.
  - We also use a **Ordering Direction** from left to right, with the left topmost pie chart being the statistics for the very first season i.e the 2009-10 season.

4.1.2 Player form over all seasons

- Here, we aim to analyze the player form over all seasons. To determine the player form, we use a crude metric of number of goals scored per minute in each competetiion.



(a) Form per season in La Liga



(b) Form per season in Champions League

Figure 2: Player Form per season

- From the visualization, it's clear that Messi was dominant in La Liga during the 2011-12 and 2012-13 season, in which he scored a record 91 Goals. We can also see that in the same time, Ronaldo was more dominant in the Champions League, averaging almost 1.6 goals per minute!
- For the visual encoding, we use the following:
  - Chart type used is a **Timeseries based Line Chart**
  - The x-axis contains the season, which is a string variable.
  - The y-axis contains the average number of goals per minute, which is a quantitative variable.
  - We use colors as a **Channel**, with Red signifying statistics for Ronaldo and Blue for Messi.

4.2 Harry Potter Dataset: Statistical and Textual Visualizations

- Here, we aim to analyze the Harry potter dataset, and focus on creating both statistical and text based visualizations.

4.2.1 Running time of each movie?

- We aim to visualize the running time of each movie in a simple plot.

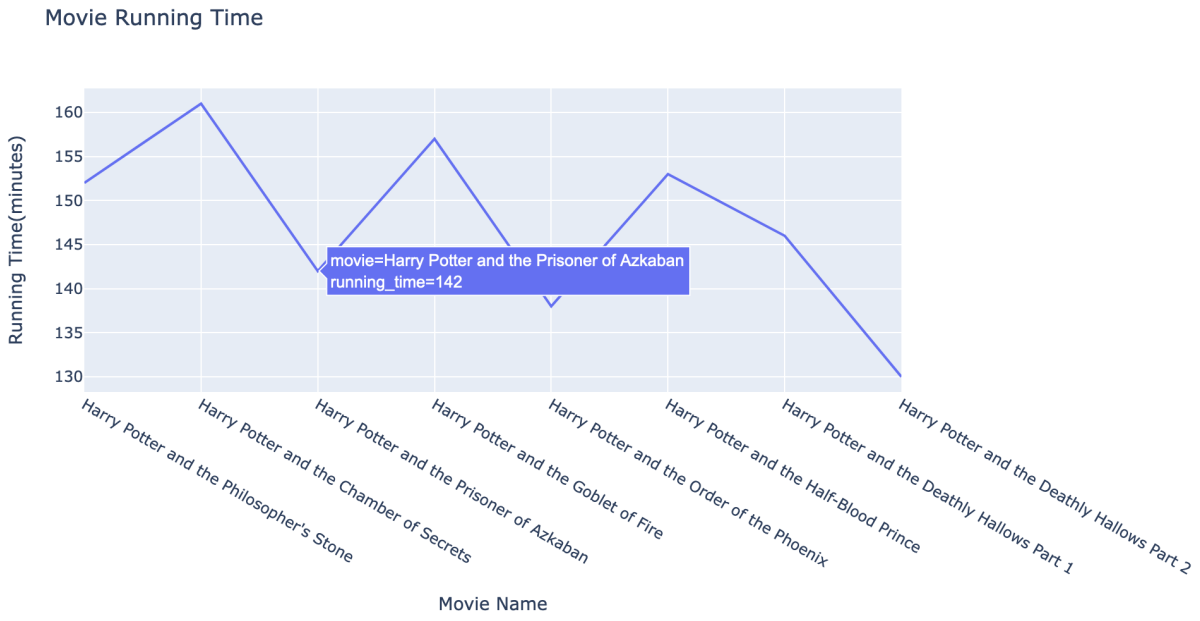


Figure 3: Running time for each movie

- From the above figure, we can see that while the first few movies focussed on having a high running time, the movies have grown shorter as time has passed, possibly to account for shorter attention spans!
- For the visual encoding, we use the following:

- The chart type used in a **Line Chart**
- The x-axis contains the name of the movie, which is a string/categorical variable.
- The y-axis contains the total running time of the movie, which is a quantitative variable, in minutes.

4.2.2 Budget vs Box Office collection

- We aim to visualize the budget vs box office collection, as an indicator of how popular a given harry potter movie was during the time it was released.

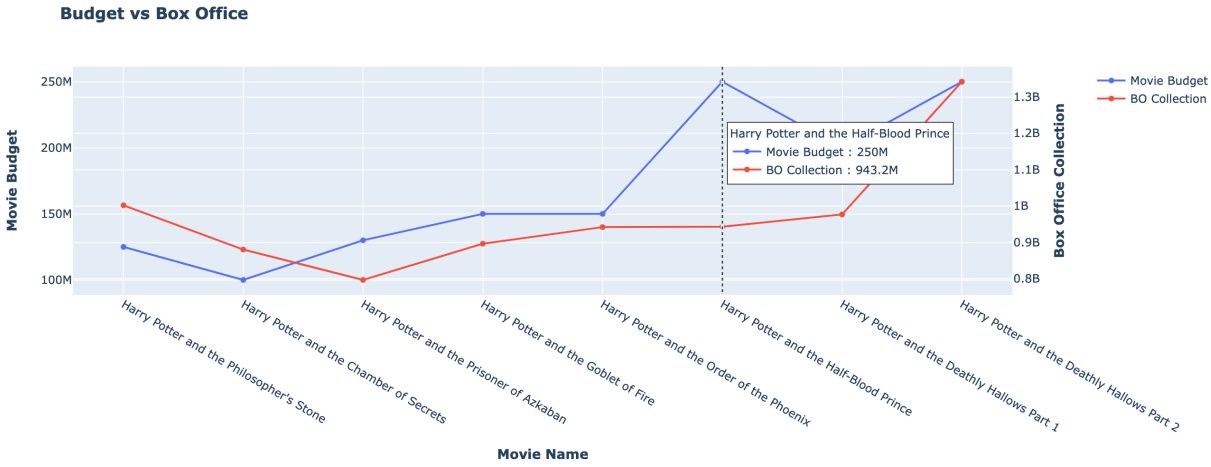


Figure 4: Budget vs Box Office Collection for each movie

- From the visualization, we see that only the first and the last harry potter movies had a collection of more than 1 Billion Dollars!
- The visual encoding used is as follows:
  - The chart type used in a **Line Chart with Dual axis**
  - The x-axis contains the name of the movie, which is a string/categorical variable.
  - The y-axis on the left, is a quantitative variable, and contains the budget of the movie. The scale is in millions
  - The y-axis on the right, is a quantitative variable, and contains the box office collection of the movie. The scale is in Billions.
  - We also use color as **Channels**, with blue depicting the movie budget and red depicting the Box office collection.

4.2.3 Who had the most number of dialogues?

- Here, we aim to determine which character had the most number of dialogues, across all movies.

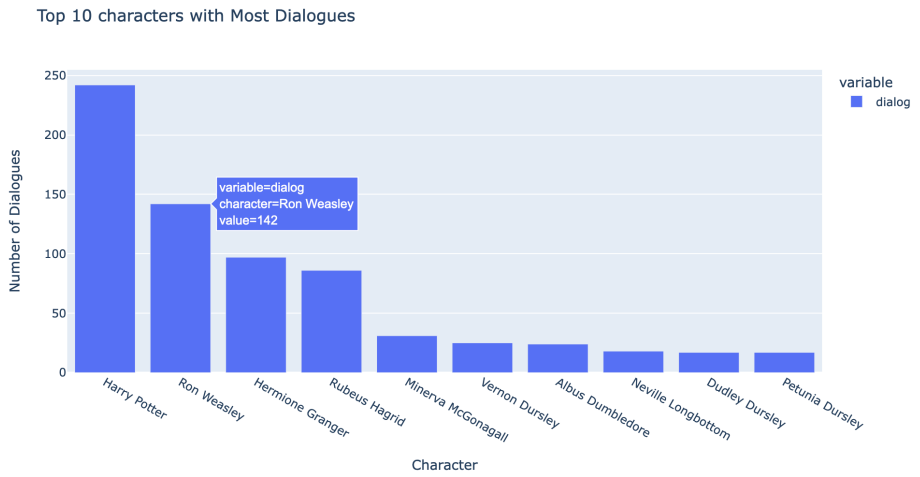


Figure 5: Top 10 characters with most dialogues

- We see that the main protagonists had the most amount of dialogues, followed by their supporting characters/mentors such as Hagrid, Dumbledore, etc.
- The visual encoding used is as follows:
  - The chart type used is a **Bar Chart**
  - The x-axis is a string/categorical variable, and denotes the name of the character.
  - The y-axis is a quantitative variable, and denotes the number of dialogues spoken by a given character across all movies

4.2.4 Word Cloud of Dialogues

- Here, we aim to visualize the dialogues spoken by Harry Potter in all movies.
- For preprocessing, we lowercase all the given words, and remove the stopwords.
- We create a contour mask using a harry potter stencil image, which gives the visualization the shape of the actor.
- Finally, we only display the top 2000 words.

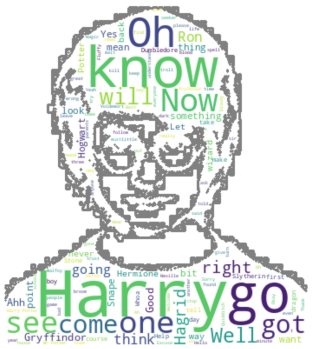


Figure 6: Wordcloud of Harry Potter Dialogues

- From the above visualization, it's clear that:
  - As expected, "Harry" is the most common word.
  - Apart from mentions of other characters, we also see mentions of the different in the movie, such as Gryffindor(the house to which Harry was assigned) and Slytherin(The house which Harry was praying was not assigned to him!)
- The chart type is **Wordcloud**. Size of the word is used as **Channel** to denote how many times the word was spoken by the character.

4.2.5 Distribution of words spoken by a character.

- We aim to visualize the different sequence of words using a wordtree.
- Here, we analyze the dialogues of Dumbledore, but focus on the word "Potter" i.e the instances where he has mentioned potter.

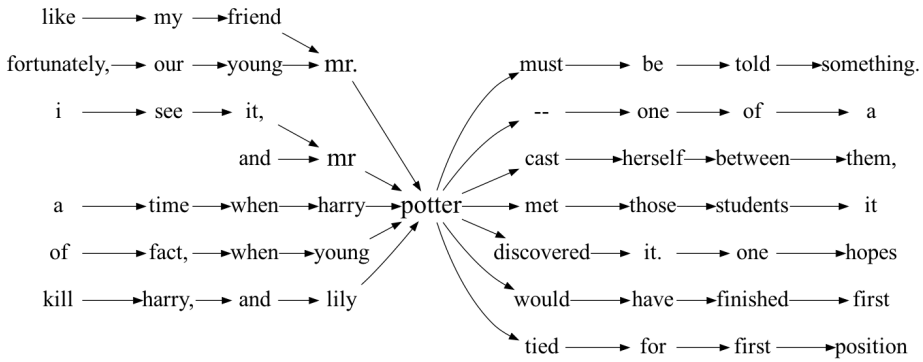


Figure 7: Wordtree of Dumbledore Dialogues for the word Potter

- The chart type used is a **Double Wordtree**

4.2.6 Topic Detection

- Here, we aim to visualize the main topics present in the dialogues by Harry Potter.
- For preprocessing the data, we convert all words to lowercase and filter words with alphabets that are atleast 3 letters long.
- We also remove the stopwords to reduce the noise in the topics.
- For creating the vectors, we use Term Frequence - Inverse Document Frequency(TF-IDF)
- Finally, we detect the topics using the Latent Dirchlet Algorithm(LDA).

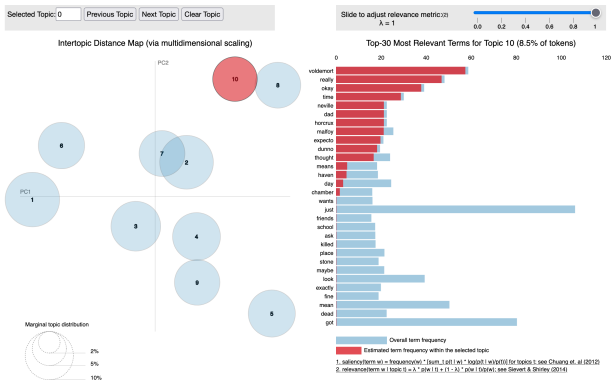


Figure 8: Topic Modelling using LDA on dialogues by Harry Potter

- From the above visualization, we see that:
  - We experiment with different number of topics, and obtained the best results when the number of topics is 10.
  - There are other topics dedicated to specific characters such as:
    - Topic number 10 is specifically for voldermort and horcuxes, which are a source of his power!
    - Topic number 8 primarily contains dialogues about Ron.
    - Topic number 6 primarily contains dialogues about Hermione.
  - We also see that, while topic number 1 is the largest, the words in this topic are general and don't point to anything specific. This could be considered as a "noisy" topic.
- This visualization is a **Meso Level** visualization, since it focus on only the dialogues of Harry Potter.
- We use circles as **Mark** to show the words in each cluster, and the size of the circle as **Channels**, to determine the number of words in the cluster. Large cluster contains more number of words.
- Finally, on the right, we use a **Bar Chart** to display the total number of occurences of each word in the given topic.

5 References

- Topic Modelling using LDA
- Introduction to Wordtrees
- Wordcloud in Python