Process Book: Once Upon a Time
David Huang

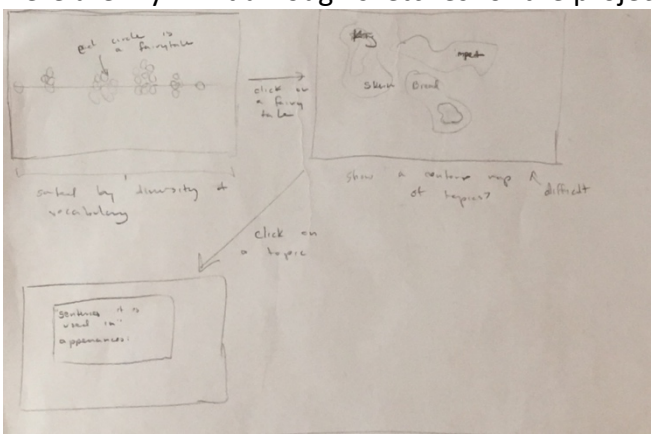
Key Questions:

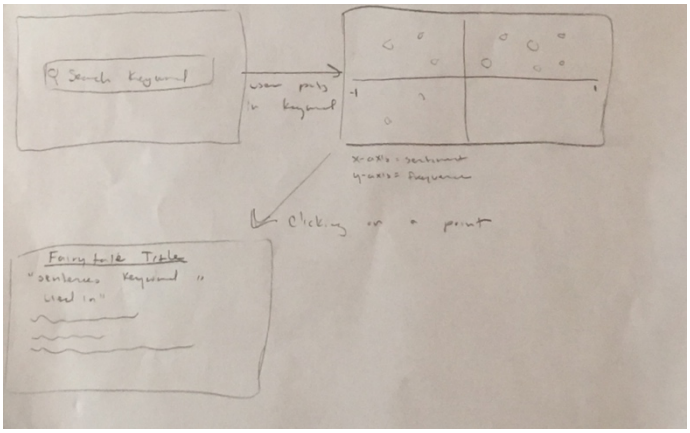
1. **What do you want to achieve with the visualization?**
   Given our dataset and all the fairy tales I read as a child, I recognized that fairy tales are a diverse genre of short stories, and I was most interested in how I could visualize that sort of diversity. Although there is no concrete measure for "diversity", I settled on the cosine similarity between stories to visualize the dataset's diversity.
   So, with this visualization, I want to show readers how strongly correlated two fairy tales are, and supplement that result with a few basic statistics to perhaps support the conclusion.
2. **What tasks do you want to support?**
   It is most important for the user of the application to be able to choose two fairy tales they are most interested in.
   Users should also be able to quickly compare fairy tales and move around the visualization easily (a dataset of 43 fairy tales is bound to take up a lot of the screen).
3. **What designs will help you achieve these tasks? Name at least two.**
   To quickly jump around from fairy tale to fairy tale and learn a little about each, I will create a matrix of fairy tales.
   Hovering over an element in the matrix will display a tooltip with basic information: titles and the cosine similarity.
   The color intensity of an element in the matrix will also serve as an indicator of the cosine similarity.
   Clicking on an element will reveal more information about two fairy tales that the user is interested in.

Quick Overview of the Project:
This project visualizes how correlated/similar two fairy tales are. Correlation is measured using cosine similarity (the angle between two vectorized documents), so 0 represents a weaker correlation and 1 represents a strong correlation between the documents. The goal of this visualization is to give the user an idea of how a fairy tale might be more similar to one instead of another. To do this, I chose to use a correlation matrix where each row and column are a fairy tale from the dataset. Color is used to give the users an idea of strongly correlated a fairy tale is with another, and other statistics like tf-idf scores, or number of words shared are used to augment the cosine similarity.

Process with Sketches

Here are my 2 initial rough sketches for the project:





I liked the idea of constructing a visualization about the usage of language in fairy tales (topics, sentiment, term frequency, etc), but I quickly found that I lacked the experience and knowledge to achieve what I wanted to do with both visualizations. Here are the problems I found with these sketches:

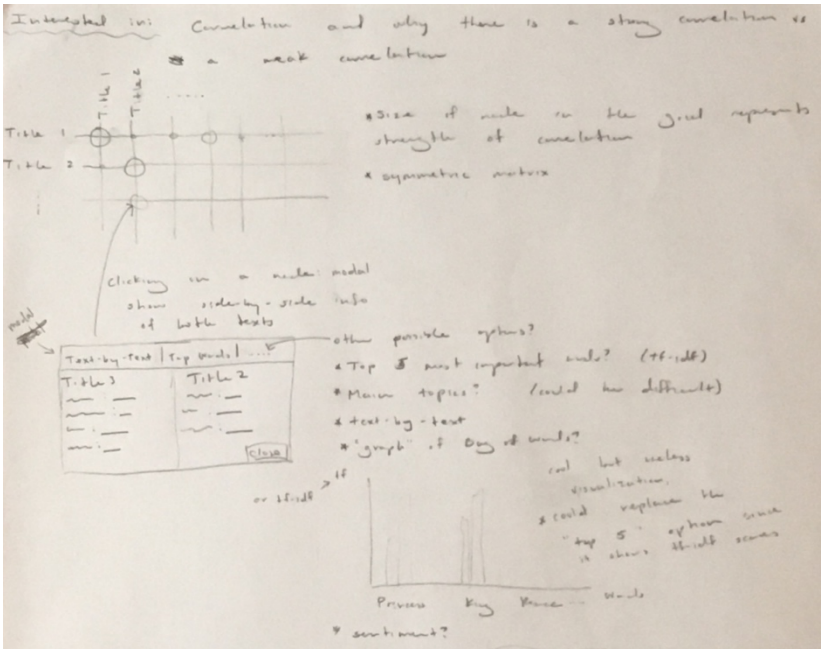
Problems with Sketch 1:

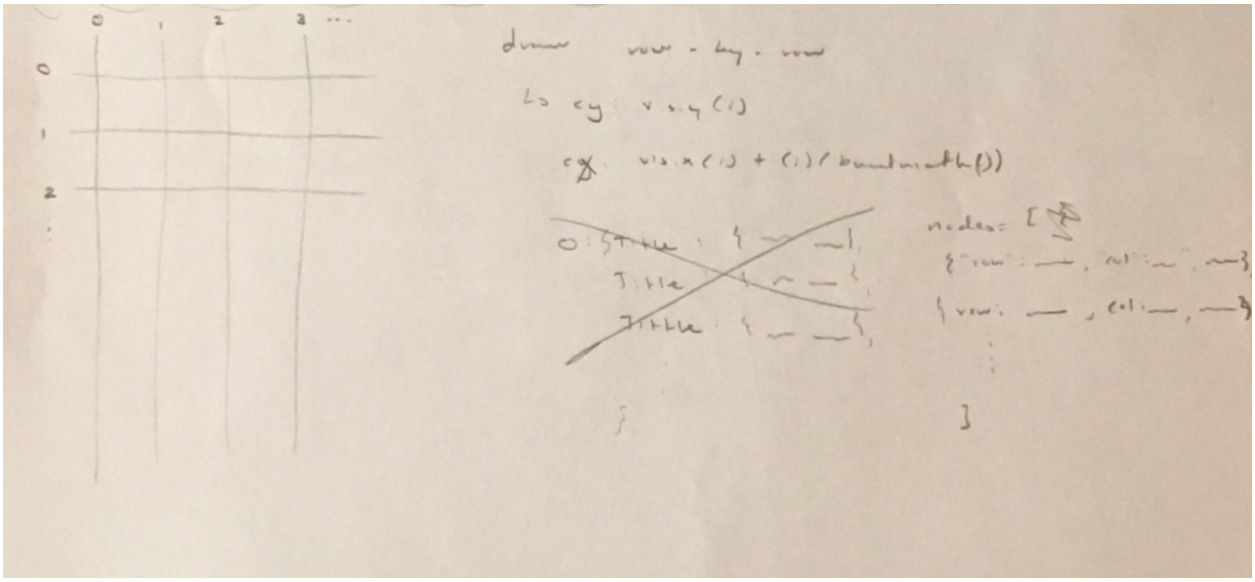
- finding topics in documents is not an easy feat
- contour maps don't answer anything compelling about the data, it's more so just giving an "overview" of the topics in the fairy tale

Problems with Sketch 2:

- it's too simplistic
- sentiment and frequency really shouldn't have any correlation, so it makes no sense to plot them
- making the user type a keyword first won't always yield itself to interesting results (e.g. if they typed "computer")

My next step was to redefine what I was interested in and draw sketches focused on that idea. I found that explicitly writing out goals and what I was interested in helped a lot in my design process, and ultimately resulted in much more clear designs than my previous two.



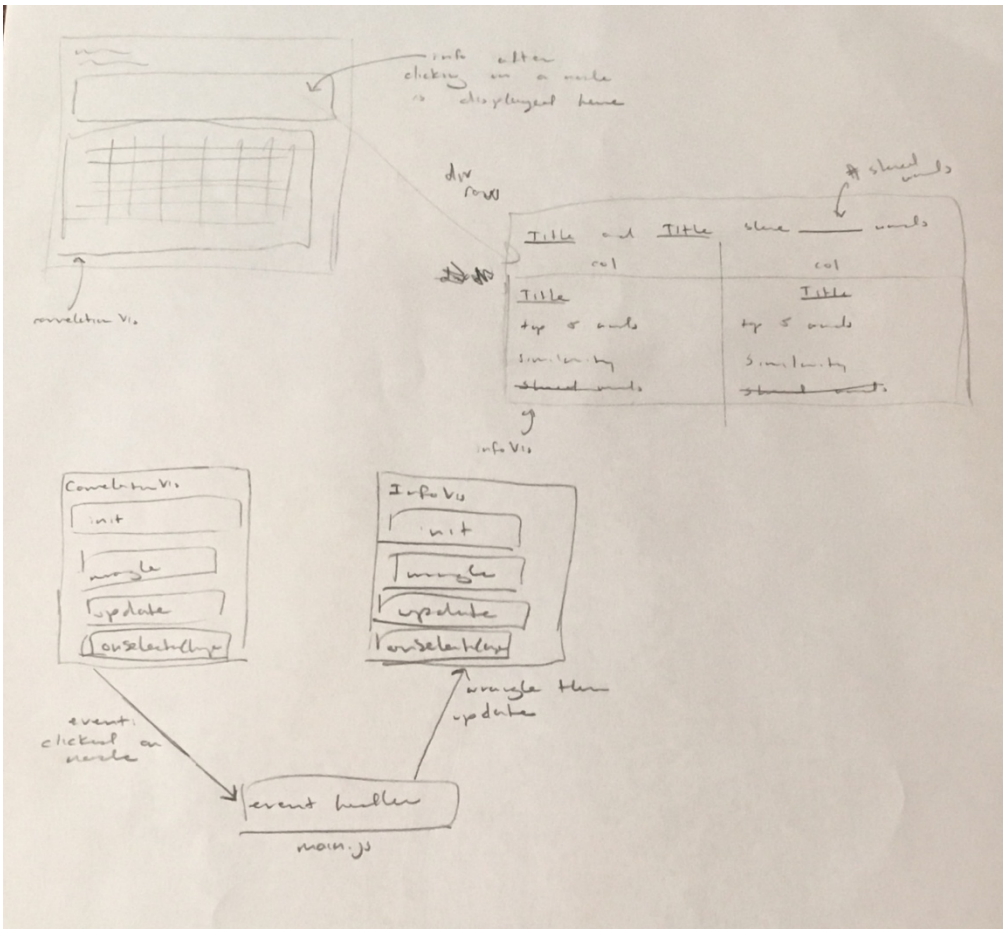


As you can see from my sketches, I originally toyed around with the idea of using size (the node's radius) instead of color to represent correlation. However, I quickly realized that the amount of fairy tales in the dataset would make it extremely difficult to contain variable radius circles within the fixed dimensions of the matrix. Moreover, some correlations have extremely small values ($< 10^{-2}$), so even if the circles were scaled up based on the cosine similarity, the visualization could still end up with small, annoying circles that would be difficult to interact with.

I also modified the click interaction after considering how tedious it could become to constantly open and close a modal just to learn more about two fairy tales. So, I chose to discard the modal and simply put the information above the matrix, all on the same page.

I used the five sheet methodology, but I found that by my third piece of paper I was already very satisfied with the sketches and chose to pursue the visualization laid out below.

All these modifications led me to my final sketch:



Final Design/Implementation

The data processing was done in Python using scikit-learn and numpy.

These fairy tales share 127 words. A correlation of: 0.1976.

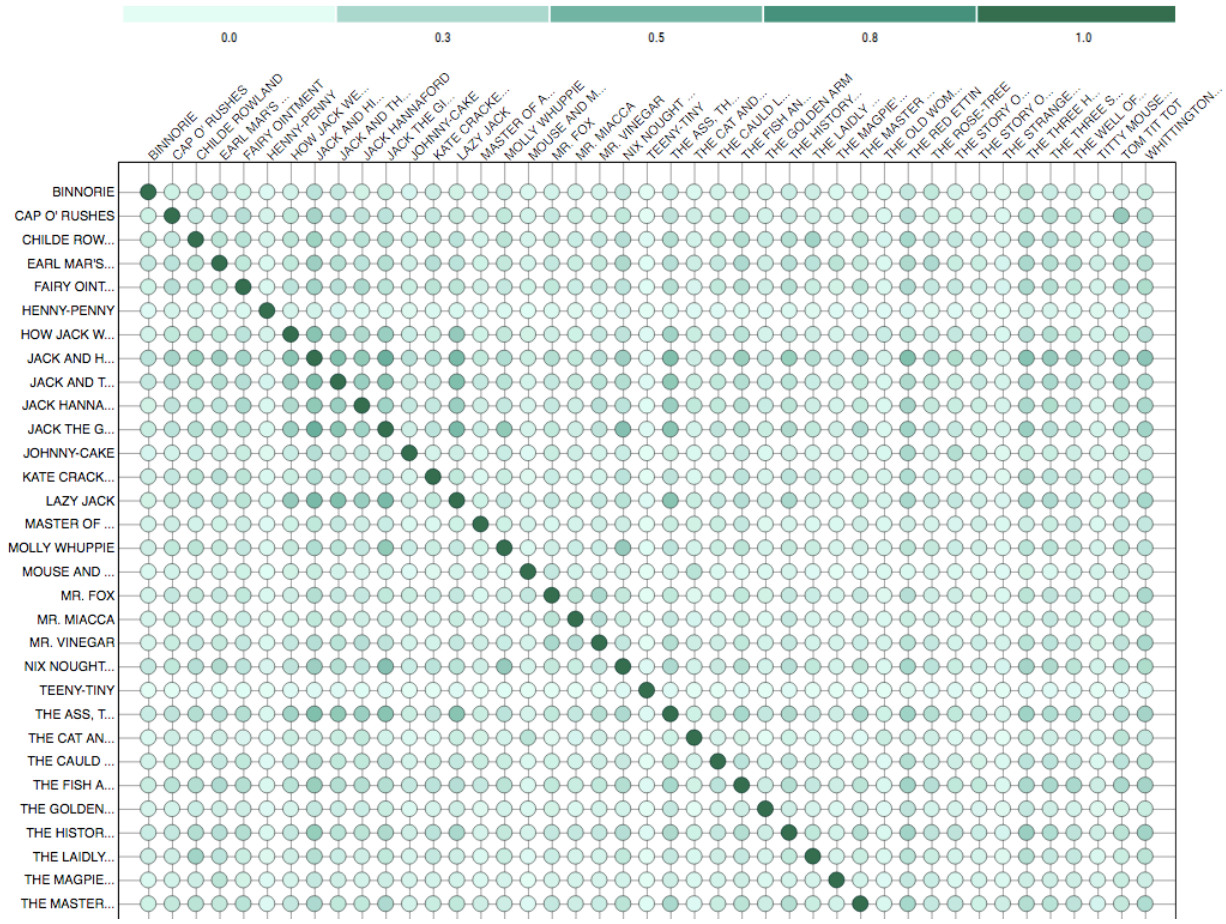
EARL MAR'S DAUGHTER

1) DOVE: 0.2882
2) FLORENTINE: 0.256
3) EARL: 0.224
4) MAR: 0.224
5) GOSHAWK: 0.192

KATE CRACKERNUTS

1) KATE: 0.6108
2) PRINCE: 0.2485
3) ANNE: 0.2114
4) NUTS: 0.2114
5) HENWIFE: 0.1905

Above is the information portion of the visualization, a body of text displaying supplementary data associated with correlation.



This is the actual correlation visualization (part of it is cut off). It is a symmetric matrix where each node represents a comparison between two fairy tales. A (circle in the matrix) node's shade of green corresponds to the value of the cosine similarity. A color scale is included with the visualization for more context.

Final Design Reasoning

Why a matrix?

When I was sketching designs, I came to the conclusion that a matrix was the most effective visualization that gave users a quick overview of the overall correlation between fairy tales. It was difficult to think of/draw other visualizations that would be just as effective as a matrix. As proof, just from glancing at the image of the visualization above, it is immediately apparent that fairy tales have a weak correlation with each other. There are only a few nodes with darker shades of green aside from the matrix's main diagonal.

Color

Color is used to represent the magnitude of correlation between two fairy tales. As explained before, it is a much more effective visual indicator than size, and it gives the user a fast overview of the overall correlation among the fairy tales.

Hovering

Following my design logic of giving users a quick overview, it felt natural to give more insight into why a node is the shade of green it is. So, I chose to display a tooltip when a user hovers over a node to reveal the actual cosine similarity.

Clicking

The goal of this visualization was to reveal how correlated fairy tales are. While the matrix and tooltips do give some insight, the information presented doesn't completely fulfill my goal. This is why I added this extra interaction: to display more data that might reveal more about the cosine similarity:

- 5 most important words from each fairy tale
- Number of words shared