Questions:
What do you want to achieve with the visualization?

In this visualization, I aim to compare the difference between English fairy tales, Russian fairy tales, and Indian fairy tales. In the visualization, I want to have a clear representation of important data settings, and I want to have an effective usage of color so that it makes it easier for the reader to follow my ideas.

What tasks do you want to support?

However, since "difference" is a rather general term, I break it down in my visualization presentation. The detailed comparison includes aspects as below:

- Analyze the frequencies of most common words in the documents by accessing the tfidf index of each of the words. Tfidf is used to see how "important" the word is to that document. Words like "you", "me", and "the" are filtered out in this process.

  Compare the frequencies of those words between English and Russian fairy tales so that we can see that fairy tales from those two countries prefer to use different set of words. For example, English writers might like to call their kid Jack, while Russians may call their characters like, Vladmir. There are also other things that may differ in fairy tales. For example, English might mention more about "forest" while Russians mention more about "snow".

- Analyze different linguistic dimensionalities of English, Russian fairy tales, and Indian fairy tales. This includes Biological Processes, Drives, Social, Cognitive Processes, and personal concerns.

  I would then compare those factors between the fairy tales so that I can show the reader how those tales differ in those aspects and conclude from this sample data in what aspect exactly are fairy tales over these three places differ.

  For some aspects, however, there might not be a big difference in some aspects for the world fairy tales. Therefore, I may suggest that some things are rather similar in fairy tales all over the world.

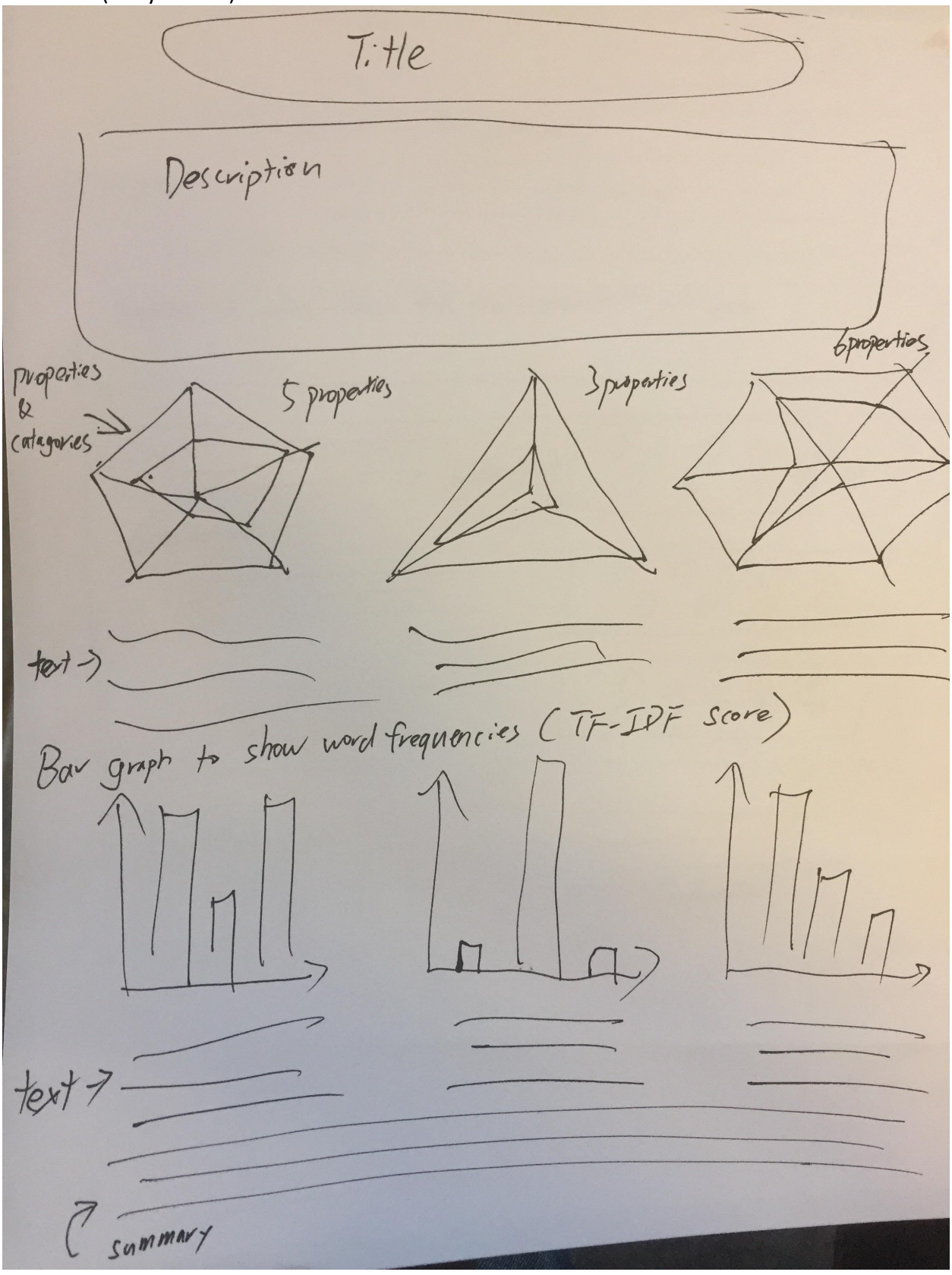
What designs will help you achieve these designs? Name at least two.

- For the tfidf analysis, I would like to use a babble graph, where each word is a babble and it has a larger size if it has a higher tfidf score. The color of the babble will depends on the linguistic tag of the word. For examples, different categories of nouns will be of different color.

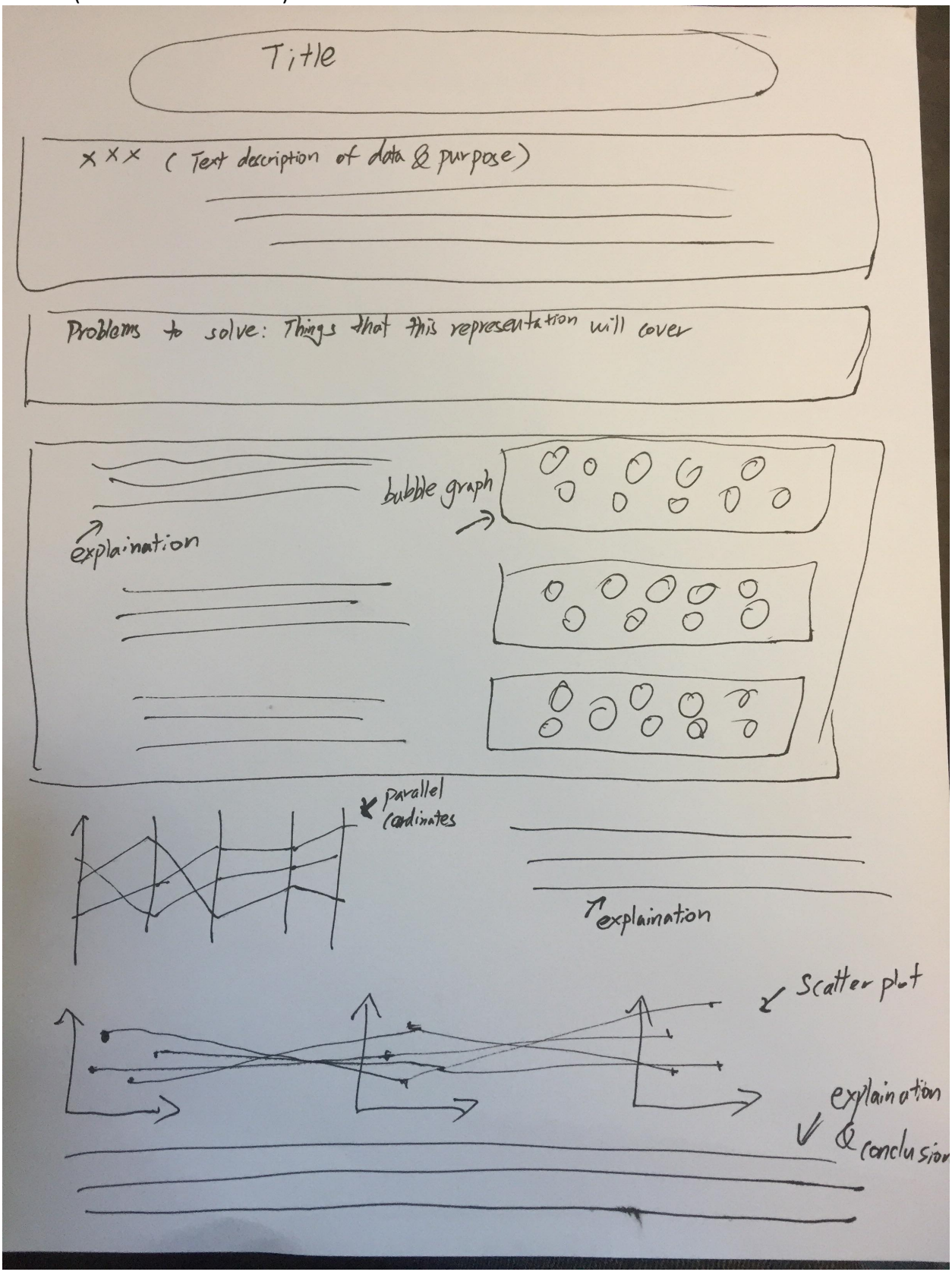
- For the linguistic factors, I will use a parallel coordinate for each one of them. I would use different color of lines to represent different countries. Also, for the factors, I would change the heat of the color based on the property it represents.

  When there is not a linear relationship of the properties so that heat color won't apply, I would use two-dimensional scatter plots and use dots to represent the country. Then I would connect the dots from the scatter plots so different plots will be connected and therefore make it easier for us to understand the data connections.

Sketches: (early Sketch)



Sketch (Sketch for final draft)



Justification for the final design

There are many reasons regarding how and why I choose my final design.

First of all, compared to my early sketch, my final draft gives the user a more dimensional view of the data. On the original draft, the user can either look at one property represented by a shape, or they look at a monotone bar graph where the three countries' data presented parallelly in a bar chart. On the other hand, in the final draft:

- The babble graph draws the reader's attention into individual words, so they can have a view of the most illuminant individual entities in the data set.
- The parallel coordinate allows the user to view multiple properties under a category all at the same time, so the user can have a broad view about that attribute. Horizontally, the user can know about the order of indexes based on its color. For example, in the order to emotions like happy->amused->delighted->neutral->sad->depressed->mad, each axis would represent one of those emotions and their color would be in an order where the closer to happy the warm the color, and the closer to mad the cold the color. Vertically, the user will be able to compare stats of different countries' fairy tales for the same attribute. For the parallel coordinate model, I unfortunately saw a page with some code from an old version of d3 (d3 version 2), when I try to learn online how to create draggable axis in a chart: https://bl.ocks.org/jasondavies/1341281. However, I did not copy the code and since we use completely different version of d3 and for different functionality, I derived my own code that fit best for the purpose of my report.
- The connected scatterplots focus in the relationship between attributes (not entities). For example, if the document mentions boys and girls, the number of boys the story mentions may not give a meaningful explanation of the data set. However, if we use the number of boys as x axis and number of girls as y axis, then these two data together give the user the idea of gender ratio in the data set. Moreover, connecting those dots from different scatter plots helps the user to track the ratio status of the same countries' fairy tale. The user just needs to follow the specific color of lines, to see all these ratios.

Second, the final draft represents the data better by using its graphical tools and get rid of some common pitfalls from the model in the old scratch:

- If the user wants to get to know the most basic attribute of the graph, nodes, he would not have a clue from the old scratch. However, in they new graph he would have an easy reference to the most common words in the dataset through the babble graph.
- In the different shapes representing attributes in the old scratch, the edges of the inner multi-edge shape do not have any statistical meaning. It only serves to make it look better but may interfere with the user's perception about the data distribution. On the other hand, to show it in parallel coordinate makes it easier for users to transit through features, and they will be like "Oh A does better than B in the aspect of X, but A lacks its strength in Y where B is good at".

- In the bar graph from old scratch, it is extremely hard to compare stats from one bar graph to another. The user has to read the label, take down the number, then look at the other graph and do some calculation. However, the connected scatter plot makes it easier to compare ratios, and parallel coordinates for attributes.

Last but not least, the final draft spread the text fields all over the page, because people generally like shorter paragraphs and spreading out the words therefore prove as an eye candy.