Leah Ajmani - lha37
Elise Colbert - emc277
Jonathan Lee - jyl88

3300 P1 Written Description

A. A description of the data.

1. *Description of Data*

Our project contains three visualizations that highlight information regarding the usage and popularity of different types of emojis around the world. More specifically, our first visualization plots different emojis based on their popularity and their corresponding sentiment, or positivity/negativity. Our second visualization shows which countries use the most of certain categories of emojis (food, sports, love, etc). Our third visualization displays the most popular emoji used per state according to our data set.

2. *Report where you got the data*

We used two data sets. The first data set was collected by researchers from CLARIN.SI analyzing emoji usage on Twitter on over 1.6 million tweets (with only 4% of said tweets containing emojis). The second and third data sets were collected by SwiftKey, a cloud service company, that analyzed emoji data across a wide arrange of categories to learn how different countries and U.S. states use emojis. Links to papers/reports are provided in the works cited. Additionally we used a map of the US from D3 that is also linked in the works cited.

3. *Describe the variables*

The variables used for the first visualization include Use Frequency and Sentiment Score. Use Frequency expresses how popular an emoji is (ex: an emoji that is used a lot has a high Use Frequency). The Sentiment score uses two variables from the dataset: Positive and Negative. The paper associated with the dataset calculated a “score” based subtracting the Negative variable from the Positive variable. We repeated this process and from there we created our own Sentiment Score that scales the difference of Positive and Negative and scales that to be from -1 to 1. So, our new Sentiment Score expresses how emotionally positive or negative an emoji is on a scale from -1 (extremely negative) to positive 1 (extremely positive) where a score of 0 would be neutral.

The variables used for the second visualization include country and category of emoji (ex: food, religion, sports, love, etc).
The variables used for the third visualization include state and emoji, correlating the top used non-facial emoji with its respective state.

4. *Filtering the data*

In terms of filtering and reformatting data, we filtered for the top 75 most popular emojis from the data set to be displayed on the first visualization. For the second visualization, we filtered for the top categories of emojis and filtered further by capturing only the top 4 emojis within each category to be displayed in our visualization. Finally, we filtered for the top emoji used per US state for the third visualization.

5. *Did you combine multiple datasets and how?*

We combined both datasets from CLARION.SI and SwiftKey by manually merging the data from the SwiftKey dataset that contained the country and emoji category information into the CLARION.SI dataset.

6. *Did you selectively choose a subset of the data to improve usability? What criteria did you choose for this selectivity?*

For our first visualization, we selectively chose a subset of the top 75 used emojis because using all the emojis provided in the dataset created massive usability/readability issues. Too many emojis overlapped with each other, making it extremely difficult to differentiate. After playing around with numbers like the top 50 and top 100, we found that the top 75 was a good medium because it showed the trend that most emojis cluster around the neutral mark on the sentiment score (Sentiment score = 0) and you can still see most of the emojis.

For our second visualization, we removed emoji categories we felt were less significant/meaningful such as “plants”, “holidays”, “sun”, and to improve usability. We wanted to show top categories from each country in the dataset, so we removed some categories that were less interesting if that country was already represented several times in the visualization. Our original visualization displayed more emoji categories so it was difficult to read and distinguish where categories cut off. By reducing the number of emoji categories used, the categories were able to become further spaced out, thus solving the distinguishing problem.

B. A description of the mapping from data to visual elements.

For each of our visualizations, we mapped the emoji data by using the emoji icons as data points on our graphs. In our first visualization, the vertical position of an emoji expressed its popularity while its horizontal position expressed its sentiment (positive or negative). We also implemented a background gradient to the first visualization to further emphasize the sentiment with red indicating negative sentiment and blue indicating positive sentiment. In terms of transformations, we used a logarithmic scale on the y axis that measured how frequently an emoji was used (popularity) in order to account for outliers. This helped better represent our data while minimizing the loss of information.

For our second visualization, we again used the emoji icons to reflect the respective emoji data. We also used color to distinguish between countries.

For our third visualization, we mapped emojis onto a map of the US to represent the most popular non-face emoji used in each of the respective states.

C. The story.

Our first visualization shows how often different emojis are used based on their sentiment. A key finding was that the most frequently used emojis are the least neutral in terms of sentiment. Even more surprising, however, is that across all of the top 75 emojis, most of them are neither strongly positive nor strongly negative in sentiment (notice how most of the emojis are found around +/- 0.2 out of a 1.0 sentiment scale). This is surprising as one might think emojis would frequently be tied to strong emotions since they are generally used to express emotion by nature.

Our second visualization shows the top categories of emojis used by different countries. It was surprising to see how each country's top emoji category reflected some subtle cultural differences. For example, it was interesting to see France using the most heart/love related emojis given they are typically seen as a very romantic country. It was also very interesting to see how Canada, Australia, and the US had some very different top emoji categories despite speaking the same language (English). Some fun discoveries were that raunchy emojis are used the most by Canadians and that LGBT and Technology emojis are used the most by people in the US.

Our third visualization shows the most frequently used non-facial emoji in each state. It was very surprising to see that the eastern states and east coast had a lot of top animal related emojis including the horse, baby chick, mouse, frog, and hatching chick for Tennessee, Kentucky, North Carolina, Virginia, Connecticut and Rhode Island respectively. It was also amusing to see how some states' most popular emojis embodied stereotypes they have. For example, Arizona's known to be very dry and it's popular emoji was the cactus. Hawaii is known for its beaches and surfing which is reflected in its popular emoji being the surfer. Also, New York's most popular emoji was the Statue of Liberty which is frequently tied to the state.

Works Cited:

https://www.clarin.si/repository/xmlui/handle/11356/1048
http://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0144296#pone-0144296-g003
https://www.scribd.com/doc/262594751/SwiftKey-Emoji-Report
http://cdn.swiftkey.net/USemojidata-SwiftKey.pdf
https://d3js.org/us-10m.v1.json