

Potter Plots

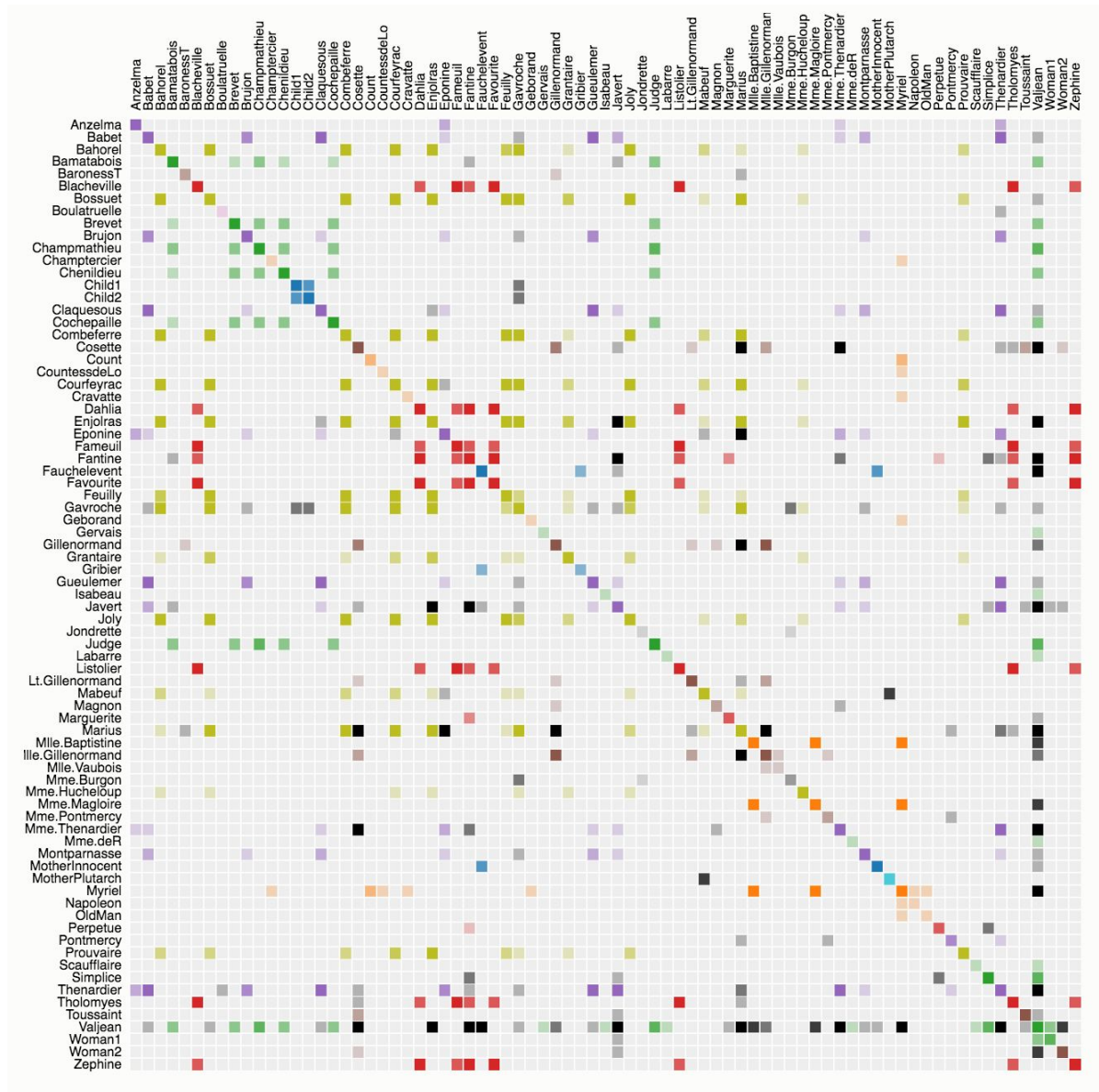
Overview and Motivation

During class, Alex showed us a visualization that plotted the interactions of *Les Miserables* characters that clearly showed the different groups throughout the book. Jessica was very intrigued by this visualization. Being a huge *Harry Potter* fan, she wanted to do something like the *Les Miserables* visualization for *Harry Potter*. She thought it would be interesting to graphically see aspects of the series that people, including herself, have fallen in love with.

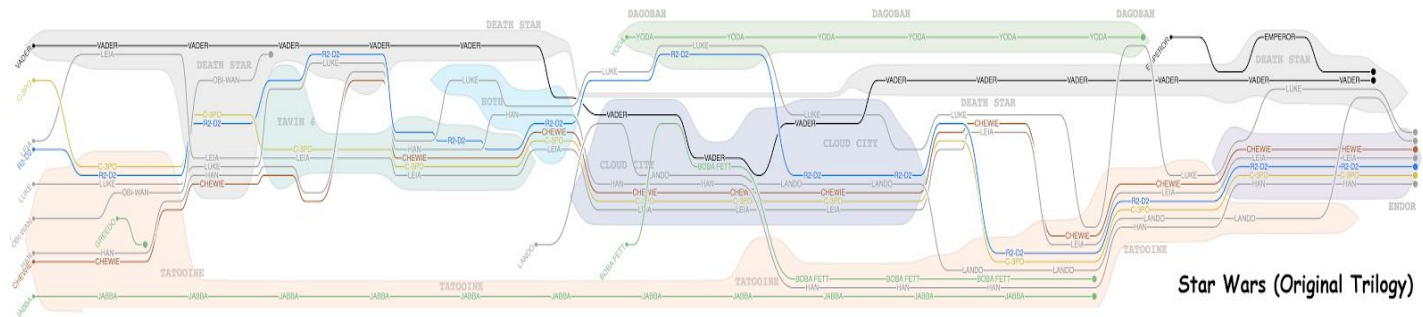
With this project, we plan on showcasing integral parts of the series like spells and deaths with other pieces of information like word counts and interactions. Our project is geared towards fan of the *Harry Potter* books in particular. We chose this focus group because many characters in the books do not appear in the movies and more details about spells and interactions could be easily gathered. We hope that our visualizations can lead users to examine the series more thoroughly and understand certain pieces more clearly.

Related Work

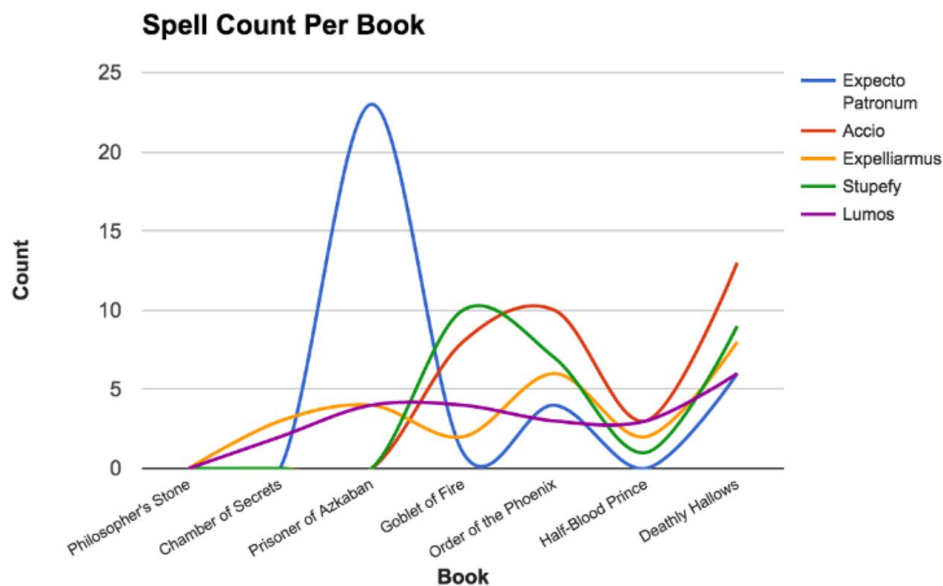
Our biggest inspiration for this project, as mentioned above, was Mike Bostock's *Les Miserables* Co-occurrence chart, shown below. The author mapped the interactions of all of the characters in Victor Hugo's book and displays them in a co-occurrence chart that can be sorted by character name, frequency of interactions, and by character clusters. This way of displaying the data really emphasizes important characters and helps users see trends. We thought something like this would be really helpful for the *Harry Potter* series since there are so many characters that appear across the different books.



When Jessica posted on the discussion forum looking for a partner, Alex posted a link to a [visualization](#) that more clearly showed the sequence of events, character interactions, and locations. An example of the output of the program developed by members of the ViDI Research Group at the University of California, Davis can be seen below. We were very excited by this prospect and really wanted to incorporate its design into our own implementation. We thought it would more clearly show the progress of the story and the characters involved.



The visualization that related most to *Harry Potter* already that spurred our imagination for this project was Ian Minoso's [textual analysis of Harry Potter](#). Minoso created simple charts for word count, location count, spell count, curses, word dispersion, sentiment analysis, grammar counting, and relationships in each book. Below is an example of his spell count graph. While these graphs contain good information, they were not interactive and more information could not be found. We thought we could build off of this template and allow interactions and more information to be gained for users.



Questions

When we started our project, we had a handful of questions we wanted to answer. These questions were:

When looking at all of the books together:

- How were the number of words dispersed throughout the series?
- How many spells were cast in each book?

- How many people died in each book?
- How often were locations mentioned?
- How did the gross revenue of each book compare to its respective movie?

When look at each individual book:

- How were the number of words dispersed throughout each chapter in a book?
- How many spells were cast in each chapter?
- How many deaths occurred in each chapter?
- What locations were mentioned in each chapter?
- How did the characters interact in each chapter?

As we started gathering our data, we realized that some of these questions were either too vague or would require too much data collection on our part. Our questions changed to be:

When looking at all of the books together:

- How were the number of words dispersed throughout the series?
- How did the number of spells cast by major characters change over the course of the series?
- How many people died throughout the series, who killed them, and what was the major cause of death?
- How did the gross revenue of each book compare to its respective movie?

When look at each individual book:

- How were the number of words dispersed throughout each chapter in a book?
- How many spells were cast in the entire book, and what distribution of those spells were cast by major characters?
- How many people died throughout the book, who killed them, and what was the major cause of death?
- How many chapters were two characters mentioned in?

Data

We were able to find txt versions of all of the books from glozman.com. Based on the data that we needed to find, we either searched through these files for key information or used scripts to extract the data for us.

We used JSON to structure our data. We currently have three different JSON files that store our data.

books.json

This file has a majority of our data. Each book has its own object, which has the book and movie grosses, individual chapters with their word count, name, the color of the book, and the spells.

To determine book and movie grosses, we used information from the-numbers.com to determine the worldwide box office amounts in millions. Determining the gross for each book was a little more difficult. We were able to find the [number of copies of each book](#) and the [total gross for all book sales](#). To get the gross for each book, we took the total gross and divided it by the total number of copies sold. This gave us an average of \$17 a copy. We were then able to multiply the number of copies for each book by \$17 to find out the total gross for each book.

For the color each book, we took an [image of the covers of the paperbacks](#) and took the hex codes from the covers for individual book colors.

The spell data required a little more work. [Wikipedia](#) has an article that lists all of the spells in the series and in what book they are used. Once we could find what spells were used in the books, we searched the text files of the books for instances of the spells. A JSON object could then be created based on the name of spell, who cast it, and then how many times they cast it.

deaths.json

The Harry Potter Wikipedia has a [comprehensive list](#) of all of the deaths in the series, who killed them, and the specific cause of death. We were able to take all of this information and create JSON objects for each one.

interactions.json

To get character interactions, it would be way to difficult to go through the text and find each one. Instead, we created a PHP script that split each text file into the individual chapters. From there the script ignores any words found in quotes, since characters could mention other characters that weren't actually in the same chapter. Once each chapter has been parsed, the number of times two characters appear in the same chapter are totalled for the end amount. This does lead to errors, especially since Harry's thoughts are not in quotes. For example, Aunt Petunia and the Basilisk appear

in the same chapter together given the script, but obviously those two don't actually appear in the same chapter and interact with each other.

Locations.json

Initially, Srivatsan tried to put together a list of locations that are mentioned in the book from the list of provided in the Harry Potter Wikipedia but this list turned out to be inadequate as many locations are simply not mentioned. In order to make the list more manageable, a decision was made to include a certain number of select locations alone in the graph. Srivatsan tried to put together this list through by generating the number of times these locations were mentioned but it was not possible to differentiate between the instances where the locations are mentioned and where the characters currently are. So the locations instead were put together manually based on where the characters move through each chapter

Battledata.json

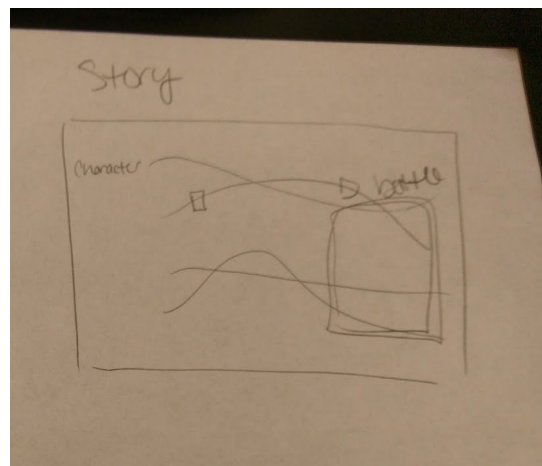
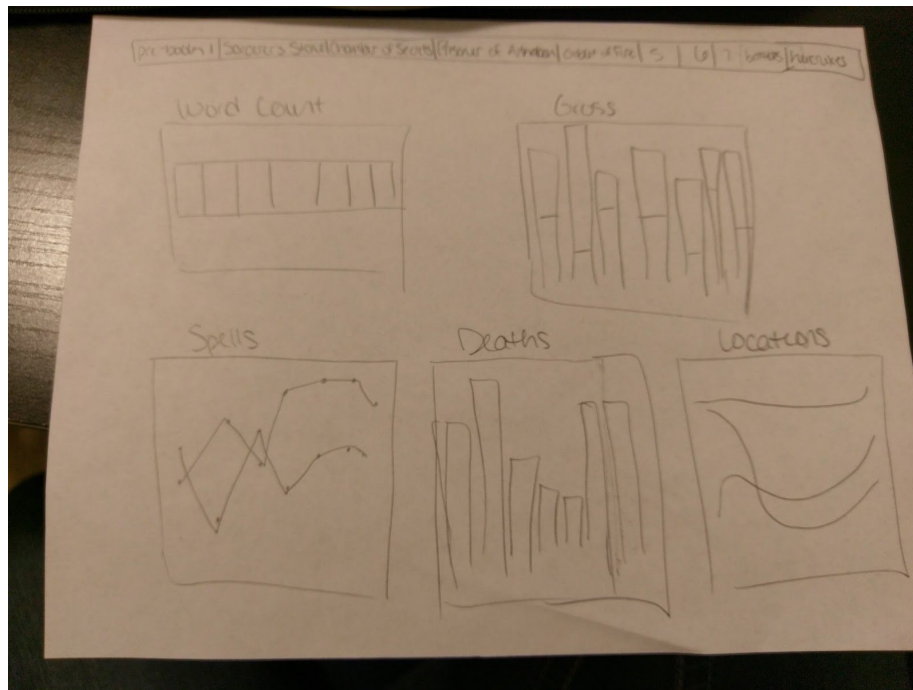
The battledata.json is put together manually for three battles that happen throughout the book. The battle is described as a series of events. Events can be of five types: Entry, Exit, Spell, MSPELL and Special. Entry events describe the entry of a character into the battle, the side which the character fights for and any modifications based on if the character is dead or physically absent. Exit events describe the exit of a character from the battle and the reason like death, running away or incapacitation. Spell events describe a spell being cast and taking effect. It mentions the caster and the target on whom the spell's effects take place. MSPELL events describe a spell whose effects are nullified or if the spell does not hit the intended target. It mentions the caster and the intended target. Special events are used to describe special magical events like Priori Incantatem and possession.

Exploratory Data Analysis

Seeing as most of this data had to be put together ourselves, there were not many ways we could look at our data before we started. The only thing that gave us an idea of what our data would look like was Minoso's textual analysis web site. His results helped us have a base point for our own data. For example, the Wikipedia page of spells did not always list all the times *Lumos* was used in every book. We were able to compare Minoso's results to our own to see if we were missing some of the key spells he looked at in his analysis.

Design Evolution

Below is our final design from our project proposal.

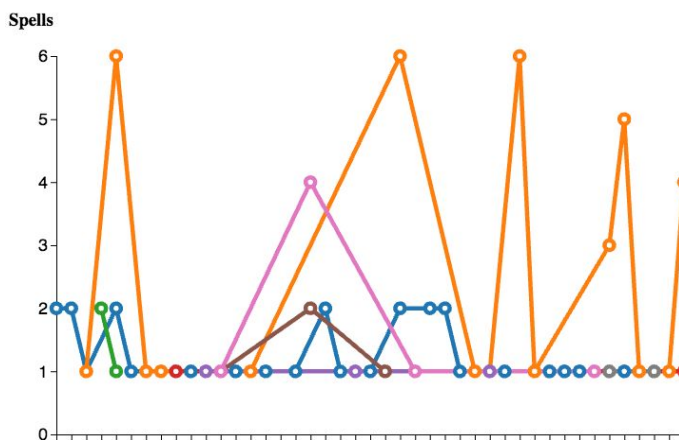


Across the top of the page, a navigation bar would allow users to toggle between the all book view or a view looking at the data from an individual book. When a tab has been selected, the colors of the text and background reverse and the font gets bigger, showing that that tab is the one currently being looked at.

We decided on using a stacked bar chart for the word counts. This way, each rectangle, representing a chapter or book, would be proportional to its actual length in the book or series respectively.

The total gross chart was also a stacked bar chart. We felt that this would be easier for users to compare the total gross of book and movie, with some comparison also available between books and movies separately. We decided to change this instead to individual bars for each piece. This is because we forgot that the seventh book was split into two movies and users may find it easier to compare a book or movie using this layout instead of comparing bars that started at different heights.

Our design for the spells changed throughout our project. At first, we had a bar chart that just showed how many spells were said in each book or how many times each spell was used in a particular book. At this point, we had a lot of bar charts and Carolina suggested changing the chart to a line chart just for some variety. This was implemented (in the `spell_by_caster` branch). To do so, the data needed to be changed from only spell counts to who cast the spell and how many times. This way there could be a y-axis of number cast, x-axis of either book or spell, and each line would represent a particular person. This, however, did not work as well as we thought it would.



To the left is how this implementation looked for the seventh book. While there are some interesting pieces to this, there are many characters that only cast one of each spell, causing the line of different colored dots at the horizontal 1.

For clarity, we think that showing individual Aster plots for each spell, with each wedge being a person

casting the spell and how much, would work much better for the individual books. A line chart could still work for the view for all of the books, since that it done slightly differently and is not as cluttered as the individual books.

When we started, we were not sure exactly how to show the deaths throughout the series. We thought that a bar chart could work for the all book view with the number of people who died per book. This would fall apart once it came to the individual books

views. To give our graphs some more interactive components, we decided to implement a sunburst chart that would show the cause of death, the killer, and the victim.

Unfortunately, we did not have enough time to implement the location chart or battle views to our satisfaction or in time for the creation of the video showing our product. Anything relating to those pieces can be found in the locations-branch branch.

We started out with the idea of using the story flow diagram from UC Davis to show the events of each chapter. However, as we began to compile the data for this, it became clear that we would not be able to get this data for each book in the specified time. We instead decided to switch to the co-occurrence chart. We felt that this visualization was interactive enough to cause a viewer to investigate many pieces in this chart. It also more accurately shows how often two characters interact without having to page through multiple views and tallying the number up themselves.

Implementation

Navigation Bar



When first coming to our website, the “All” section has been selected. This is shown by the black text and border around that rectangle. To switch to individual book views, a user simply needs to click on the rectangle with the name of the desired book. This will invert the colors of the background and text, increase the font size, and add a border around that rectangle. All data for the visualizations are updated when a new section is selected.



Word Count

Word Count

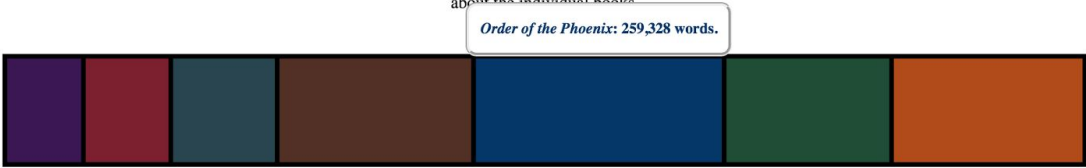
Over the course of the series, 1,117,878 total words were written. Hover over the rectangles below to learn more about the individual books.



The word count view displays a rectangle that is representative of the entire series (as in the figure above) or the entire book (as in two figures down). The text above the chart states how many words the rectangle represents. Hovering over a smaller rectangle shows a tooltip that names the book or chapter with the total number of words in that section.

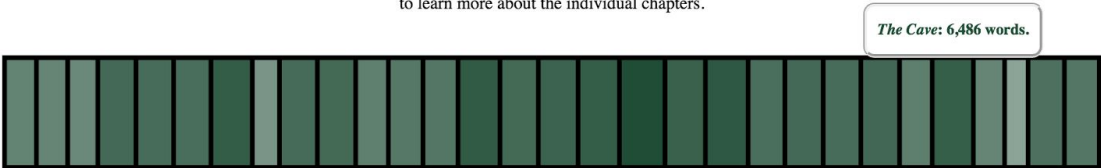
Word Count

Over the course of the series, 1,117,878 total words were written. Hover over the rectangles below to learn more about the individual books.



Word Count

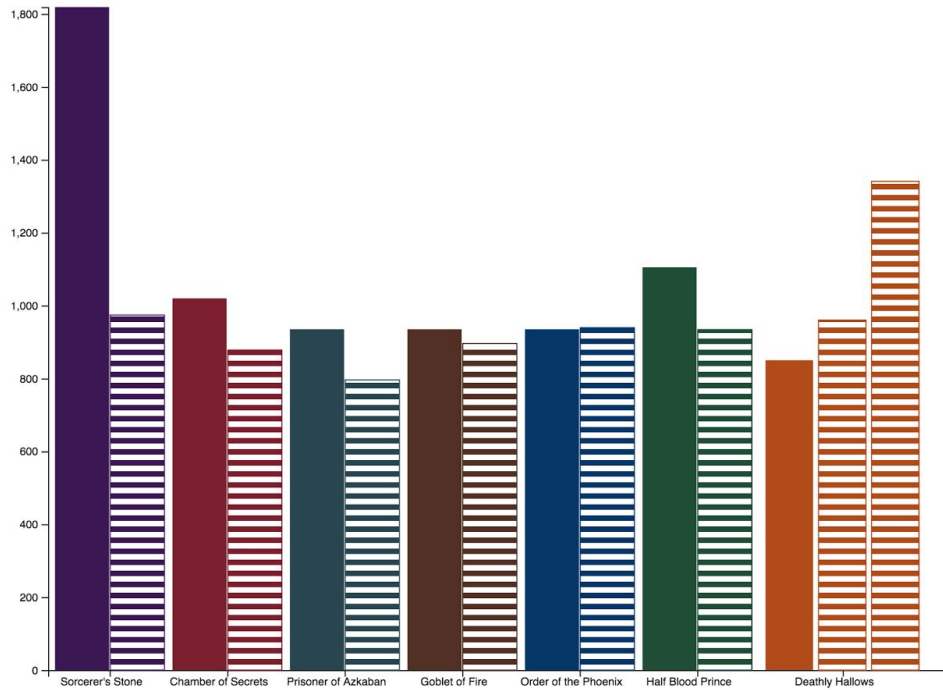
Harry Potter and the Half Blood Prince has 30 chapters totalling to 173,244 words. Hover over the rectangles below to learn more about the individual chapters.



Book vs. Movie Gross

Book vs Movie Gross

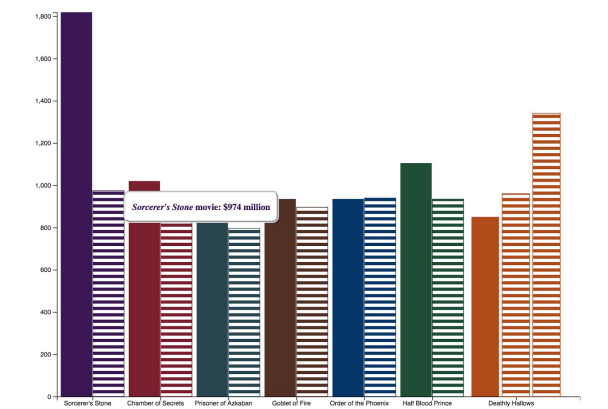
The total grosses of all seven *Harry Potter* books and films are shown. Books are solid colored bars, while movies are striped bars. In total, the books made \$8,449 million and the movies made \$8,678 million. Hover over the bars to find out more information about each book or movie.



The book and movie gross chart shows the gross in millions made by the entire series. The striped texture represents the movies associated with each book, while the flat textures are the books themselves. Hovering over an individual bar shows the name, whether it was a book or movie, and the gross in millions, as shown below.

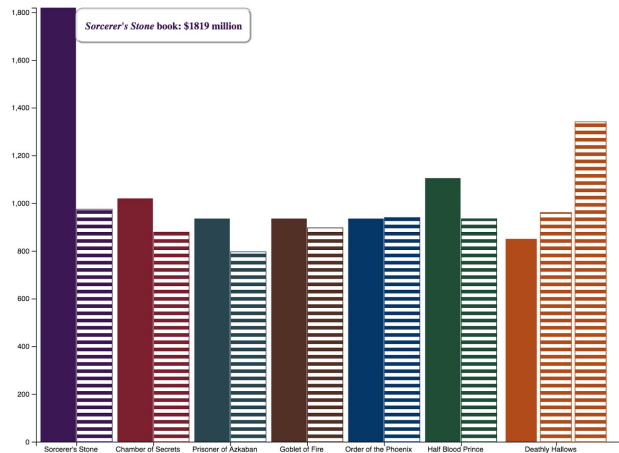
Book vs Movie Gross

The total grosses of all seven *Harry Potter* books and films are shown. Books are solid colored bars, while movies are striped bars. In total, the books made \$8,449 million and the movies made \$8,678 million. Hover over the bars to find out more information about each book or movie.



Book vs Movie Gross

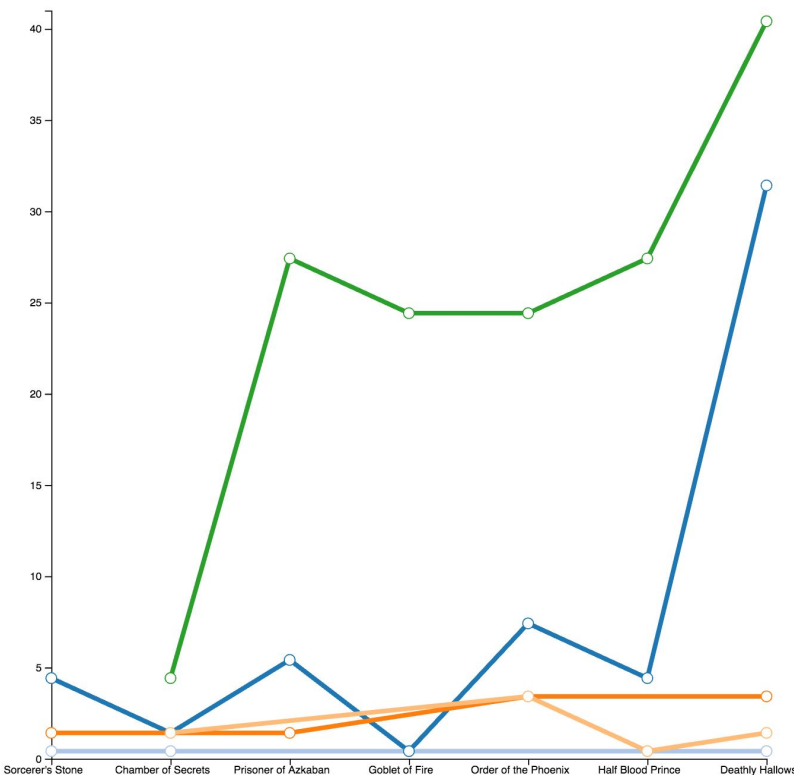
The total grosses of all seven *Harry Potter* books and films are shown. Books are solid colored bars, while movies are striped bars. In total, the books made \$8,449 million and the movies made \$8,678 million. Hover over the bars to find out more information about each book or movie.



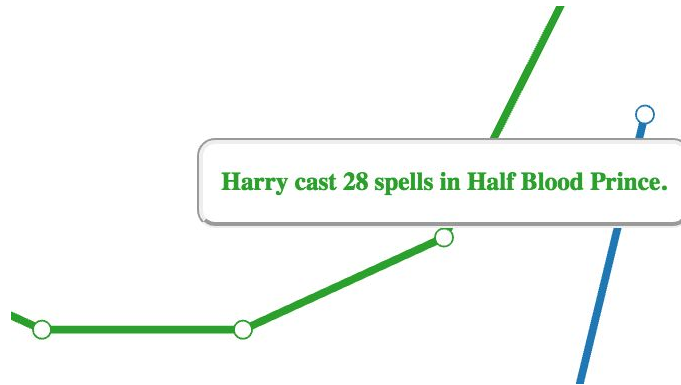
Spells (All series view)

Spells

In all seven books, 360 total spells were cast. Below, the characters who verbally cast spells in more than 3 books are shown. Hover over a circle to learn how many spells were cast and who cast them for each book.



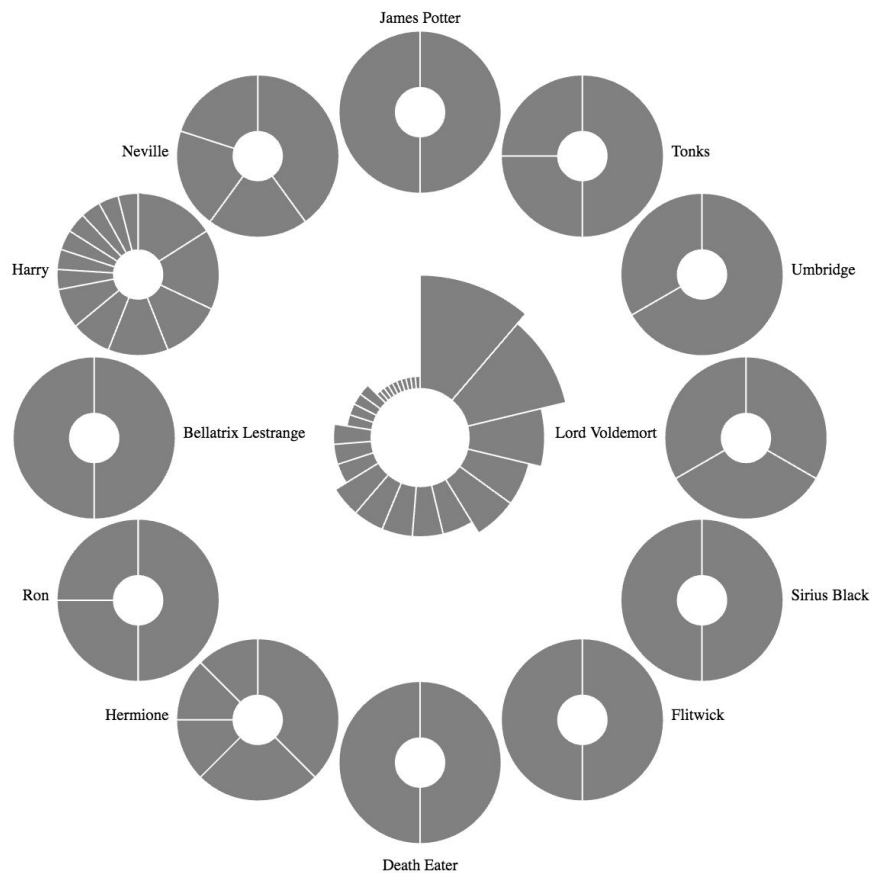
The spell view for the entire book series is displayed as a line chart. Each line represent a character that verbally cast spells in more than 3 of the books. This shows five characters: Harry, Hermione, Ron, Snape, and Hagrid. Hovering over the individual circles shows a tooltip that says the character name and how many spells they cast in each particular book. This can be seen below.



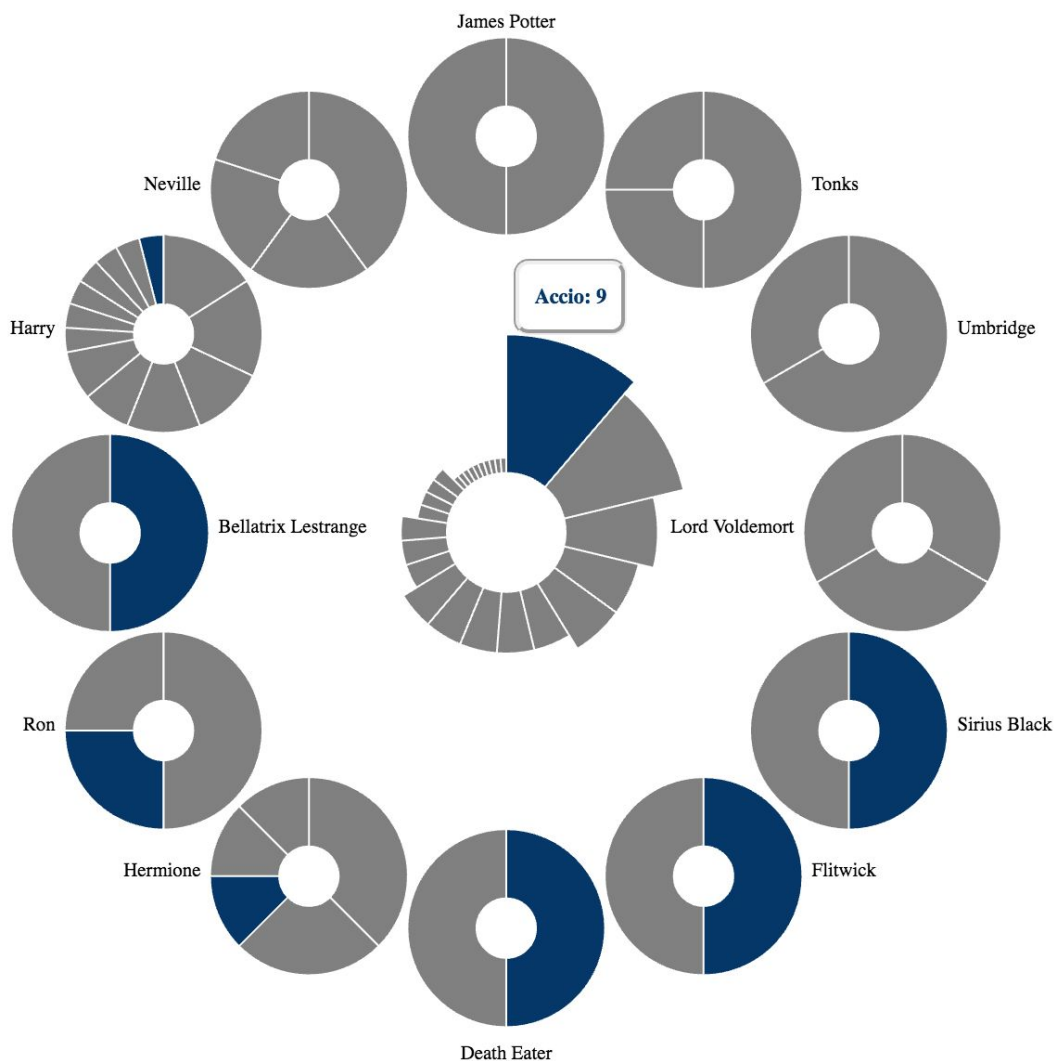
Spells (Individual Book View)

Spells

In *Order of the Phoenix*, 80 total spells were cast. The distribution of spells is shown in the middle, while characters that cast more than 1 spell are shown around the overall spells. Hover over a piece of any of the charts to view the number and corresponding spells on other charts.



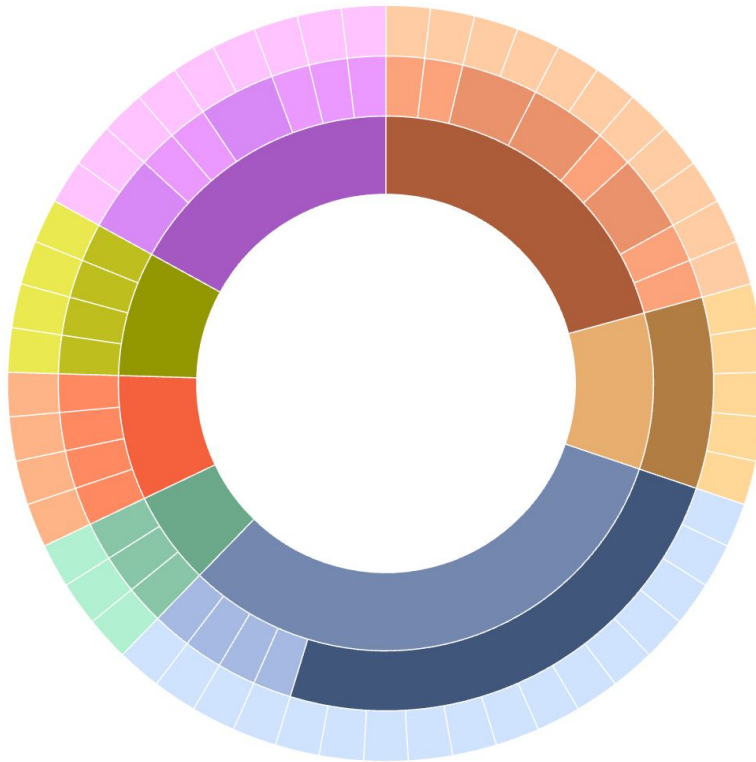
To display all of the spells cast in each book, an aster chart displays wedges that represent the individual spells with sizes based on the number of times that spell was cast. Surrounding that center chart, individual pie charts show each character that verbally cast more than one spell in that book (in *Sorcerer's Stone*, characters that cast more than 0 spells are shown). When a user hovers over a wedge, whether in the center chart or in the character charts, that wedge is highlighted, along any other wedges that represent that spell in any of the other charts. A tooltip shows the name of the spell and the number of times it was cast in relation to the view the user is hovering over. For example, if *Lumos* was cast 6 times in total, with Harry casting it 2 times and Hermione casting it 4 times, if the user hovered over the central chart wedge, the tooltip would say “Lumos:6.” Hovering over the *Lumos* wedge in Harry’s chart would display, “Lumos: 2.” The screenshot below demonstrates this.



Deaths

Deaths

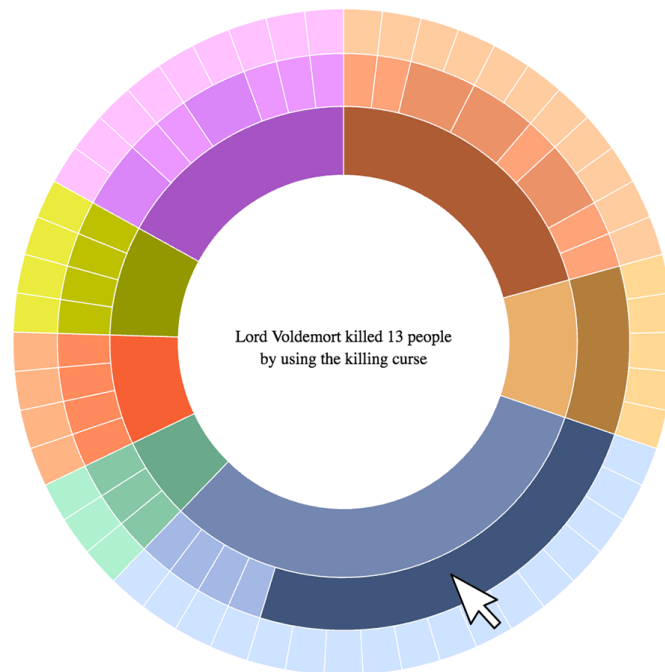
53 total people died over the course of the series across 7 different categories. Hover over the wedge pieces to learn the cause of death and killer.



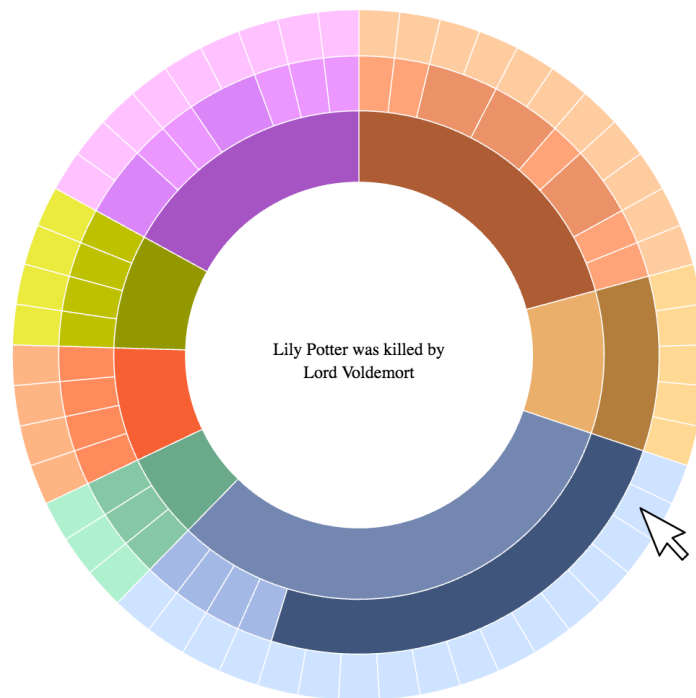
A sunburst chart was used to represent deaths. The chart splits all of the deaths in the entire series or book into seven categories (or as many categories of deaths that occur in that book). These categories include: natural causes, the killing curse, burning, a curse, a wound, unknown causes, or other deaths that did not fit into these categories. The inner wedges each represent one of these categories. Hovering over each wedge says the number of people that died in that category in the middle of the chart, as shown below.



The next layer symbolizes the different killers that fall under that category. The killer may be unknown, the actual name of the person, or, in the case of a natural death, natural causes. The center of the chart displays how many people that killer killed in that category, as shown below.



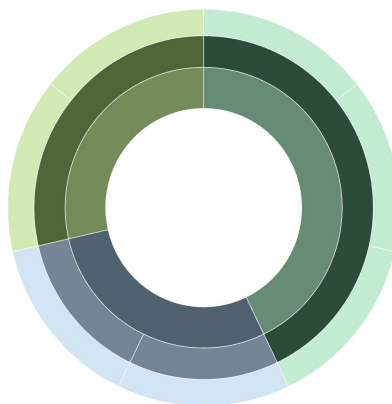
The outermost layer of the chart represents the individual people who died by a certain killer in a certain category. The victim's and killer's names are displayed in the center of the chart as shown below.



Going to an individual book changes this view slightly from what has been shown previously. Only categories of deaths that occur in that particular book are shown. Also, the colors of the chart change to a color scheme that is in line with the color of the book. For example, the chart for *Half Blood Prince* is shown below with its color scheme.

Deaths

7 total people died in *Half Blood Prince* across 3 different categories. Hover over the wedge pieces to learn the cause of death and killer.

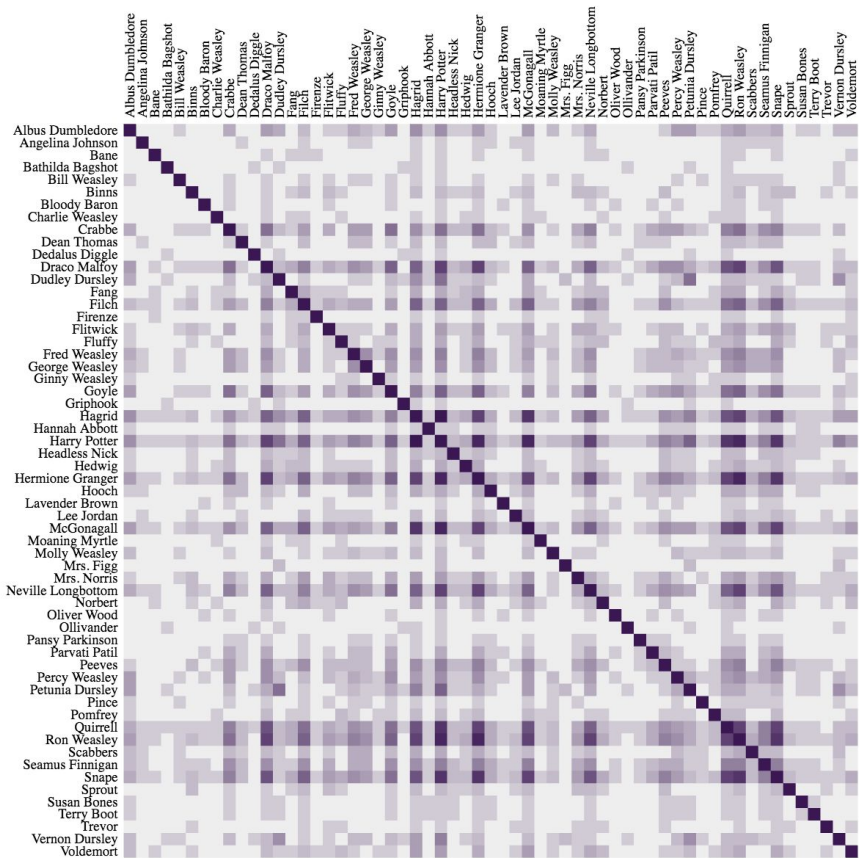


Story

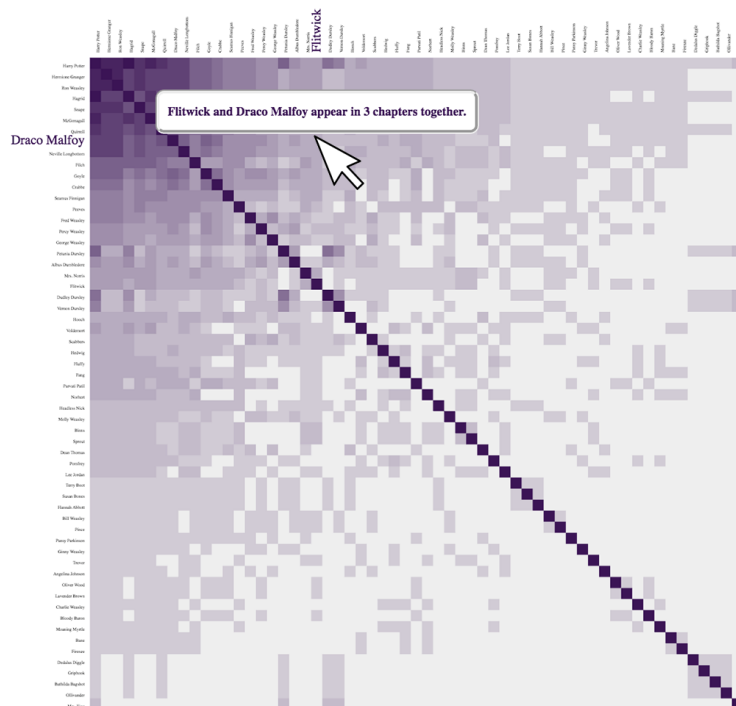
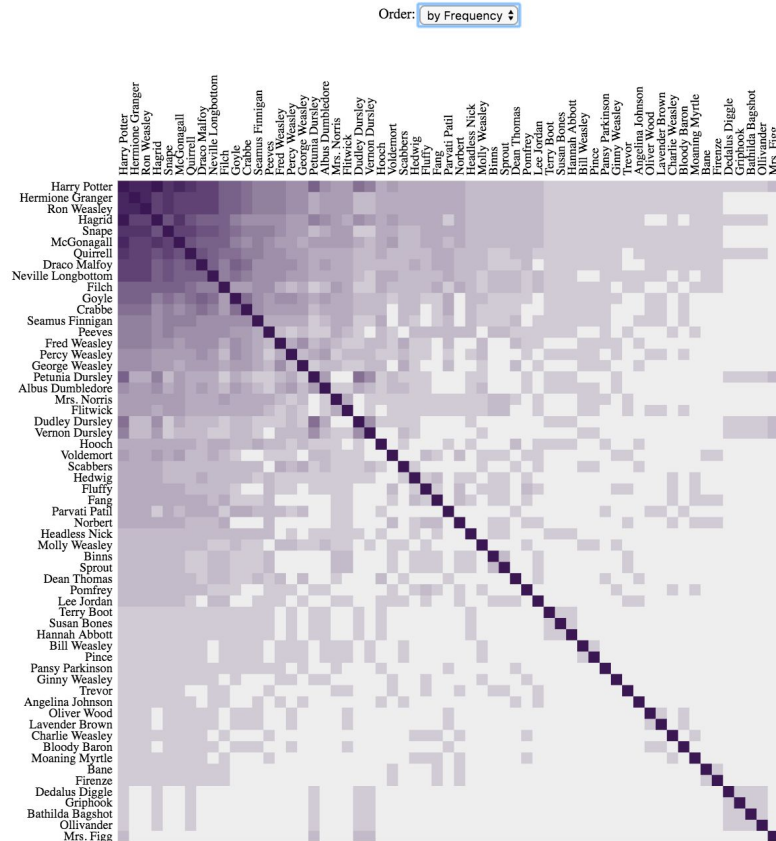
Story

Sorcerer's Stone includes 17 chapters. This co-occurence matrix shows how many chapters two characters are mentioned in together. Hovering over a cell can tell you how many chapters two characters are mentioned in together. You can reorganize the data by name or number of chapters using the dropdown menu.

Order: by Name

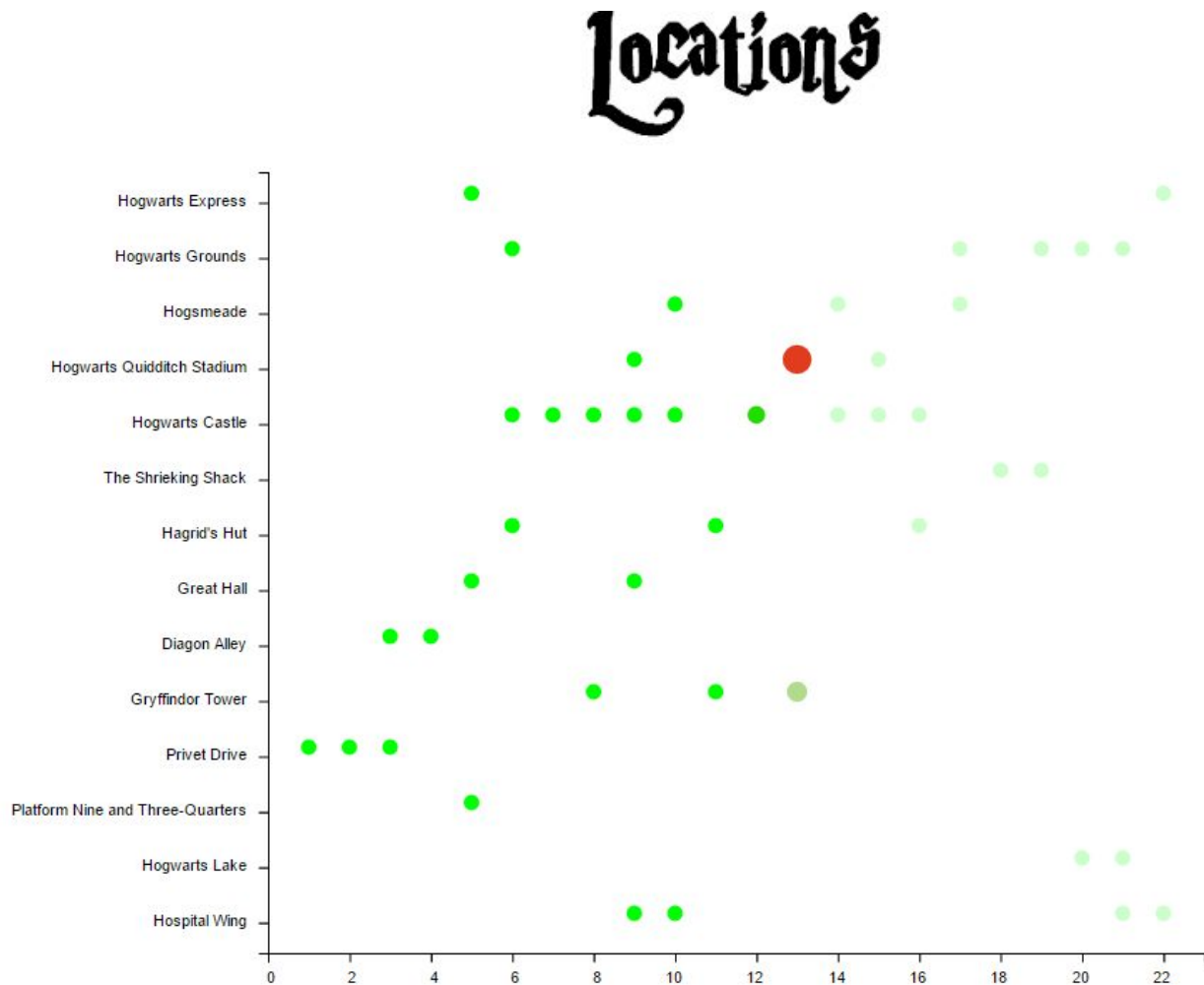


For each book, a story view shows the number of chapters two characters are mentioned in together. The total number of chapters is listed in the text above the chart so users can compare the number characters are mentioned in and the total. The initial set up for this view is to be organized alphabetically by name. This can be changed, using the selection box, to be organized by number of chapters mentioned in. The result from this is shown below.



To learn the specific number of chapters, users can hover over the colored cells. Cells that have no coloring or are in the diagonal do not show a tooltip. All other cells show a tooltip stating that character X and character Y are in Z chapters together. To make it easier to read, the character names on both axes are highlighted. The figure the left shows an example of this.

Locations



The location charts are present in the location branch of the repository and have not been added to the video. The locations chart is a continuous animation where Harry moves through the locations in each book. The locations are indexed by chapter and the current location where Harry is, is denoted by the red circle. The circles that are not fully opaque are the locations that have not been visited yet. The lime circles represent where locations that Harry has visited already. The location chart appears only for individual books and is hidden when all is selected.

Evaluation

We learned a lot about the *Harry Potter* series through our visualizations. Each visualization answers one of our key questions and even gives us more information than we expected.

Our word count chart accurately shows how words were dispersed throughout the series and individual books. We think our chart even adds more meaning to this by using the stacked bar chart since it shows the exact percentage of the whole through the width of each book/chapter. When we added the text above the chart, a whole new understanding of our data occurred. Before, the width was an arbitrary number that wasn't displayed. By adding the total number of words in the series/book, it really made it clear just how much content there was. Seeing that the chart represents over one million words is incredible.

This chart could benefit from a better design. While intuitive for the information, there could be a more eye grabbing way to display this data.

Seeing as we were unable to find the gross sales of each book, we think it is fair to say that our chart comparing the book and movie gross presents entirely new data. The visualization itself is very powerful. The fact that the first book has made more money than any movie (not counting the combined gross of the two *Deathly Hallows* movies) really shows that reading has not dwindled as movies have been produced. In particular, when totally up all of the book grosses and comparing it to the movie grosses, the movies only made 2.6% more than the books did. This is mind blowing. While that 2% is \$200 million, we think it really speaks to how important the book series is.

Bar charts are relatively bland and this chart isn't as eye catching as some of the charts that follow. We could have possibly implemented this as a bi-directional bar chart and it would have shown the total of book and movie combined and allow for easy comparison between books and movies separately.

The two different spells charts give us different representations of the same data, and thus give us totally different information. The all series spell chart shows that only five characters verbally say spells in more than three books. Given how prevalent spells are in the series, one would think there would be more characters that fit this criteria. Hermione is the only character in the entire series that casts a spell in every book. We expected Harry would be the most common, and indeed he casts more spells in every book than everyone else, except in the first book. We feel this really emphasizes Hermione's importance to the series.

In this visualization, it is hard to tell if when two lines cross and there is a circle if the character below actually has a data point there. For example, in *Goblet of Fire*, a circle

marks that Hermione cast one spell, which resides on the line representing Hagrid. It is impossible to tell if Hagrid also cast one spell in this book or not. Representing this data differently, or possibly showing all overlapping circles, could make this data more clear.

The individual book spell charts tell us how often spells occur in each book and who casts them, instead of just a count of all the spells. We think that both of these metrics really give an insight into the content of the book itself. As the series progresses, more and more “evil” characters appear and more and more curses are present. Something that we also noticed is that the most prevalent spell in each book is either something learned in that book or something that is important to the plot. The first four books’ most prevalent spells are spells that were learned and serve an important purpose to the plot. The last three books’ most prevalent spells more give an idea of what is happening in the story. For example, in *Sorcerer’s Stone*, the most prevalent spell is *Wingardium Leviosa*. This spell is learned here, and it also is used when fighting against the troll, which is where the three main characters become friends. This spell also signifies that the characters are learning magic and at this point everything is whimsical and non-threatening. In comparison, the most prevalent spell in *Half Blood Prince* is *Petrificus Totalus*. Used mostly in battles, it shows that the characters need to use spells to protect themselves, the world is no longer so whimsical.

Currently in this view, a tooltip only displays the spell count relative to the chart being hovered over. It would be beneficial to have information on all of the spell counts in the different charts in this one tooltip to accurately compare. The arrangement is also heavily dependent on the number of characters. Right now the maximum number it can handle is 12.

The *Harry Potter* series isn’t exactly known for the high number of deaths that occur, but when seeing that 53 total people died throughout the series, it really points out the importance of the death in the series. Something we found particularly interesting while compiling our data was the number of people who died from burning. It was not something that we considered as a category before looking at the data, but we realized that that category was indeed needed. I think it is also interesting that a third of deaths occur from the killing curse and all of the killers were in league with Lord Voldemort (though Snape’s use of it is debatable in that context).

One thing we hoped to implement was the ability to switch between sorting by cause of death or killer. We unfortunately did not have enough time for this, but we feel like this would add new meaning to our data.

The visualization that inspired this project also gives information about the series. When sorting just by name, the characters that appear with the most characters immediately stand out because of the cross that is formed from their row and column. For example, in *Chamber of Secrets* Draco Malfoy is mentioned in at least one chapter with every character that we compare against. His prevalence in this book, and other books, showcases just how important of a character he is. This visualization also makes it very clear how as the series progressed the stories became more complicated with more and more characters. The first book's chart has 59 characters, while the last one has 106 characters.

The visualization we based this on also sorted its characters by cluster. We did not have enough time to sort all of our characters into clusters but this would give another way for users to compare the data and see correlations. The chart is also limited in the fact that as more characters are added, the smaller each cell gets and the harder it is to distinguish what cell corresponds to whom. While the tooltip helps in this situation, a better visualization may help with this confusion.