Advanced Data Vis Project Progress Report

Maks Cegielski-Johnson
u0836524

April 1, 2018

1 Current Site

The current website can be found at

`https://makscj.github.io/adv-vis/final/website/`

There are four interesting components to interact with, from left to right:

- T-SNE embedding of GloVe vectors, with interaction possible with each word. Clicking on a word adds it to a collection of "selected" words for futher investigation. Any selected words will be highlighted in the the 2-dimensional plot discussed next. This svg can also be zoomed/panned for easier exploration.
- A 2-dimensional visualization of a pair-wise plane in the original 50-dimenisonal GloVe vector space for the words. This has a joint interaction with the "rubix" controller discussed futher below.
- A search box that allows the user to search the T-SNE space for a specific word, dimming all points except for the result, and flashing the result point 5 times in red.
- As mentioned above, the "rubix" controller which represents the cartesian product of all of the dimensions in the original GloVe space. Changing the selected square will update the 2-dimensional plot mentioned above. Each square can be clicked on to show a different slice of the original space. This can also be controller with arrow keys for easier interaction.

2 Project Summary

1. **What is an overview of your project?**

   Visualizing word vectors with high dimensionality techniques.

2. **Why is the project worth pursuing?**

   Word vectors are commonly used in NLP research, but since the dimensionality of vectors is quite large, it would be useful to have a tool that allows for easy exploration of this space to collect insights.

3. **What are your project objectives?**

   Create a visualization that allows user interaction to explore the space of high dimensional word vectors.

4. **What are the questions you would like to answer?**

   How can visualization be used to successfully determine whether two words are similar, or what the degree of similarity is between two words.

5. **What data will you plan to use?**

   Word2Vec and GloVe vectors, given enough time, perhaps WordNet similarities.

6. **How can we evaluate how successful your project is once it is completed?**

   Since this tool is rather exploratory, if correct insights can be made with my tool. If my tool shows two words being similar that aren't similar, is there an underlying similarity that isn't clear between these words, or is my tool incorrect.

3 Timelines

- ~~Week 1 (March 5) - Project proposal~~
- ~~Week 2 (March 12) - get the data, prepare it for visualizing. set up Python frameworks~~
- ~~Week 3 (March 19) - try out different HD visulizations~~
- Week 4 (March 26) - try out different HD visulizations
- Week 5 (April 2) - try out different HD visulizations / try out TOPO visualizations
- ~~Week 6 (April 9) - make website / d3 visualizations~~
- Week 7 (April 16) - make website / d3 visualizations
- Week 8 (April 23) - finish website / prepare presentation, report

3.1 Timeline Status

I have made some D3 visualizations earlier than originally planned because of some of the fruitless outcomes of trying some HD visualizations. I appear to be on track, and I just need to continue exploring the HD visualizations we discussed in class to see what other insights I can provide to the data.

This task has been more challenging than I thought, because I haven't been able to find a mapping of the T-SNE data using keppler mapper, and because the original dataset has 50 dimensions it is harder to visualize these using parallel coordinates or a scatter-matrix.