Data Project Report
futureme

Esther Cho
Esa Eslami
Tommy Keswick
Alberto Pepe
Pinar Yoldas

Concept

Originally the dataset was chosen before the concept was worked out. The site futurme.org is a pretty basic web service that allows you to write yourself a letter to be delivered at a later date, up to 50 years in the future. They currently host nearly 400,000 letters, about 30,000 of which are public. By writing a letter to your future self you might, for example, remind yourself of something you are scared of forgetting. or perhaps, you might remind yourself of something you are trying to forget and you want to check on your progress. We would like to capture and display instances of remembering/forgetting patterns in daily cyber/meat time/space. Futureme is a good place to get some data from and play around with it. Initially we scraped some data, calculated word occurrences, filtered out some major stop words and obtained a cloud of recurrent words. Words like "remember" and "forget" show up in the top.

From the 30,000 emails sent to your future selves, we selected the ones in which terms like *remember*, *remind*, *forget*, and *forgot* were highly recurrent. We ended up with about 7,000 emails.

The overall concept of the project is to attempt to give viewers a "sense" of the emails from futureme.org without them having to read all 30,000 public messages. The particular visualization we have chosen to display for the class was arrived at after the data were analyzed for certain trends and with certain parameters in mind. The futureme.org dataset is quite rich, so this visualization represents one possibility that we were able to complete within the time constraints of the project.

Method

The public emails were collected from the site using a python script. The contents were then analyzed in various ways. More python scripts were used to make lists and counts of word frequency, length of time between composition and delivery date, and word count of individual emails.

To show one possible way of analyzing the data about word frequency, the most common words were classified into positive and negative meanings. For example, *love* and *happy* were given positive values while *hate* and *sad* were given negative values. That this method completely removes context is recognized, but is still useful to show methods of classifying data such as these. With the lists of negative and positive words an overall "mood" was determined for each of the emails based on ratios of the occurrence of the different words. An R script was used to calculate the ratios.

Another way that was used to look at the data from the emails was through the distribution over time. Time lags vary from 1 day to 50 years. Email volume spikes at 1, 6, 12 months and then goes down considerably. The time periods were broken down into months so that there are up to 600 possible periods the emails fit into. By plotting these occurrences, trends can be seen about when users are trying to remind themselves of something.

R was then used again to make visuals out of both the mood and the time. We then assigned values to colors (dark red: positive, pink: negative). Why pink? No particular reason: we want to represent these colors using balloons. Red balloons seem to be available in more shades than others. To aid visualizations we divided data from the 7,000 emails into 99 sections. Each section will express the mood for about 700 emails and hopefully we'll get some sort of pattern when looked at as a whole.

Setup in Physical Space

We wanted to come away from this project with more than an slideshow presentation on a projector. By using balloons and making the visualization very large and out in public, we hoped to express some of what the emails expressed by just using the data from the whole set.

After much deliberation and shifting of ideas, the final presentation will be set up using fishing line, four shades of balloons from red to pink, large nails, printouts of the R data visualizations, and text from some of the emails. All of this will be set on the grass in front of Broad Art Center using the fourth story balcony railing as well.

One balloon will be attached to one length of fishing line for a total of 99 balloons and 99 lines. Each line will attach at a single point on the railing of the fourth story balcony of Broad. The other end of the line will be nailed into the grass at varying distance from the others.

To determine the placement of the nails in the grass, as well as the placement of the balloons on the line we are using the time distribution of the emails. Since most emails were sent to the recent future-- 1, 6, and 12 months-- there will be clustering towards the beginning of the bunch. The lines will attach to the ground closer together and the balloons will be closer together and higher up on the lines. It will sort of resemble a large 3-D histogram. The few emails that get sent much further out in time will be further away on the grass and their balloons will be lower to the ground representing a smaller number of emails.

Attached to each of the lines at a height noticeable to passersby will be a printout of the R visualization. Written on top of this visualization will be a representative phrase from an email. For example, *please, forget about the past. move on.* or *remember: life is a waste of time, time is a waste of life, get wasted all the time, and have the time of your life.*

Preservation Information

To create again:
Python version 2.4.0
    process.py happysad.py tokenize.py file_builder.py cloudify.py wordfreq.py
R version 2.5.0 with gplots, gtools, and gdata libraries
    future.r getcolors.r getmeans.r plotloop.r
text editor
futureme.org website with same structure *or* emails_remembering_forgetting.txt

To preserve data:
text files containing month distribution
text files containing email snippets
JPEG files for preserving R visualizations