Magic is Everywhere

NADIEH

I spent a large portion of my youth reading comic books, especially "Donald Duck" and "Asterix." But at age 11, I picked up Terry Goodkind's *The Wizard's First Rule* from our local library and was instantly hooked on the fantasy genre. I was so grasped by the feeling of disappearing inside these strange and magical worlds that authors created through their words. And this hasn't changed; fantasy is still the only fiction genre that I enjoy reading. So for this topic, I knew that I wanted to focus on fantasy books.

As diving into the words of a book or series itself could potentially be a copyright issue (say, my favorite one, *The Stormlight Archive* by Brandon Sanderson), I looked for a different angle. I've always felt that the titles of fantasy books are somewhat similar: either something about magic (makes sense), or some "name/object" of some "fantasy place," such as *The Spears of Laconia*. Maybe I could dig into the trends and patterns of these titles?

I had to get my hands on a whole bunch of fantasy book titles and scraping the web looked like the fastest way to do this. On Amazon I found a section that showed the top 100 fantasy authors from that day. I wrote a small web scraper function in R with the `rvest` package that automatically scanned through the five pages on Amazon (20 authors per page) and saved their names. However, I couldn't find an easy way to get their most popular books and the Amazon API seemed too much of a hassle to figure out.

Most big company APIs seem like too much hassle to me for that matter...

Luckily, Goodreads has a very nice API. I wrote another small script with help from the `rgoodreads` package to request information about the 10 most popular books per author, along with information about the number of ratings, average rating, and publication date for each of the 100 authors that I had gotten from the Amazon list.

Data Can Be Found in Many Different Ways

While I didn't come across a single public dataset with a bunch of fantasy book titles, I knew that there are other ways of finding data than simply looking for structured CSV or JSON files. Instead I figured out what *was* available: the Amazon author list to get popular fantasy authors and the Goodreads API to retrieve information about those authors. Combining those resources, I was able to create the dataset of fantasy book titles.

Next came the trickiest part; I had to do text mining on the titles, which in this case consisted of text cleaning, replacing words by more general terms, and clustering of similar titles. For the text cleaning I made a few choices. For one, I only kept the authors that had a *median* number of ratings per book that was above 20. Furthermore, I wasn't looking for any omnibus—a collection of several books—or books written by many people. For this (although not perfect) I looked at how often the exact same book title appeared in my list and took out those that appeared more than twice. Furthermore, I removed all books with terms such as "box," "set," or "edition," making sure to manually check all the deletions. Finally, I scrapped books with no ratings. This left me with 862 book titles from 97 different authors.

Now the data was ready for some text cleaning by removing digits, punctuation, and stop words (which are some of the most common words in the language, such as "a," "is," "of," and carry no meaning to interpreting the text). I did a quick word count after the title cleaning to get a sense of what words occur most often in book titles. As these are words, I couldn't resist visualizing the results as a word cloud (see Figure 5.1). The bigger the size of the word, the more often it appears in titles (the location and angle have no meaning). I was very happy to see how often the word "magic" occured!

I also removed some very specific words to this particular dataset of books, such as "Part."

Fig.5.1

The words occurring most often in the 862 fantasy book titles. The bigger a word's size, the more often it occurs.

I wanted to look for trends in these words. However, for a standard text mining algorithm, the words "wizard" and "witch" are as different as "wizard" and "radio," even though we humans understand the relationship between these words. I first tried to automatically get hypernyms of each noun in the titles, but that sadly didn't give me good enough results, the terms weren't general enough or already overgeneralized. I therefore set about doing it manually and replaced all ±800 unique words across all titles by more general terms, such as "name," "magic," "location," and so on.

A hypernym is a word that lies higher in the hierarchy of concepts. Like "fruit," which is a hypernym of a "banana."

Manually Add New Variables to Your Data

Manually enriching your data, because either doing it perfectly with the computer isn't possible or takes too long, or because the extra data is unstructured, is something that you need to embrace when doing data analysis and creating data visualizations.

In this case, after trying an automated route, I *manually* converted each unique word from all the titles into a more general term. This variable in turn became the main aspect that defined the location of the books, thus it became quite important and worth the time investment!

I loaded this curated list back into R and replaced all the specific title words with their general ones. The final data preparation step was to turn the set of fantasy book titles into a numerical matrix, which could then be used in clustering analyses. I won't go into the details of how this was done, but if you're interested, you can google for "Document Term Matrix."

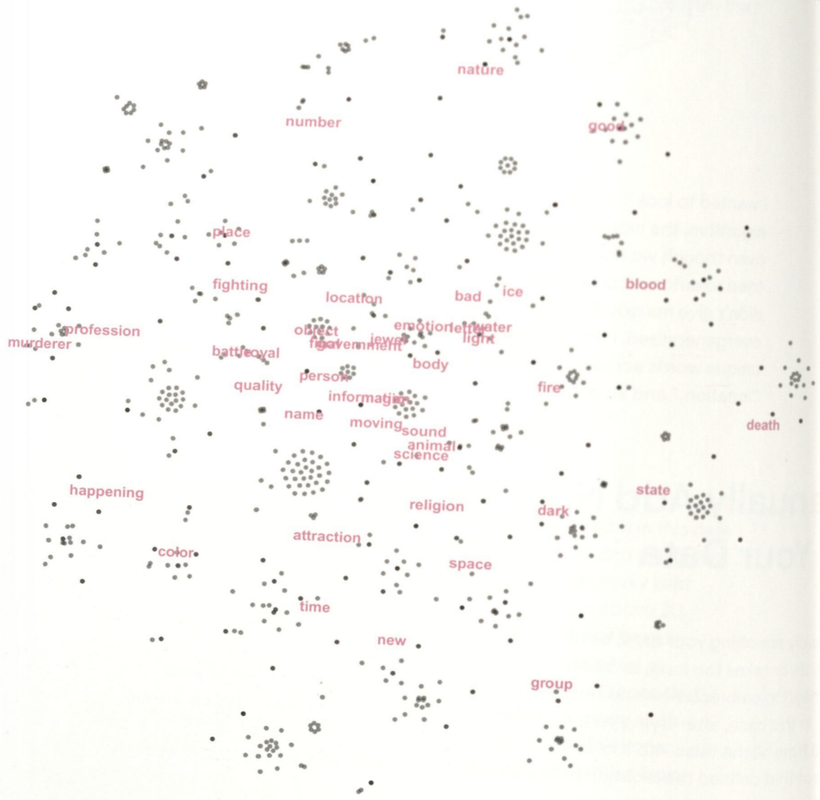
I tried several clustering techniques on the books, such as K-means, Principal Component Analysis, and tSNE, to see which result would visually give back the most insightful result. My goal with the clustering was to get an x and y location for each book, where similar titles were grouped or placed together in a 2D plane. Inspecting the resulting visuals for each technique, I found that tSNE gave back the best grouping of titles; books were spread out nicely and there were clear clusters of different topics.

I placed the ±40 most occurring words/terms on top of the tSNE plot, in their "average" location when looking at all the positions of the books that contained that term. While not perfect, this gave me a sense of where certain topics were located. A hotspot of books were present in most terms, but a few stragglers in other locations pulled most terms toward the middle.

I'll start using the word "terms" now to describe the most common replaced words of the book's titles, such as "magic," "name," or "nature."

Fig.5.2

The clustering result of running a tSNE on the book title words, with the "central" locations of the ±40 most common words plotted in pink.



My final data step was to prepare an extra variable that would be used to draw a path between all books from the same author. Many books didn't have a publication year from the Goodreads data, so I couldn't do it chronologically. The next best thing was to draw the shortest path between the books, so the length of the lines in the final visual would be minimal. Thankfully, I could use the "Traveling Salesman Problem" approach (imagine a salesman wanting to travel between cities in the shortest distance possible). With the TSP package in R, I calculated the order in which the books should be connected.

Sketch

Throughout the data preparation phase I was thinking about how I wanted to visualize the results. From the get-go I already had the idea to put the books in a 2D plane, placing similar titles together. But how could I get an interesting shape for the books, other than just a boring circle?

Since I was looking at titles, I thought it would be fun to somehow base the resulting "shape" of a book on the title as well. I could split the 360 degrees of a circle in 26 parts, one for each letter in the English alphabet, and then stack small circles on the correct angles, one for each letter in a title. I would then connect all the letters from a word in the title with curved lines, sort of "spelling it out." In Figure 5.3, you can see where I was still deciding if I wanted the lines connecting the letters, to only go around the outside or through the middle of the circle as well (the top right part of the sketch).

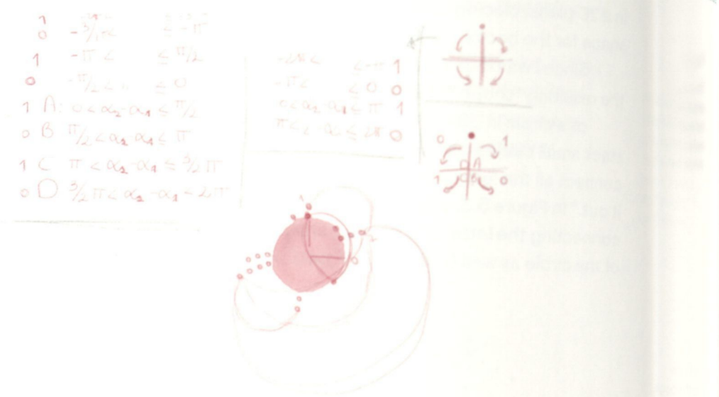
Fig.5.3

Figuring out how to visualize the book "marks" themselves in an interesting manner.

After having had so much fun with SVG paths during previous projects, I wanted the lines between the letters to follow circular paths. One of the elements that I had to figure out for these SVG arc paths was the `sweepflag`. Not too difficult, but much easier to figure out if you draw it (see Figure 5.4).

Fig.5.4

Trying to figure out how to get the SVG arc sweepflag signs correct for starting positions in each quadrant of the circle.



Skipping ahead a bit to a point where I had already started on a simple visualization of the books (plotting the book circles, nothing fancy), I looked into the most common terms of the titles again, such as "magic" or "royal." From my explorations during the data phase I knew that placing them in their exact "average" location just wasn't quite right. I wanted to have them more on top of their hotspot and not be pulled towards the center by a few books in other locations. Therefore, I created simple plots in R that showed me where books relating to a certain term were located. See the pink circles that belong to books with the term in their title that's above each mini chart in Figure 5.5, such as a clear grouping of books whose titles relate to "fighting" on the left side of the total circular shape.

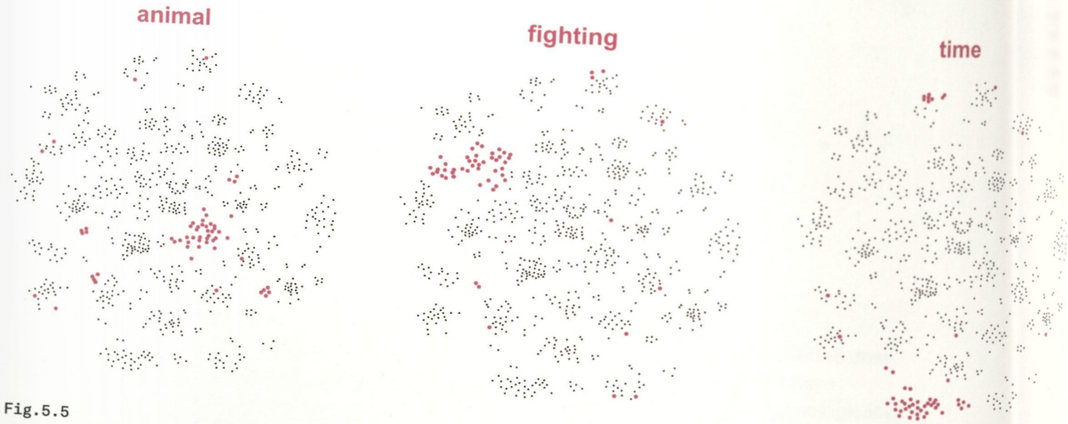


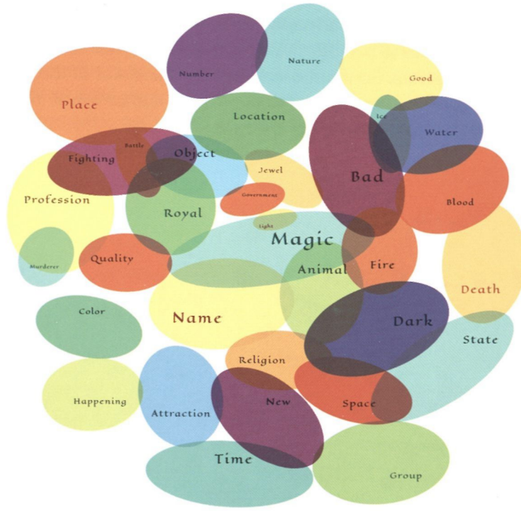
Fig.5.5

Seeing where books that relate to a certain theme/term fall within the whole.

I then took the tSNE plot with all the books into Adobe Illustrator, and together with the charts from R, I manually (yes, again) drew ovals for the top 35 terms that had a clear hotspot. This resulted in the arrangement that you can see in Figure 5.6, which I could use as a background image behind the book circles.

Fig.5.6

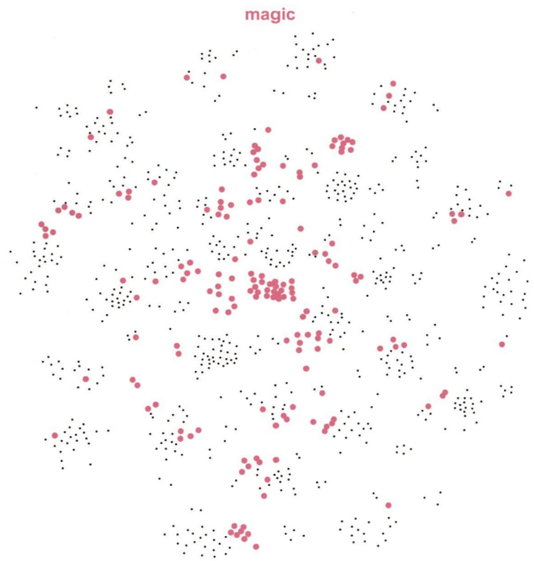
The landscape that reveals itself when looking at the hotspot locations of the most common ±30 terms



What this exercise also taught me was that "magic" is found practically everywhere throughout the circular shape (Figure 5.7), hence the title of the final visual.

Fig.5.7

Books that have a title that refers to "magic" are found practically everywhere in the tSNE result.



With the x and y locations of the tSNE result finally attached to each book title, I could begin to code the visual with `D3.js`. Thanks to past experience with creating circular paths between two locations (such as the lines in between two circles in my previous project about European royalty) it started out rather painless. A simple addition to the visuals was to size the circles in their area according to how many ratings the book had gotten. Furthermore, I used the thickness of the path that connected books from the same author to denote the author's rank in the Amazon top 100; the thicker the path, the higher the rank.

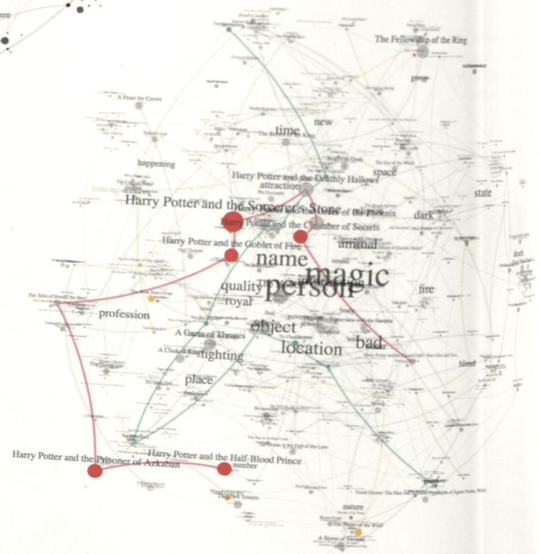
People on Goodreads definitely like J.K. Rowling: those circles became huge!

Fig.5.8

Placing the books in their tSNE locations in the browser while sizing the circles and line thickness by number of ratings and author rank, respectively.

Fig.5.9

Coloring the circles and connecting lines of some of my favorite authors and book series.



To make the visual a bit more personal, I chose five authors (my three favorite authors, plus two other authors that I enjoyed a particular series from) and marked these with colors, both in terms of their circles and paths (Figure 5.9).

Those authors are Brandon Sanderson, Patrick Rothfuss, and J.K. Rowling, plus Terry Goodkind and Brent Weeks.

Next, I focused on the book "shape," adding the small circles, one for each letter in a title, around the main book circle and then connecting the letters of one word with a path. I had some fun with these arcs, trying to make them swoosh a bit. However, the paths were getting much too big, obscuring titles and other books, so I tuned it down a bit. (Figure 5.10 is still my favorite visual result.)

As you can see from the previous images, many titles overlapped even though the tSNE algorithm did a great job of separating the books in terms of main themes. Since it was important to be able to read the title of a book, I had to adjust the positions of the circles to reduce overlap.

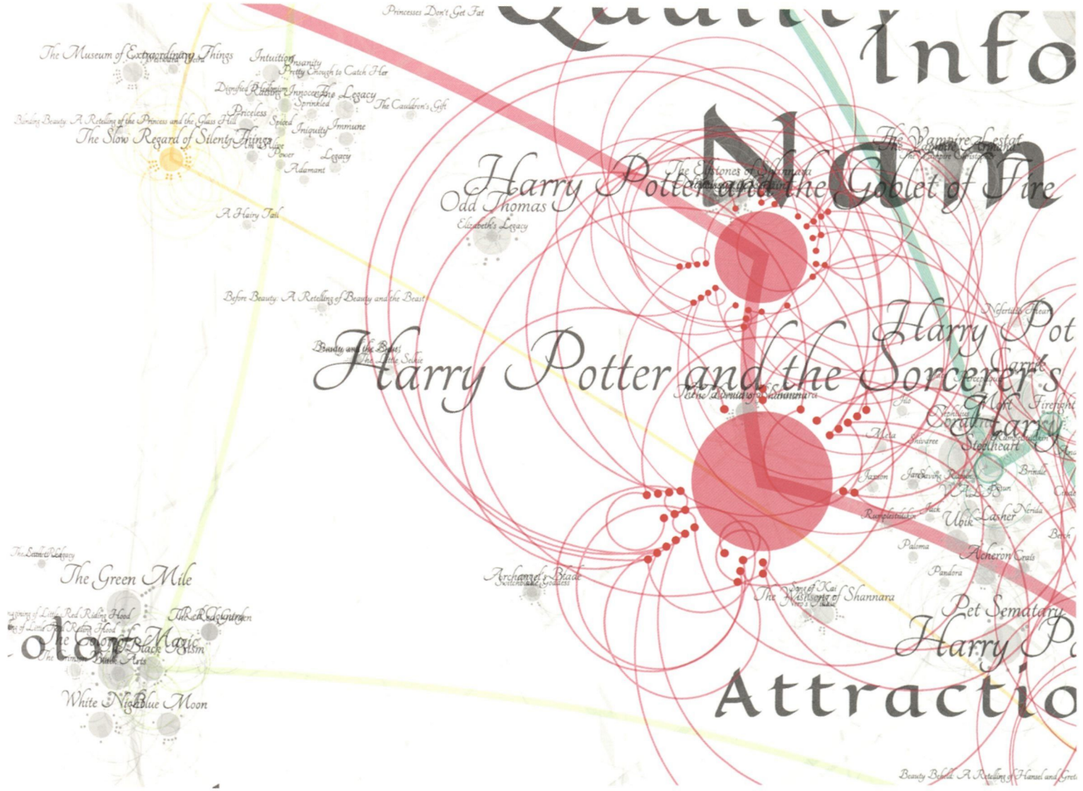


Fig.5.10

Creating each book's visual mark with the title "spelled" out with the smaller circles around the main one, connecting each word in a title with swooshing arcs.

So, yes, for the third time in this project alone, I took the manual approach. I wrote a small function that made it possible to drag each book around, saving the new location in the data. I've since come to love the fact that you can actually save data variables into your local browser so that, even after a refresh, the books would reappear on their moved locations! (Search for `window.localStorage.setItem()`)

Using the handy `d3.drag()` it's very difficult to SVG groups across page.

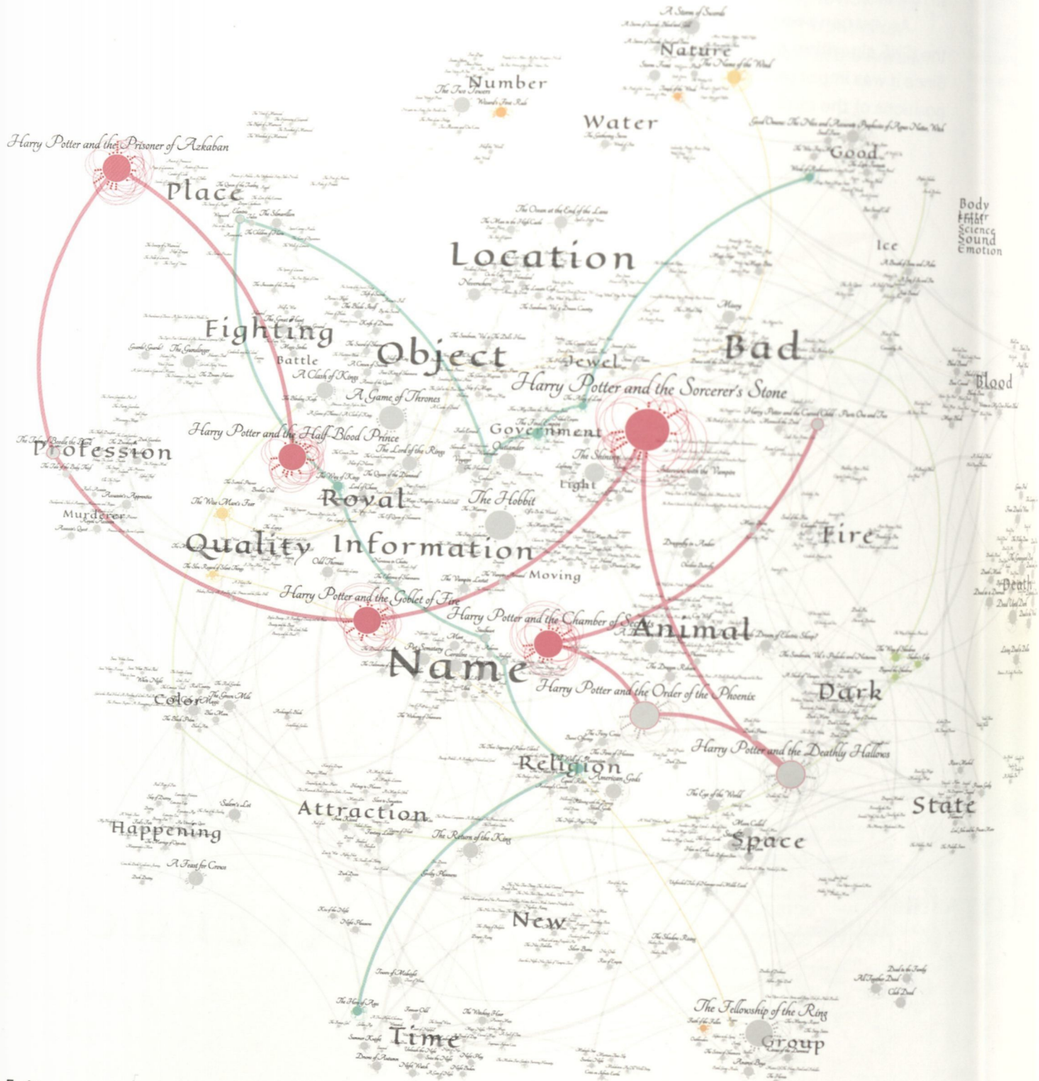


Fig.5.11

Having moved the books to reduce the worst textual overlaps.

↳ Precalculate "Visual" Variables

For this project I precalculated "visual variables" to add to the dataset, the x and y pixel locations, first gotten from the tSNE clustering and then fine-tuned by dragging them around to prevent (title) overlap. This made it possible to immediately place each book's circle onto its final location. However, I also precalculated in what order the books by the same author should be connected to have the shortest line. It would've made no sense to have each viewer's browser have to calculate almost 100 shortest paths, if the outcome is always the same.

Surprisingly, it only took me about an hour to slightly shift the ±850 books into practically non-overlapping locations. Taking these updated x and y positions into R, I created a new CSV file that became the dataset that is used for the final visual result.

With the books done, I focused on the background locations of the most common terms, such as "magic," "blood," and "time," using the hotspot-oval SVG from Figure 5.6. The advantage of using an SVG image is that I could still make changes to the ovals with JavaScript and CSS. I felt that blurring all the ovals extensively to merge the colors around the outsides would be the way to go (Figure 5.12). Thankfully, I found that it looked even better than I had imagined. (/◕ワ◕)/*:・゜✧

Fig.5.12

Using SVG blur filters to make the background ovals smoothly blend into each other.

As a bonus, having this (blurred) background gave me the option to make the book circles, swooshes, and lines a nice crisp white, instead of the boring grey they were before.

Fig.5.13

To explain what the smaller dots and swooshes around each book's circle meant, I created an animation that shows a book's title being "spelled out" (Figure 5.14)

Finally, I only added a minor interactive element; when you hover over a book, it highlights all the books by that author. With the online page done, I also turned the visual into a static print version, updating the layout to one I felt more fitting for a print.

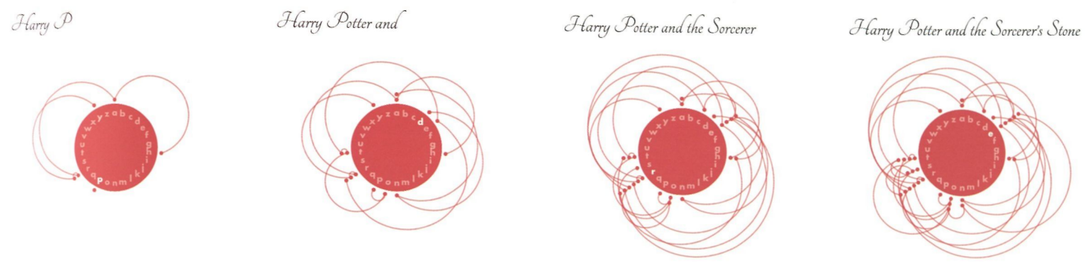


Fig.5.14

A few different moments from the animated legend explaining how to interpret the smaller dots and swooshes around each book's main circle.

Reflections

I'm very happy with the end result; I love rainbow colors, and the blurry aspect reminds me of fairy dust. Perhaps it's more data art than dataviz though. (*^▽^*)ﾉ The code part was thankfully not very difficult this time. The most intricate things to program were the swooshes. The whole project was really more about going back and forth between the data in R, other elements in Adobe Illustrator and the main visual in JavaScript, and D3.js taking more time than expected.

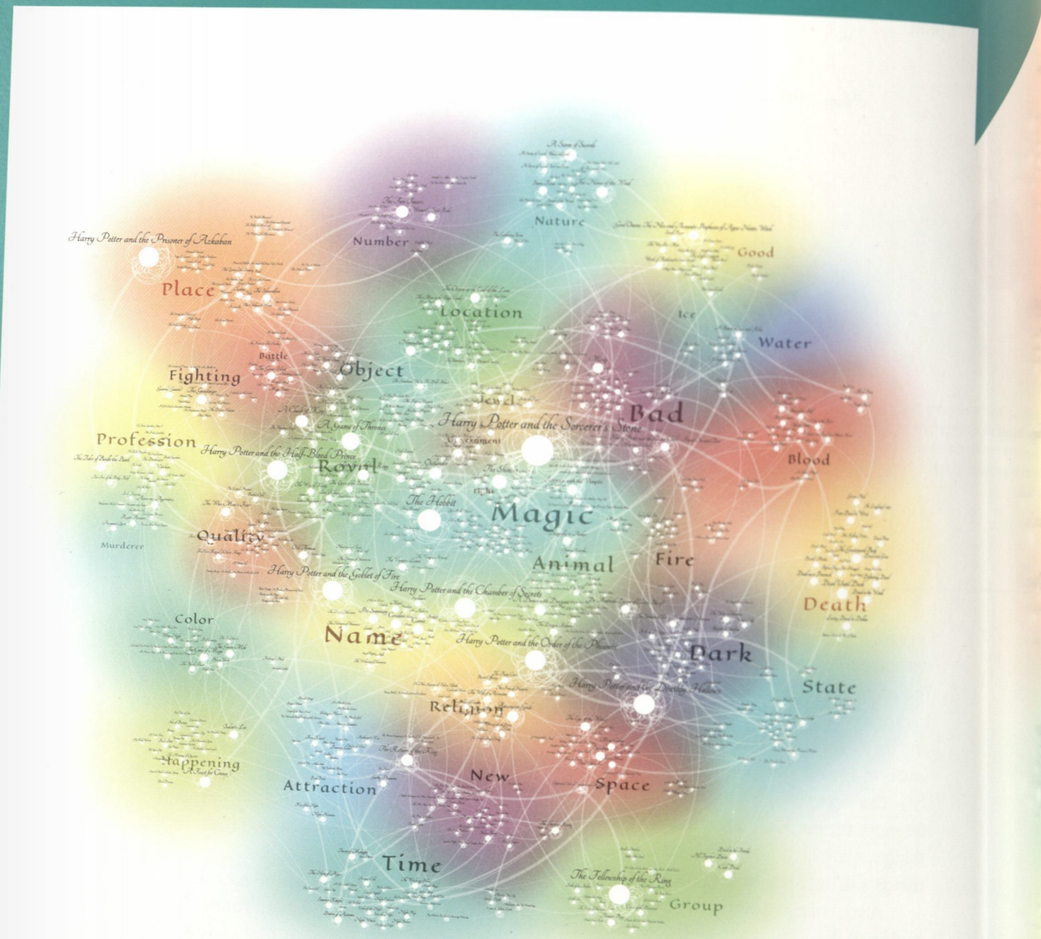
During the entire process of creating the visualization I noticed that there are far more terms that relate to bad things, such as "blood," "death," and "fire," than things relating to good aspects, such as "light." Maybe references to evil and bad things sell better/create more interesting stories?

Also, I had expected that many authors would probably be fixed within a certain region of the map, all their books following the same title themes. However, that turned out to be false. Most authors are actually spread across the map. Only a few really stick within one location; Charlaine Harris is quite fixed on including "death" in her titles, for example.

Finally, although the interactive hover makes it easier to explore this visual, in the end I prefer the static poster version. It's large enough so the small details of the swooshes and tiny dots around each book's main circle can clearly be seen, and even the smallest book titles are legible. It's both nice to look at and hopefully invites you to dig in and find insights into the world of naming a fantasy book.

Magic is Everywhere

↳ MagicIsEverywhere.VisualCinnamon.com



Magic is everywhere

Investigating patterns in Fantasy book titles

The titles from the top 10 books of the top 100 best-selling fantasy authors on Amazon were collected. Using text-mining the titles were analyzed for general subjects or terms, such as **fire**, **royal**, **time**, & more. Finally, these titles were clustered in a 2-dimensional plane, which placed books with similar themed titles together.

Created by Nadieh Bremer | Visual Cinnamon

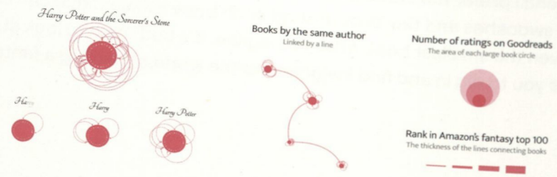


Fig.5.15

The final poster/ print based version of "Magic is Everywhere."

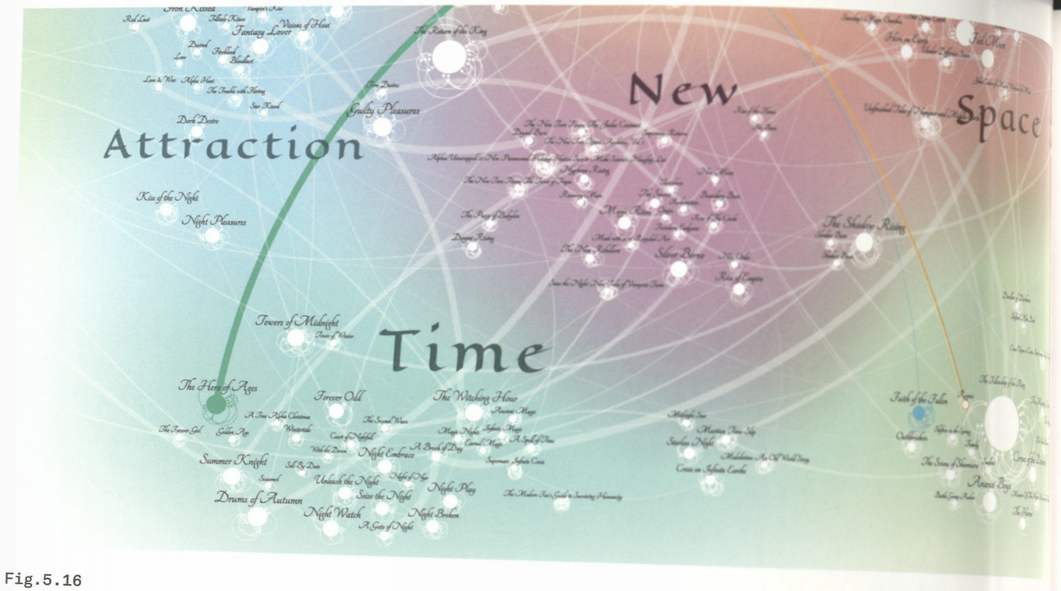


Fig.5.16

Zooming in on the lower-left section of the map that focuses on the themes of "time," and "new."

Fig.5.17

Hovering over a specific book circle will highlight the author, their rank in the top 100, and all of the other books by the same author.

Fig.5.18

The title that I made for the online version.

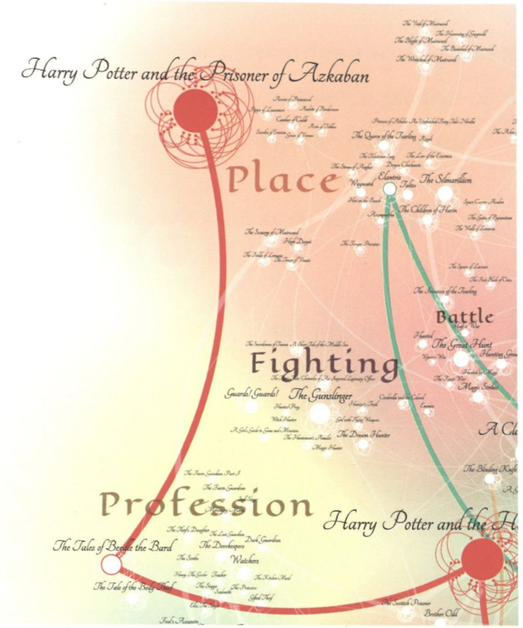


Fig.5.19

Zooming in on the upper-left section where the third Harry Potter book stands out from the other books.

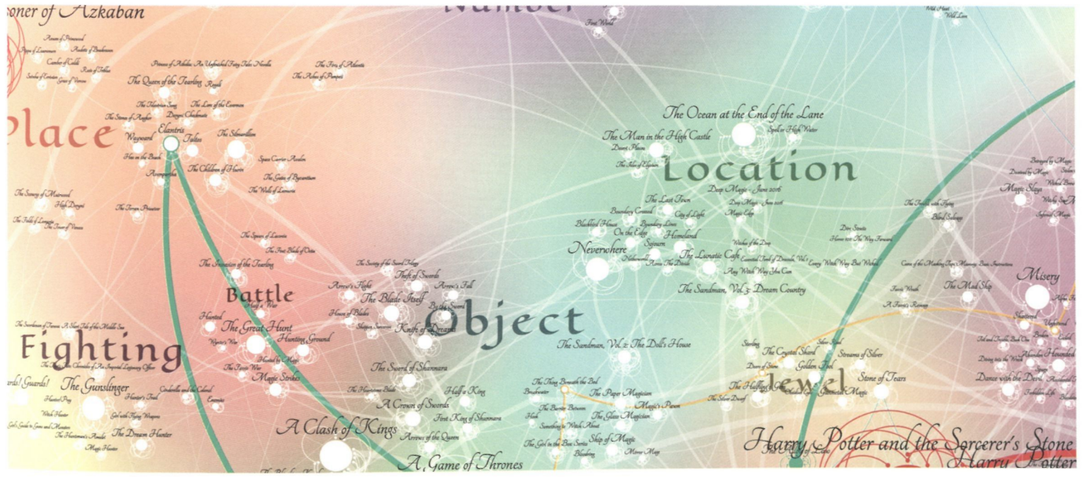


Fig.5.20

Zooming in on an upper-middle section of the map that focuses on the themes of "object," and "location."

characters, conversations, and themes below to explore them. Take advantage of the fact that some characters, conversations, or themes will disappear as you filter down; their co-appearances and co-occurrences are often times just as interesting as the songs that are left.

If you get into a bad state, reset.

↓

Ambition
a1 a2 a3 a4

Personality
p1 p2 p3

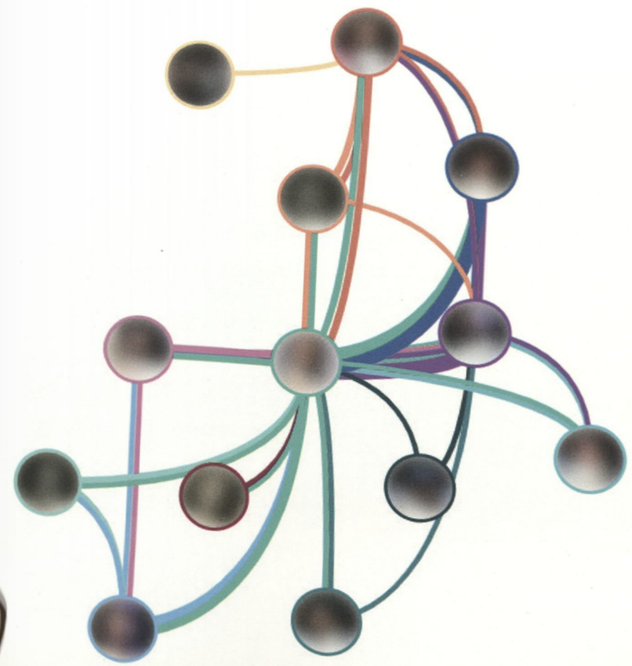
Contentment
c1 c2 c3

Miscellaneous
m1 m2 m3

Legacy
l1 l2 l3

Relationship
r1 r2

Death
d1



Every Line in Hamilton

SHIRLEY

In the summer of 2016, I got really, *really* obsessed with *Hamilton: An American Musical*. It was quite a unique experience because all of the show's lines and dialogues were contained in songs, so I could get the whole plot by listening to the cast recording. I had it on repeat for months, and it got to a point where I was analyzing lyrics and searching for recurring themes throughout the musical. At one point, my boyfriend (now husband) suggested I turn it into a data visualization. I was really resistant at first ("that's beyond obsessive!"), but eventually gave in ("ok, I guess I *am* that obsessive.") I had been talking to Matt Daniels from *The Pudding*—a collective of journalist-engineers that work on visual essays—about working on a story together and pitched the idea to them. I wanted to create a visual tool to analyze character relationships, recurring phrases, and how they evolved throughout the musical—and they agreed.

I had originally budgeted one month to work on the project, but it ended up taking three months on and off. It took so much time and was so all-encompassing that I didn't have the time to work on a project for the "Books" topic, and I asked Nadieh if I could turn my *Hamilton* visualization into a *Data Sketches* project. It was a musical, but I made the point that I had created the dataset using *Hamilton: The Revolution* (a detailed book about the creation of the musical, lovingly referred to as the "Hamiltome" and co-written by Lin-Manuel Miranda, the creator of *Hamilton*), and Nadieh thankfully agreed. (; · ∀ ·)ﻭ