Harry Potter is a series of seven fantasy novels written by British author J. K. Rowling. The series was originally published in English by Bloomsbury in the United Kingdom.
Since the release of the first novel, the books have garnered immense popularity, positive reviews, and commercial success worldwide. They have attracted a wide audience of both young and adult readers and are often considered cornerstones of modern young adult literature. The original seven books were adapted into an eight-part film series by Warner Bros. Pictures.
The success of the books and films has allowed the Harry Potter franchise to expand with numerous derivative works, a play named Harry Potter and the Cursed Child, a studio tour in London, a digital platform on which J. K. Rowling updates the series with new information and insights, and a pentalogy of spin-off films premiering in November 2016 with Fantastic Beasts and Where to Find Them.

Development of the visualization

We managed to agree on the Harry Potter theme rather early in the semester. With such a large theme, the number of possible visualizations increased quite fast. After some research and brainstorms, we managed to reduce the number of visualizations we wanted to make to 6 + 2 optional and here comes the datasets hunt. After having seen what Google suggested, we agreed to use "only" 3 different datasets based on characters, character information and movies. We also included the book textual content in our data.

With the data in hand, we started the exploration and cleaning phase which turned out to not be as easy as expected (we will explain this later on, in the challenges section). After having standardized and merged the datasets, we focus on the story we wanted to tell, or more exactly in our case, how we should organize the visualization to have a good pace while providing a good amount of information to the user. That's where we agree on the main lines of our visualizations: general information about the wizarding world, the social section (with analysis of the houses and graph of characters), the wizarding world in itself (spells and wands) and finally a deeper analysis about book and movie text.

With all this in mind, we started to provide sketches of the different visualizations we would like to implement. We agree on the style of the website: a "slide" design. This way, each part of the visualization can be separated from the rest, to avoid having a cluttered and dense website, while providing a

smooth scrolling experience. Overall, the implementation went smoothly and we closely followed our initial idea. The last mile was quite intensive as we had to merge everything together while improving the website visual appearance.

Challenges

We have faced some challenges since milestone 1. We were lucky and we found a lot of data available in the Harry Potter theme. This lets us exploit our imagination and be creative as to show a lot of different things. However, the number of cool datasets found comes at the price of the complexity to merge them. As an example, we found one dataset describing all the characters in the series and another which is a graph of interaction scrapped on the Harry Potter wiki. Merging these two in a single one was absolutely non-trivial as some of the names were different, e.g. some of the dataset wrote "Lord Voldemort" and others "Tom Riddle" or "Tom Riddle (Voldemort)". This complexity required solving each of the issues by hand.. Similarly, some of the character picture links were outdated in the found dataset which required to provide each one of them with a new link by hand.

Another challenge was to limit ourselves to a subset of data. The best example is the number of characters, from the first dataset we got more than 700 characters which is really too large to be displayed on a graph or on the Hogwarts house visualization. We had to choose how to limit ourselves to a subset of characters that are interesting to show.

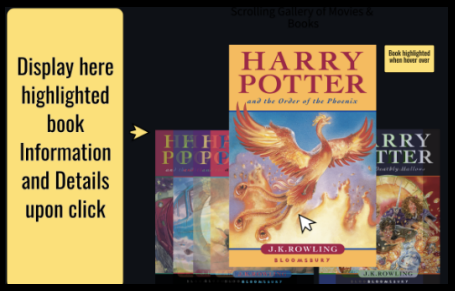
The final challenges were technical ones. Making different parts of one visualization worked as one was really challenging and the difficult to find documentation of SIgmaJS made the affiliation graph visualization really hard to implement. Another challenge as we previously mentioned was the merging of our works. The visualizations were distinct enough that we could separate the work, with each of us working on his own parts. A lot of them were done using D3.js but it turned out that we were sometimes using different versions of D3 (e.g. the one from the exercise sessions may not be the same updated version found on the web). This wasn't easy to notice at first and some of the visualizations were not working because of that. Finally we figured out the problem, used the updated version of D3, updated our code and we managed to make it work. The only downside to this is that it took some time that was allocated to code cleaning so a few files might still need some work, but otherwise the rest of the code is clean and well organized.

Finally, one of the remaining issues and the most annoying one, is that due to the size of the data we have to show, the behavior of the website on smaller devices and maybe other browsers is not guaranteed to be as expected. We test our implementation on both a MacBook Pro using Google Chrome and a Lenovo Legion using Firefox. The best user experience was to use the website in fullscreen mode.

What did we change from the original plan ?

Generally speaking, the different visualizations we implemented closely followed the sketches we made for milestone 2. The main structure is unchanged but we may have diverged from the original sketch for some visualizations due to technical issues. Additionally, we did have the time to implement the two more advanced visualizations we proposed. This provides a nice and complete website to fully dive into the Harry Potter world.

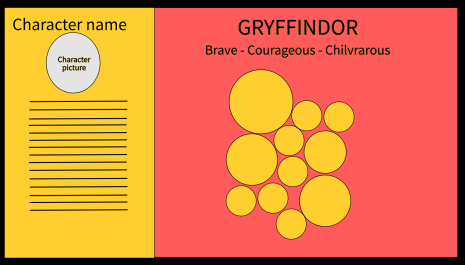
Movies and books description



As an introduction, we wanted to provide a fair amount of information regarding the Harry Potter series globally. As explained in milestone 2, we implemented a visualization allowing us to display information when hovering the book or movie cover.
In contrast to what was planned, we didn't use a fish-eye effect. While it was a nice effect, the problem with fisheye is that it sees the whole gallery of books as one instance, and cannot pinpoint which book we're hovering over, so we can display its info. We managed to implement this visualization using only HTML, CSS and JS and no other libraries. In the final version, we simply hover over a book or movie cover to be able to display relevant information related to it.

Hogwarts houses

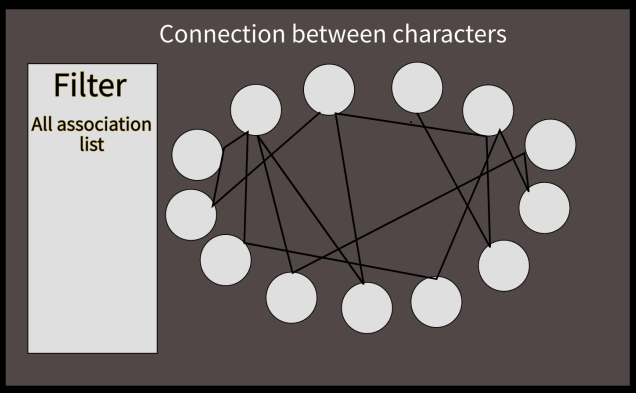
Hogwarts houses are such a central subject in Harry Potter that providing a way of exploring the 4 houses was one of our main goals from the start.



Compared to the sketch visualization, the implemented one doesn't diverge too much. We indeed manage to build a house selection menu redirecting to all the characters sorted in that house. Furthermore, the bubble plot was smoothly implemented using D3. The only thing that changed from the sketch is the fact that we changed our mind from the side bar information to a built-in modal component by Bootstrap that provides us directly all the handling and display that we need.

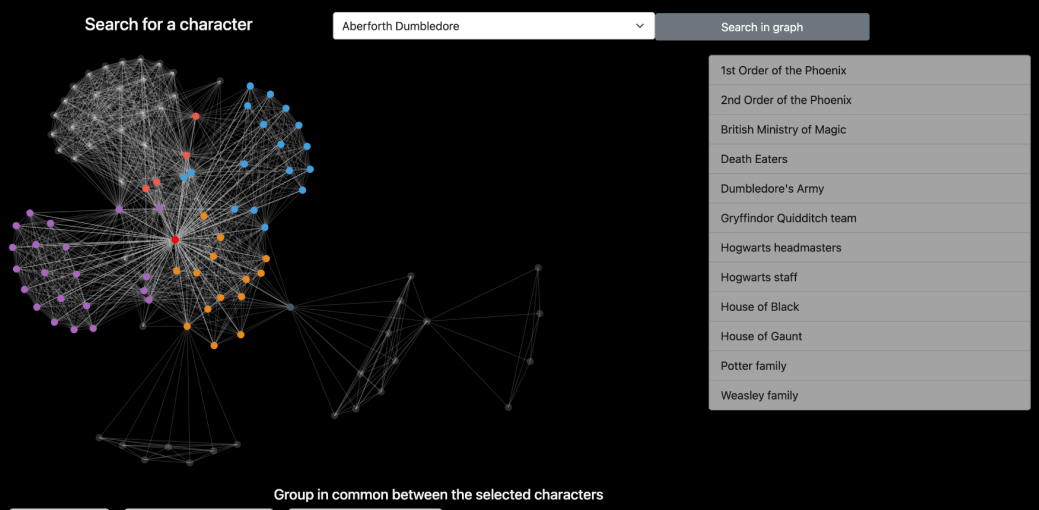
Overall, we think that this visualization provides a cool, simple and good-looking way of exploring all the Harry Potter houses.

Graph



This was by far the visualization we underestimated the most. The first idea was to display a graph with the different groups of people in the Harry Potter world, with filtering possibilities, as explained in the milestone 2.

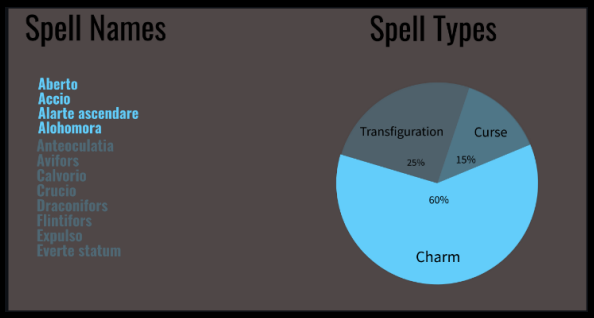
The complexity comes from the low-level library that we used, SigmaJS, which takes us a long time to figure out how we should use it the correct way. Another limitation was that initially, we planned to use all the associations of people available in Harry Potter. However, this would have included Hogwarts and the 4 houses. Including this in the graph makes it unreadable due to the huge amount of edges added.

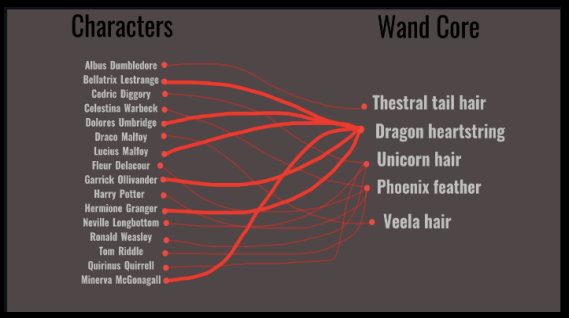


That is why we arbitrarily chose 12 associations to be shown on the website (as shown above). A final complexity comes from the interaction of the different components, i.e. the filter should interact with the graph, the search bar should also interact with the graph and the graph should interact with the "Common association" section. These interactions add a non-expected amount of complexity and it takes us some time to understand and determine the easiest and cleanest way of doing this.

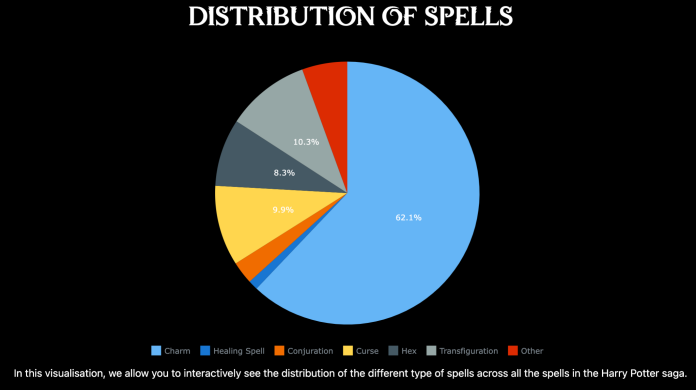
The last touch we added that was not planned at all, is the search bar. Without it, it would have been really hard to know where each character is represented on the graph and thus providing a simple search feature was a plus for this visualization. The nodes have been color coded by the community obtained from running the Louvain method on the graph of interaction.

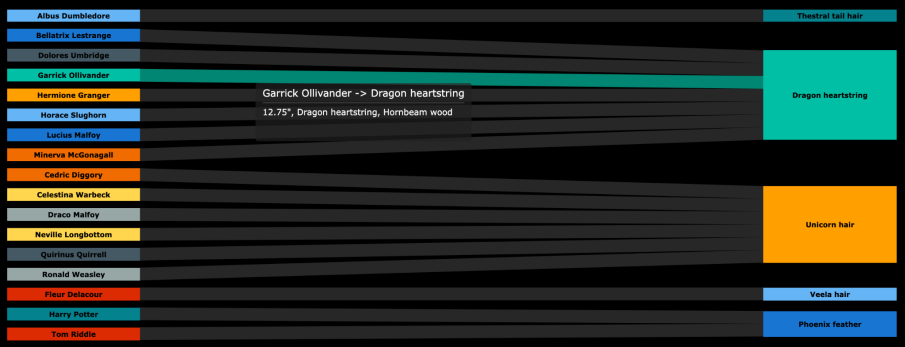
Wands and spells





Each wand of the wizarding world is unique and contains one powerful magical substance. The objective was to use the small number of wands we know entirely (length , core and wood) to show cool visualizations. The plan was to display one Sankey diagram to map the wand to their core component and additionally to make a pie chart corresponding to the distributions of spell types used throughout the books. These simple and good looking visualizations would have provided us good insights about the magical world. The Sankey turns out to be much more complicated than expected. Emna managed to find a working implementation which turned out to be incompatible with D3 and therefore a large number of our other visualizations. Alessio eventually found an alternative that was compatible with the other viz, still good-looking and allowing us to display cool hover data.

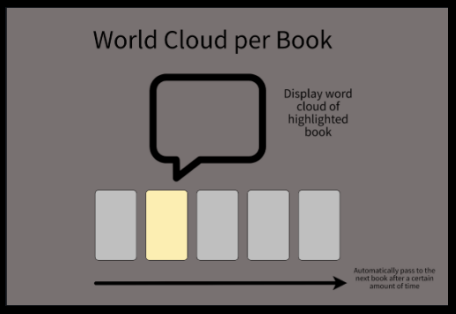




Word clouds

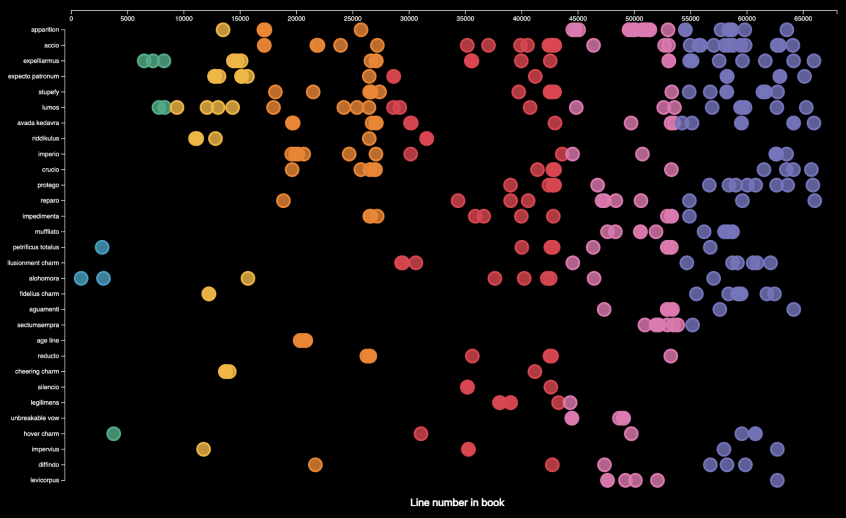
This visualization was smoothly implemented and followed exactly the sketched plan made for milestone 2.

We use a TF-IDF based method to extract the words and n-grams that are the most important for each of the books. Using the extracted words, coupled with D3 word cloud layout library, we manage to implement this first visualization diving us into Harry Potter textbooks. The colorful results provide insights about the content of each book and we can see that this simple information extraction technique is working and provides meaningful results.



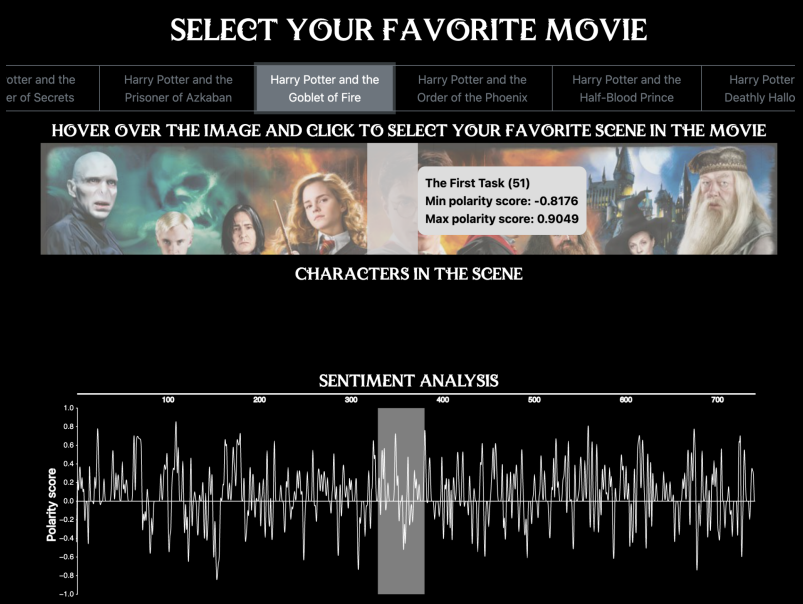
Spells launched by books

Since this was a supplementary visualization, we didn't make any sketch for the second milestone (neither for the second supplementary visualization).



The objective of this viz is to provide a cool and entertaining way of knowing when each spell has been launched throughout the book series. Each circle represents a spell launched in a given line of the book (x-axis) and a line of circle corresponds to each time the same spell has been launched throughout the book series. We color code the book of the saga to have a more meaningful representation of the timeline. Additionally, hovering over one of the circles gives the ability to see the context in which the spell was cast.

Sentiment analysis of movie script



This was the second supplementary visualization and by far the most difficult one. Again, we don't have any sketch since it was a supplementary visualization. The goal was to provide a graphical way of exploring a sentiment analysis of the movie screenplays. The high level of complexity comes from the fact that we have several components that need to interact with each other. In that order, we can select the movie, hover over the chapters which cause the corresponding part of the polarity score plot to be highlighted while displaying some information about the chapters. Additionally, the amount of data we have to present here is really high and we had a hard time figuring out what was the best way of doing it.

Despite its complexity, we are proud of this plot as it allows a deep and complete exploration of the sentiment over all the movies.

Contributions

Milestone 1

For the first milestone, we all participated to find an idea for the project and the datasets. Alessio did the exploratory data analysis notebook. Stephane and Emna cleaned it.

Milestone 2

For the second milestone, we all looked for ideas of visualizations. After putting everything in common, we split the visualization between us :

- Stephane will do the word clouds, the books and movies detailed visualizations and the introduction slide with the key figures.
- Emna will do the Sankey diagram for the favorite characters to their wand types. Additionally, she will work on the pie chart showing the percentage of spells of each type.
- Alessio will work on the Hogwarts houses and the characters inside them. Additionally, he will also work on the graph of character association in the wizarding world.

Each one of us worked on sketching their attributed visualizations and also on the final deliverable for this milestone.
Alessio also worked on extracting different features from the text data and merging the different datasets into usable ones, as well as creating the basic web server structure. Stephane worked on extracting a list of words for the word cloud visualization. This list of words will contain the words that are present in a particular book and not in the others, based on TF-IDF techniques.

Milestone 3

For the third milestone, everyone implemented his/her attributed visualizations. Alessio did the two optional and more advanced visualizations (namely sentiment analysis and the spells timeline). In addition to the technical implementation, Alessio worked on the process book with Stephane. After being done with his part, Stephane worked on beautifying the website and fixing some layout issues as well as recording the screencast. Finally, Alessio implemented Emna's visualization after the issues she encountered.

General assessment

Overall, we managed to split the work based on people's strengths. We were often in touch using Telegram and we met around twice a month to update the other on the advancement of the project. We finally worked hard during the last week of the semester to clean the visualizations, the repository and finish everything.