ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

DATA VISUALIZATION

PROCESS BOOK

MILESTONE 3

Antoine Daeniker
antoine.daeniker@epfl.ch - sciper:287913

David Desboeufs
david.desboeufs@epfl.ch - sciper:287441

Jérémie Frei
jeremie.frei@epfl.ch - sciper:247316

3rd June 2022

1 INTRODUCTION

The Harry Potter universe is a reference the modern literature. Indeed, it is without doubt one of the most famous sagas ever written. After seven bestseller novels and eight blockbusters in the span of 15 years, its roots in pop culture are still very much alive.

Its rich narrative and elaborated plots makes the whole universe difficult to grasp. This is why we had the idea of clarifying some key properties of these masterpieces.

2 DATASET

Our dataset changes many times during our project but we explain more about this in the faced challenges section. But this is our final dataset that we used in our visualisation :

- the seven books of Harry Potter in $.txt$ format
- Positive and negative words
- Spells in $.csv$ format

The principal resource were the seven books that we explore in many different ways and many times. We realized that our word analysis can bring out many different information about the books that would be impossible to see by just reading them. Either by searching specific words or by finding which words come out under certain searching conditions, we always got the result we could expect, since the team members know the story and/or have read the Harry Potter books.

The positive/negative words file has two columns (positive, negative) and we use this data to produce the sentiment analysis graph.

Concerning the spells dataset, the columns are :

- Name : The name of the spell
- Incantation : How the spell is cast
- Type : The spell type
- Effect : What the spell does
- Light : Potential light that appears when the spell is cast

We used the the Name and Incantation to search specific spells in the books and the Type column to classify them in our bubble chart.

In the beginning, we had data about characters and our first idea was to display a genealogy graph of the families of the Harry Potter universe, but we had issues about how to connect characters between them. Again, we explain those issues in the faced challenges section 4.1.1.

Moreover, we had another dataset about all potions in Harry Potter. Unfortunately, there wasn't enough shared information between potions and hence it would have been difficult to make relations between them in order to produce a nice visualization.

3 VISUALISATION

3.1 SPELLS BUBBLE CHART

We use our Spell data and the seven books for this visualisation. The first stage of the bubble chart displays the different types of spell in our dataset. The size of the bubbles differs depending on the number of different spells of this particular type. The second stage (that appears by clicking on one of the bubbles) displays all different spells of a particular type. The size of the bubble depends on the number of times that the spell is cited or cast in all books. That is why we used the name and the incantation of each spell since some spells are only cited or cast.

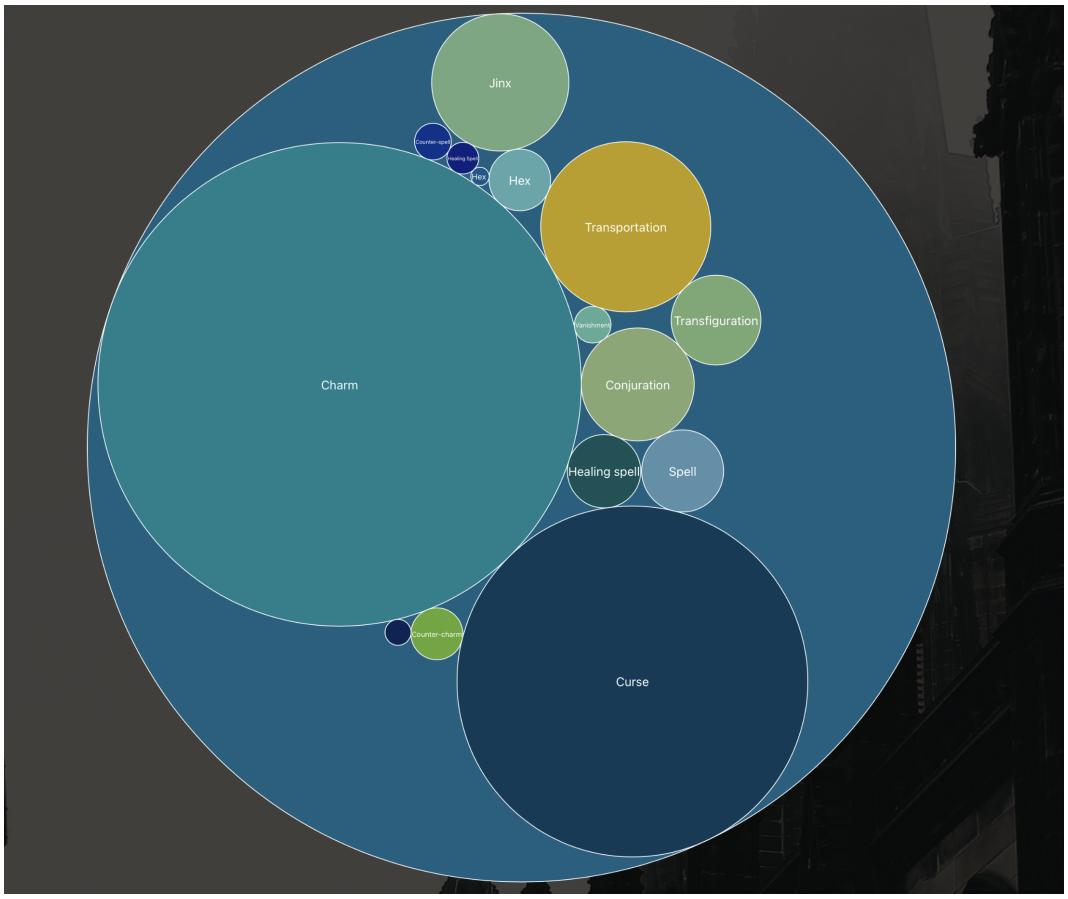

FIGURE 1
Our Bubble Chart For Spell Occurrences

3.2 SPELLS GRAPH

The first spell visualisation was just to display a raw representation of each spell in the book, like a list. After a meeting with the professor, we found out that showing spells evolution through the books could have a nice look and may allow to visualize some hidden aspect between certain spells. To do so, in the same way as in the bubble chart, we sort all spells by number of occurrence, and then take only the top five in each book. By displaying them on a graph, we could see if some spells appear in only one book or in many books.

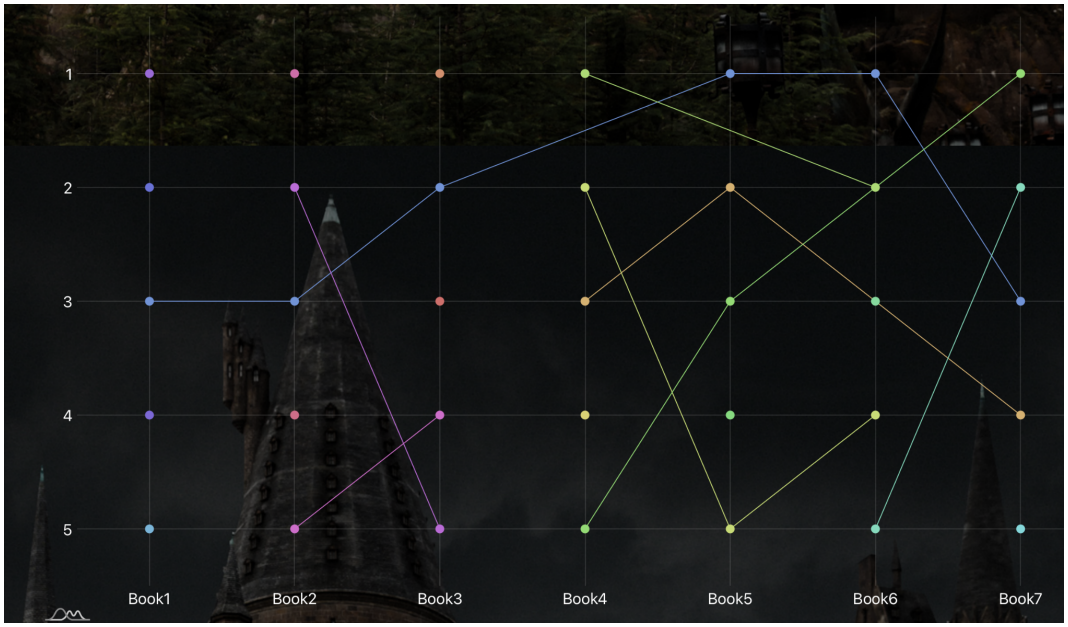

FIGURE 2
Our spells graph evolution

3.3 WORD CLOUD

First, we wanted to display a 1-gram words occurrence for each book without any other processing, except the removing of all stopping words like "the", "is" or "you". We found out that some words like "Harry", "Hermione" or "Ron" are used many times in each book compared to other important words in the story. And so they crushed other words in our visualisation. Hence, we decided to regroup all books and get all the words' occurrence. Now that we have this, we take each word's occurrence of each book and remove the top ones that appear in the total words and hence we have all specific words concerning each book.

Also, after a conversation with the professor, we thought that it would be interesting not only to gather 1-gram word but also 2-grams, 3-grams and 4-grams words to have more specific expressions concerning the books[1]. Here the managing of stop words was a bit different than only removing them from books. We explain more about this in the faced challenges section 4.1.3.

For each word cloud, we only display between 20 and 50 words and expressions.

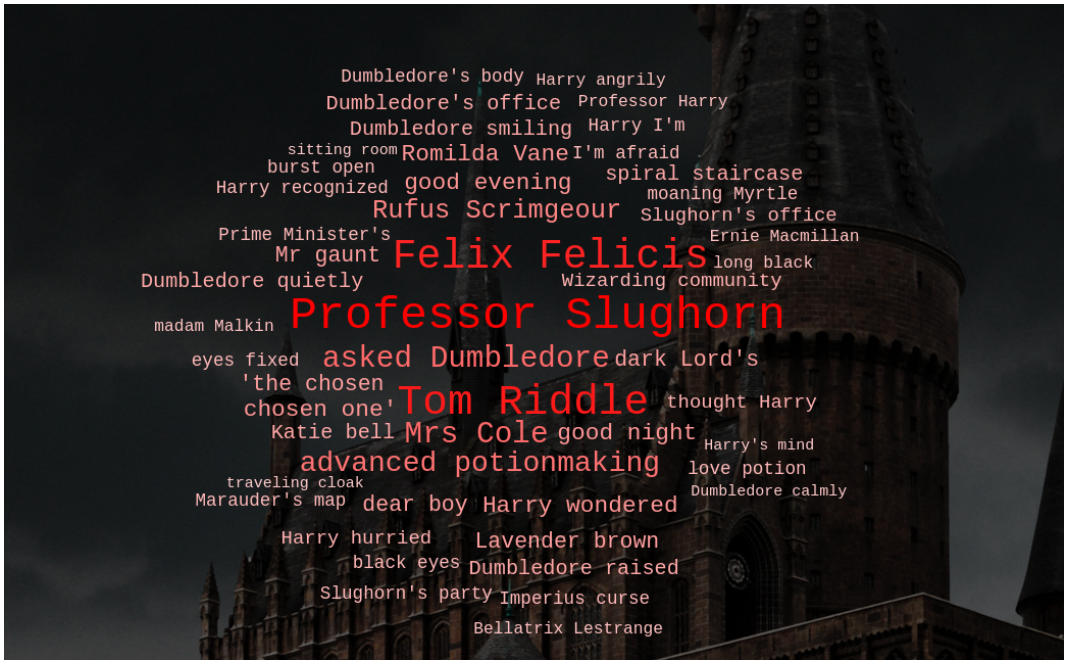

FIGURE 3
Harry Potter and the Half Blood Prince - 2-Grams

3.4 SENTIMENT ANALYSIS

For this visualisation, we use again all the books and the positive/negative words dataset. To produce this, we took each book, separate them in 10 parts of equal length, and for each words of each part, we check if this word is positive or negative (or neutral, meaning that it doesn't appear in our positive/negative words dataset), and add +1 or -1 to a counter. After this, we get seven lists (one per book) of 10 values, those represent a "degree" of emotion. We decide to uniform these values and bring them to a range of -50% to 50% for visualisation purposes.

4 FACED CHALLENGES

In this section we describe the difficulties we encountered during the implementation of our project.

4.1 DATASET PROCESSING

As depicted below, we struggled a bit to produce the different datasets needed to create the visualizations.

4.1.1 CHARACTERS DATASET

One of our original ideas was to plot a genealogy graph of Harry Potter characters, see Figure 4. The main problem was the fact that we did not manage to find a dataset that contains genealogical information. We first started to create one from scratch using Harry Potter's wiki page[2]. Figure 5 shows an insight of the resulting *.csv* file. We then realised this idea was too ambitious since there are approximately 140 characters in the books. After having a discussion with the professor, we decided to drop this idea.

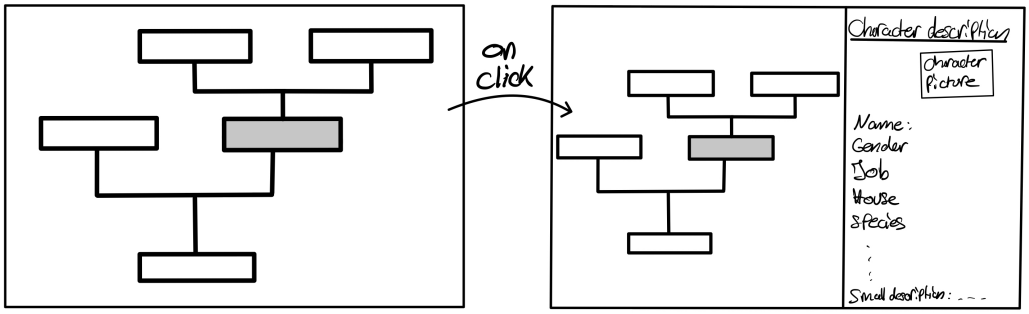

FIGURE 4
Genealogy Graph

Id	mother_id	mother_name	father_id	father_name	Name
1	12	Lily (Evans) J. Potter	13	James Pother	Harry James Potter
2	31	Molly (Prewett) Weasley	32	Arthur Weasley	Ronald Bilius Weasley

FIGURE 5
Linked Genealogical IDs

4.1.2 POTIONS DATASET

One of our very first idea was to create a sankey diagram using the potions dataset, especially columns Name, Ingredient and Difficulty level. But we found at that the ingredient were very specific for each potion and the only column with common value was the Difficulty level one. In any case, many of those data were missing and hence we decide to drop this idea.

4.1.3 WORDCLOUD DATASET

One of this section's challenges was to choose the right number of top words in the set of grouped books to to display the most specific N-gram words for each group. This number needed to be not too small, otherwise we would have some basic name or expression that appear too many in all book and doesn't have a specific role in the particular book, but also not too big, otherwise we will lose the spirit of the book.

Also, another challenge, especially concerning the 2, 3 and 4-grams, was the managing of the stopping words, which is different for the 1-gram words where we just needed to delete them from our list. For example, we didn't want to keep the expression of the form "*person shout*" or "*profesor X said*", ... But we wanted to keep expression like "The chamber of secret" or "sword of Gryffindor" which both contain stopping words. So we decide that for the 2-grams words, if both words are stopping words, we remove them. For 3-grams, if there was 2 or more stopping words and that the last words of the expression was the word "*said*" (since those kind of expressions where the ones that appear at the top in many books), we don't take it into account. Lastly, for the 4-grams, the condition was : if there are more than 3 stop words and the last words is a stopping words, we don't take it into account.

4.2 AMCHARTS LIBRARY

For all our visualization, we used the *amCharts5 library* which offers easy declaration of many types of charts. Nonetheless, even if creating charts is simplified, the amount of data preparation required to match the data structures taken by amCharts is huge and sometimes

tedious. Also, these data structures are formatted in a way which is not friendly to give additional data for more information (i.e. tooltips).

Our data pre-processing turned out to be incompatible with our target charts which made our task much time-consuming.

5 Separation of Work

We depicted below who was responsible of the principals tasks.

5.1 Antoine Daeniker

- Spells bubble chart: data processing
- Spell graph: data processing
- WordCloud: data processing, visualisation

5.2 David Desboeufs

- Genealogy graph: data processing
- WordCloud: buttons visual and logic
- Website layout organisation

5.3 Jérémie Frei

- Spells bubble chart: visualisation
- Sentiment Analysis: data processing, visualisation

6 Conclusion

This project allows us to create a visualization of our favorite saga. We also had the opportunity to explore different aspects of the books that we never thought of.
Since most of work consist of words analysis we thought it is unavoidable to cite the quote below.

> Words are, in my not-so-humble opinion, our most inexhaustible source of magic. Capable of both inflicting injury, and remedying it.
>
> - Albus Dumbledore

REFERENCES

[1] Wikipedia contributors. *N-gram — Wikipedia, The Free Encyclopedia*. [Online; accessed 3-June-2022]. 2022. URL: https://en.wikipedia.org/w/index.php?title=N-gram&oldid=1073019765.

[2] Fandom. *Wiki Harry Potter*. 2022. URL: https://harrypotter.fandom.com/fr/wiki/Wiki_Harry_Potter (visited on 3rd June 2022).