Beth Carlson

Professor George Legrady

Transforming Data

19 December 2011

SMS Visualizations

Data visualization is described as the study of the visual representation of abstract data. Data visualizations can be successful or unsuccessful in relating the abstract data set to the viewer. In order to explore the concepts behind what makes a visualization successful or unsuccessful, I have created several textual visualizations in order to compare and contrast. The data set consists of four years worth of SMS messages extracted from an iPhone. Using "Many Eyes," an experimental program created by IBM Research and the IBM Cognos software group, I created four types of textual visualizations. The different types include a word tree, a phrase net, a tag cloud, and a word cloud. Through this examination I plan to determine which approach produces the most effective visualization of the SMS message data set.

Before we look at each type, we must determine what constitutes a successful and effective data visualization. First and foremost, the data visualization chosen to represent a particular set of data must have purpose. Ben Fry, a data visualization expert and co-founder of the open source programming language Processing, wrote a book called *Visualizing Data: Exploring and Explaining Data with the Processing Environment* in which he explains this concept: "Each set of data has particular display needs, and the *purpose* for which you're using the data set has just as much of an effect on those needs as the data itself" (Fry, 2). Therefore, the question arises of what I am trying to accomplish by visualizing four years worth of SMS messages. Another way of putting this, Fry explains, is to ask What is the question?, before visualizing a set of data: "The most important part of understanding data is identifying the question that you want to answer. Rather than thinking about the data that was collected, think about how it will be used and work backward to what was collected" (Fry, 4). After the question is determined, one must choose how to represent the data collected. As mentioned above, the way in which the data is represented is as important as the data itself. Again, Ben Fry:

> The Represent state is a linchpin that informs the single most important decision in a visualization project and can make you rethink earlier stages. How you choose to represent the data can influence the very first step (what data you acquire) and the third step (what particular pieces you extract). (Fry, 9)

The next step is determining the representation and functionality of the visualization.

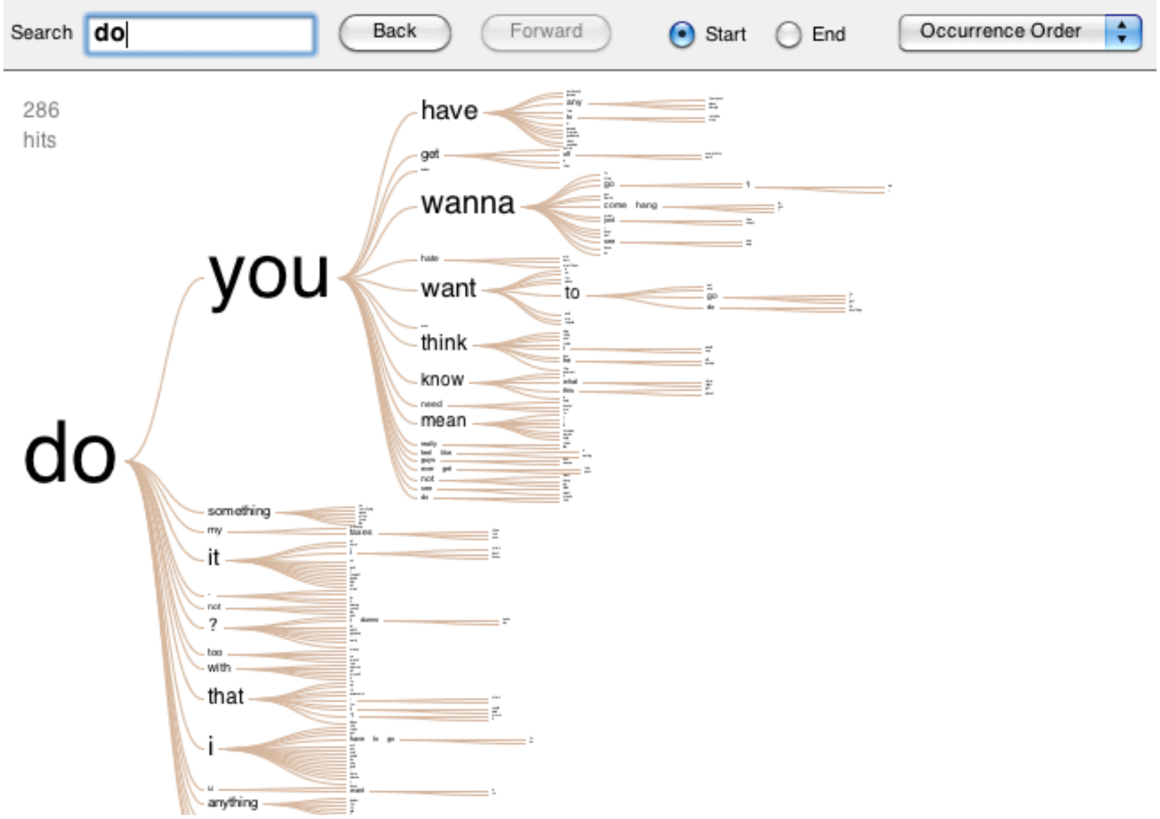
In his paper entitled "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," University of Maryland professor Ben Shneiderman describes the basic principle of a useful visualization as, "Overview first, zoom and filter, then details-on-demand" (Shneiderman, 337). He refers to this as the "Visual Information Seeking Mantra." Shneiderman goes on to outline the seven task-domain information actions that are required for a useful data visualization:

> **Overview**: Gain an overview of the entire collection.
> **Zoom**: Zoom in on items of interest.
> **Filter**: Filter out uninteresting items.
> **Details-on-demand**: Select an item or group and get details when needed.
> **Relate**: View relationships among items.
> **History**: Keep a history of actions to support undo, replay, and progressive refinement.
> **Extract**: Allow extraction of sub-collections and of the query parameters. (Shneiderman, 337)

A successful data visualization should allow for all of these actions – they are integral to the effectiveness with which a visualization can represent an extremely large set of data. It is important to be able to view the data set represented as a whole, as well as zoom in in order to examine certain parts in more detail. The visualization should also allow for the data set to be filterable, relatable, and extractable. In other words, the visualization should allow the user to interact with a large set of data in a seamless and effective way – a way that would not be possible without the visualization. And last but not least, a successful data visualization should be aesthetically pleasing. Shapes, sizes, and colors all play a role in creating a useful, well-designed visualization.

Now that we have outlined the aspects required for a successful data visualization, let us examine the four SMS visualizations to determine how they perform.

Visualization A: Word tree



Search do

● Start ○ End

Occurrence Order

286 hits

Figure a.1

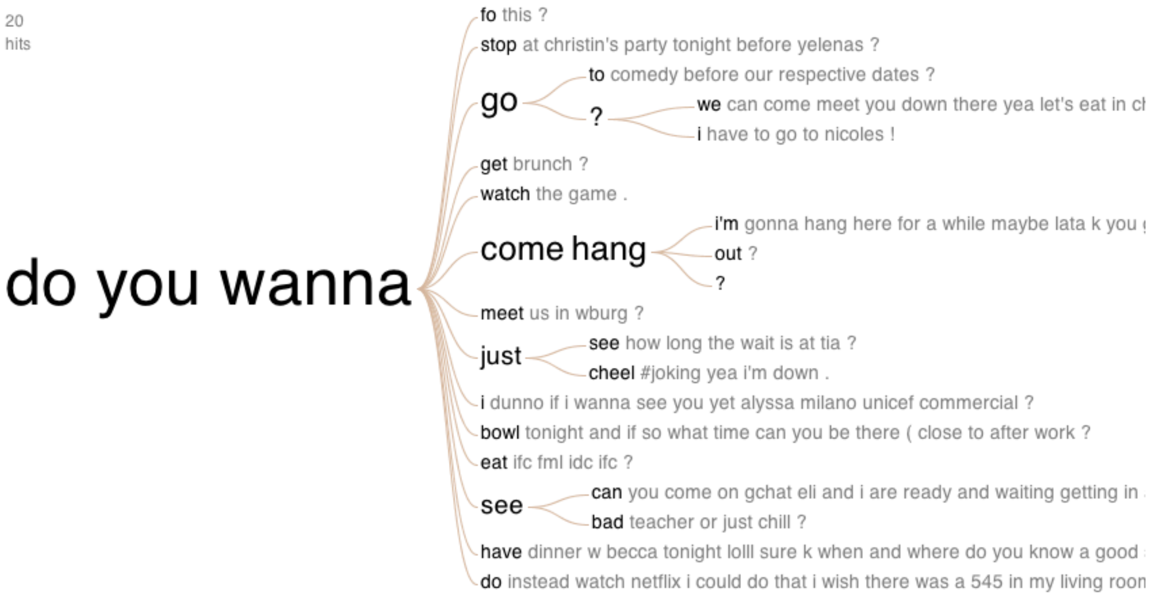
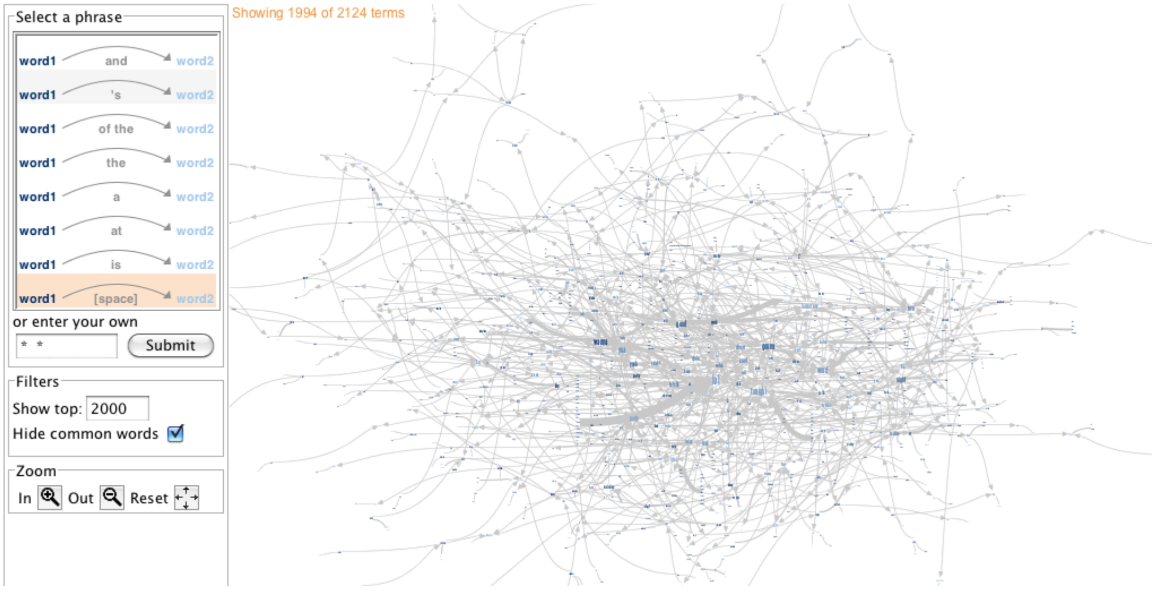

20
hits
do you wanna
fo this ?
stop at christin's party tonight before yelenas ?
go
to comedy before our respective dates ?
?
we can come meet you down there yea let's eat in cl
i have to go to nicoles !
get brunch ?
watch the game .
come hang
i'm gonna hang here for a while maybe lata k you
out ?
?
meet us in wburg ?
just
see how long the wait is at tia ?
cheel #joking yea i'm down .
i dunno if i wanna see you yet alyssa milano unicef commercial ?
bowl tonight and if so what time can you be there (close to after work ?
eat ifc fml idc ifc ?
see
can you come on gchat eli and i are ready and waiting getting in
bad teacher or just chill ?
have dinner w becca tonight lolll sure k when and where do you know a good
do instead watch netflix i could do that i wish there was a 545 in my living roon

Figure a.2

Visualization A is one of the more successful visualizations of SMS messages, but it is not perfect. The reason for this is that it does not provide an overview of the entire data set before zooming in on specific words and phrases. The purpose of visualization A is to allow the user to view all of the sequences following a certain word. The user types a word into the search field and a word tree appears with all of the sentences following that word. The user can then click on any part of the word tree to zoom in and view each sentence. When the user is choosing a word to zoom into, and hovers the mouse over that word, it changes from black to pink. This color change adds to the interactive feeling of the visualization. This visualization allows for details-on-demand, but there is no way to filter out uninteresting items. Visualization A does well relating items in the data set to one another, because the purpose of this visualization is to show the relationship among different words to form sentences. There is also a back button that allows the user to move backwards step-by-step along the zoom path. There is an option to share the visualization with other users, but you cannot extract sub-collections or the query parameters.

Visualization B: Phrase net



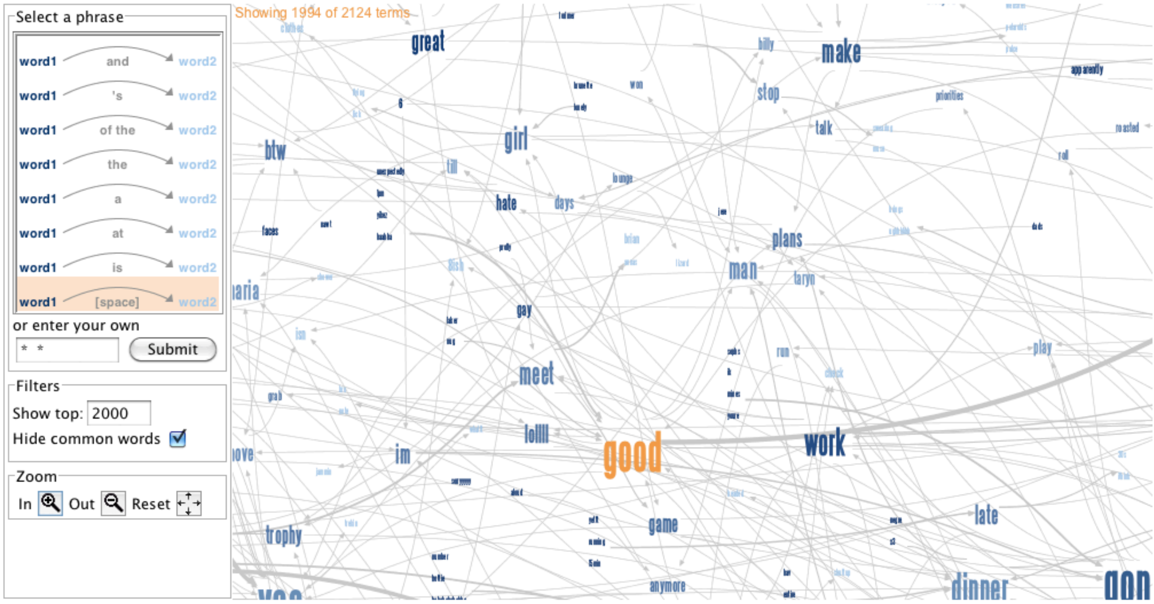
Showing 1994 of 2124 terms

* *

Show top: 2000

Hide common words ☑

Figure b.1



Showing 1994 of 2124 terms

* *

Show top: 2000

Hide common words ☑

Figure b.2

Unlike visualization A, visualization B begins with an overview of the entire data collection. Due to the large size of the data set however, you can see in figure b.1 that it appears as a blurry shape in which the individual words are too small to read. Once the user zooms in, they are able to access details-on-demand, such as the number of occurrences of a particular word. On the left hand side of the visualization, the user can filter the data set by choosing different connection words like 'and', 'it', 'the', etc. The user can also determine whether or not to visualize the entire data set, or a smaller sub-collection of words. Like visualization A, visualization B does well relating the data because its purpose is again to show the relationships among different words to form phrases and sentences. There is no step-by-step historical feature in visualization B like in visualization A, but the user can clear any filters applied in order to begin a new query of the data set. And again, there is an option to share the visualization with other users, but you cannot extract sub-collections or the query parameters. In visualization B, the colors chosen play a role in the representation of the data. The more occurrences of a particular word, the darker shade of blue in which it appears. And when an individual word is selected, it appears orange. More so than visualization A, visualization B utilizes thoughtful color scheme decisions that aid in the representation of the data set.

Visualization C: Tag cloud

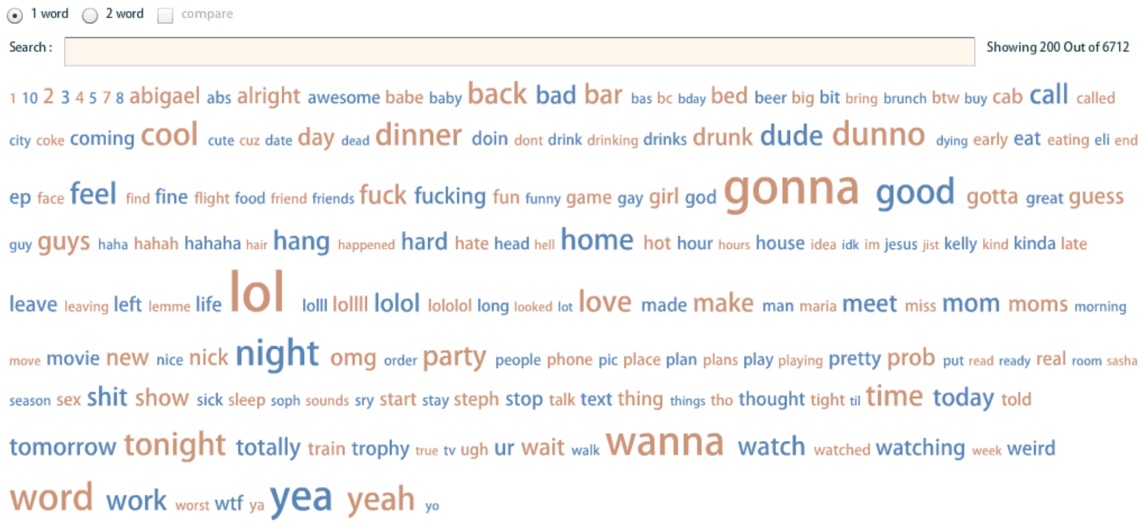
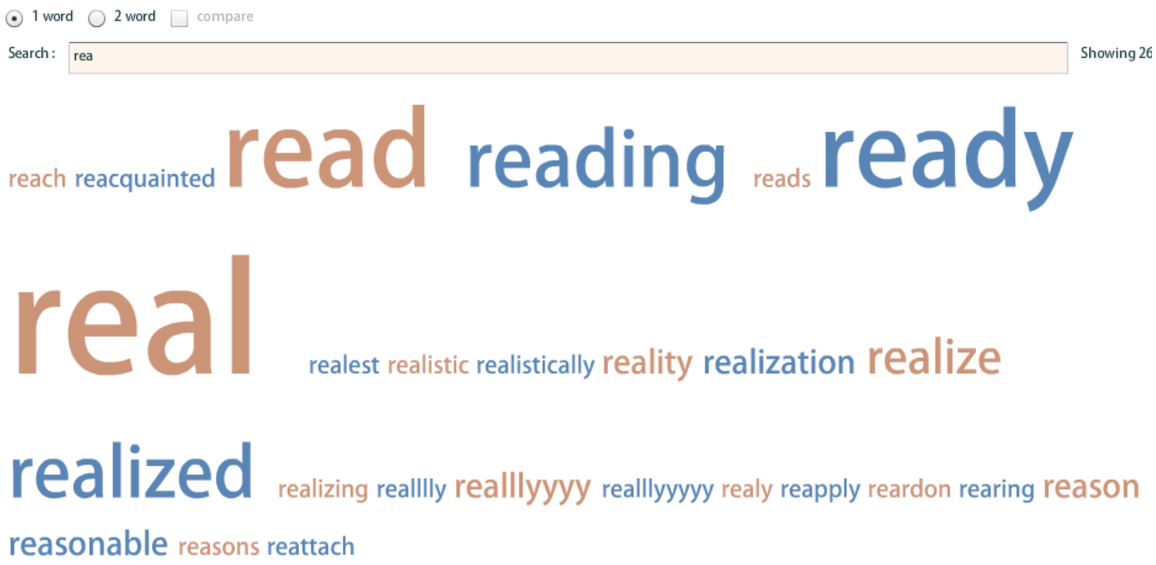

● 1 word ○ 2 word □ compare
Search :
Showing 200 Out of 6712
1 10 2 3 4 5 7 8 abigael abs alright awesome babe baby back bad bar bas bc bday bed beer big bit bring brunch btw buy cab call called
city coke coming cool cute cuz date day dead dinner doin dont drink drinking drinks drunk dude dunno dying early eat eating eli end
ep face feel find fine flight food friend friends fuck fucking fun funny game gay girl god gonna good gotta great guess
guy guys haha hahah hahaha hair hang happened hard hate head hell home hot hour hours house idea idk im jesus jist kelly kind kinda late
leave leaving left lemme life lol lolll lollll lolol lololol long looked lot love made make man maria meet miss mom moms morning
move movie new nice nick night omg order party people phone pic place plan plans play playing pretty prob put read ready real room sasha
season sex shit show sick sleep soph sounds sry start stay steph stop talk text thing things tho thought tight til time today told
tomorrow tonight totally train trophy true tv ugh ur wait walk wanna watch watched watching week weird
word work worst wtf ya yea yeah yo

Figure c.1



● 1 word ○ 2 word □ compare

Search : rea

Showing 26

Figure c.2

Visualization C is less successful overall than visualizations A and B. One of the standout distinctions is the lack of purpose behind the colors chosen to represent the data set – the words simply alternate between pink and blue, with no variation in shade. All that comes to mind when viewing this color scheme is a visualization of baby names, with blue representing boy names, and pink representing girl names. In addition to the unhelpful colors, visualization C also fails to offer an overview of the entire data set, or the ability to zoom in. The viewer must utilize the filtering feature in order to dive deeper into the data. Only by looking at words that start with 'rea' for example (as shown in figure c.2), can the user zoom in. The only successful aspect of visualization C is its ability to provide details-on-demand. If the user hovers the mouse over a particular word, a pop up box appears with a list of every occurrence of that word, as well as the sentence in which it was used – the more occurrences, the larger the word. Other than this however, visualization C offers little else. There is no option to view the history of searches within the visualization other than just clearing the search field to start over. And again, the user is not able to extract sub-collections or the query parameters.

Visualization D: Word cloud

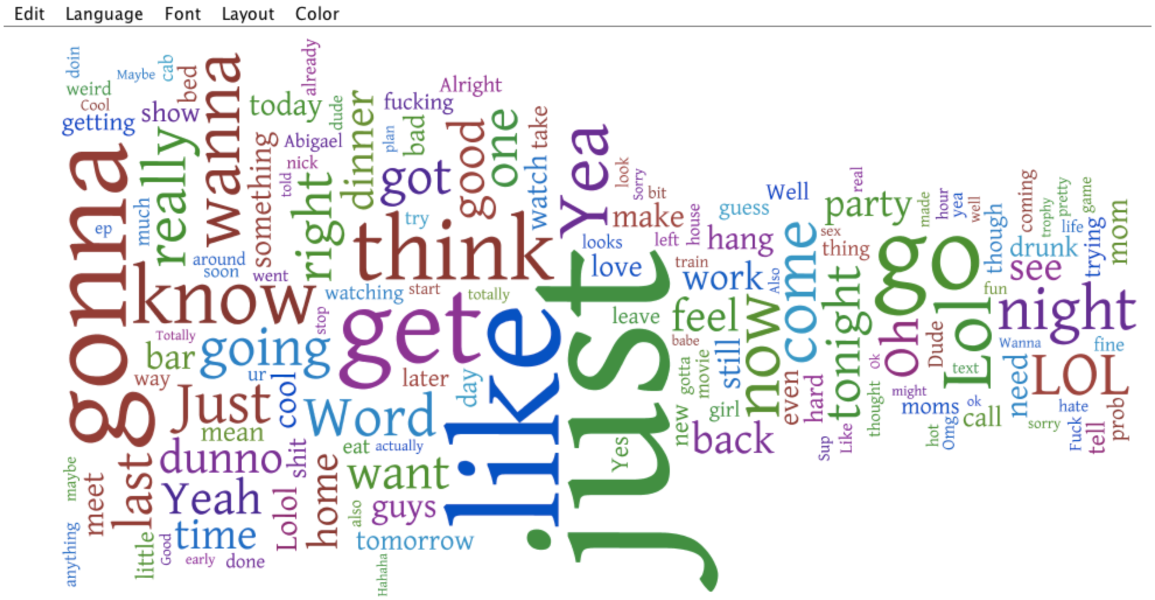


Figure d.1

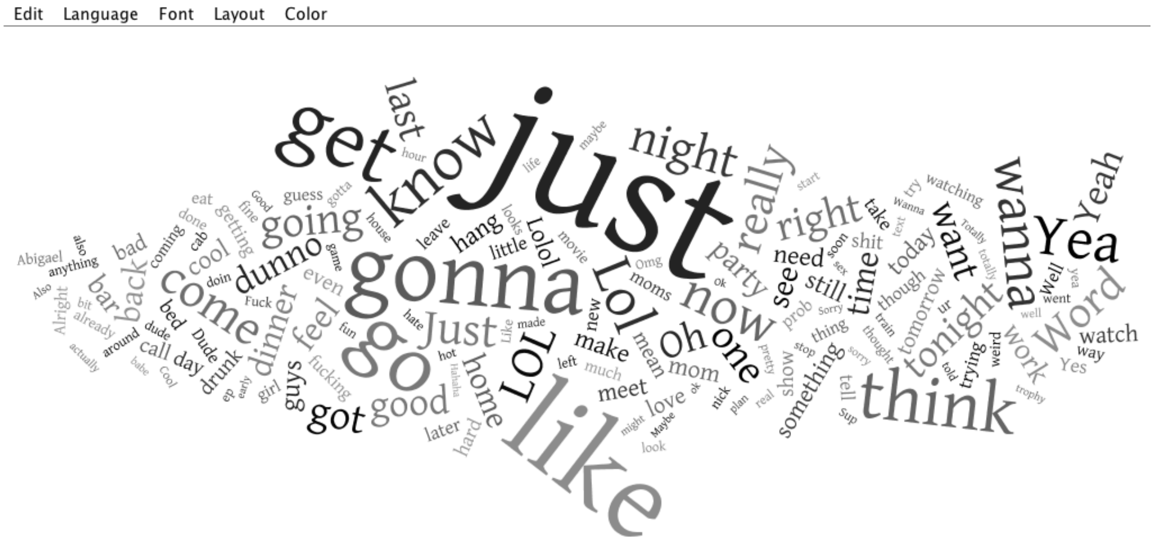


Figure d.2

Visualization D is even less successful than visualization C. Once again, the only visual representation of the data is that the more a word occurs in the data set, the larger it appears in the visualization. Like in visualization C, the color choices in figure d.1 do not aid in the representation of the data set. The color scheme is an attempt to make the visualization aesthetically pleasing, but if the color choice serves no purpose, it is irrelevant. The user does however, have the option to change the colors, and in figure d.2

you can see a more useful choice of colors. In this representation, the more the word occurs in the data set, the darker it appears. Visualization D offers no way to view the overall data set, zoom in, filter, not does it offer any details-on-demand. The only interactive aspect is that the user can choose a general shape they want the word cloud to take. The information does not relate to itself other than in size, there is no historical aspect other than 'undo' and 'redo', and again, the user is not able to extract sub-collections or the query parameters.

After examining all four data visualizations, it is clear not one of them is a perfect visual representation of the SMS messages. However, different aspects from visualizations A, B, and C can combine to create a semi-successful visualization based on Shneiderman's criteria. The one aspect missing from all four visualizations is the ability to extract sub-collections or the query parameters. In *Visualizing Data*, Fry provides an explanation for why the visualizations I created were not completely successful in representing the SMS messages: "There are dozens of quick tools for developing graphics in a cookie-cutter fashion in office programs, on the Web, and elsewhere, but complex data sets used for specialized applications require unique treatment" (Fry, 2). The IBM program "Many Eyes" is in fact, one of these quick tools, and therefore these visualizations are not uniquely tailored to the data set at hand. That is not to say however, that they were not helpful in understanding what is required to successfully visualize a large set of data. Through the compare and contrast of these visualizations, as well as the writings of Shneiderman and Fry, I have learned what factors, from their presence and their absence, are necessary to create a successful data visualization.

Works Cited

Fry, Ben. "The Seven Stages of Visualizing Data." *Visualizing Data*. Sebastopol, CA: O'Reilly Media, 2008. Print.

Many Eyes. IBM Research and IBM Cognos Software Group. Web. 19 Nov. 2011. <http://www-958.ibm.com/software/data/cognos/manyeyes/>.

Shneiderman, Ben. *The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations*. University of Maryland. Web.