Advanced Data Vis Project Proposal

Maks Cegielski-Johnson
u0836524

March 7, 2018

1 Team Members

There is a paper I wrote with Vivek you might want to look at:
Visual Exploration of Semantic Relationships in Neural Word Embeddings. Shusen Liu, Peer-Timo Bremer, Jayaraman J. Thiagarajan, Vivek Srikumar, Bei Wang, Yarden Livnat and Valerio Pascucci. IEEE Transactions on Visualization and Computer Graphics (InfoVis 2017), to appear, 2017.
http://ieeexplore.ieee.org/document/8019864/

Maks Cegielski-Johnson

u0836524

2 Introduction

I would like to explore the space of visualizing high dimensional word vectors, such as those generated by GloVe or Word2Vec. The motivation for this project is two-fold for me:

- Understanding word similarity is useful for my thesis project
- I want a project that is well defined, allowing me to focus on the actual visualization rather than having to try dealing with the data. In past projects, I have picked a data set which made the hardest part of the project not the actual visualization, but making sure the data was clean and easy to use. Instead of doing this again, I am picking something simple and fun to ensure I can focus on the visualization.

3 Project Objective

We have discussed many different high-data visualization techniques, and I would like to explore the application of these techniques when used in conjuction with word vectors.

There are many techniques to work with; Please think about specific ones you want to focus on.

4 Data

When Googling "word vector visualization" the first link I came across was a Kaggle competition to visualize Word2Vec with T-SNE. I can use the dataset from this competition to visualize with T-SNE, as well as extending the dataset to other high dimensional techniques.

I also have access to the GloVe vectors through the NLP research group at the University, so I can perform the same techniques on these as I will with word2vec.

5 Background

Googling "word vector visualization" brings up a lot of results of using T-SNE for the task, but looking at the outcomes of these visualization, many look messy and not very useful.

If you want to use t-SNE, please think about how to evaluate the effectiveness of the technique/quality of the results.

There has also been work done in the concept of $king - man + woman = queen$ with vectors, so I would like to extend this task to a visualization.

6 Technical Contributions

I would like to try as many different HD visualization techniques on this task as I can, more than just T-SNE which seems to be the state of the art. Perhaps even seeing if there is a way to create a topology in order to use the techniques we dicussed in the TOPO lectures.

7 Expected Outcomes and Deliverables

I expect to create a website which would allow a few different views for word2vec. Firstly, perhaps a general view of the entire space, and secondly, the ability to query two different words and determine visually how similar they are and in what dimension. This would hopefully allow different insights into the vector space through exploration. I could also try incorporating WordNet in order to create topologies in the space that allows more insights than just the raw space and queries.

Vivek and I have talked in general of applying mapper to NLP data;
you could explore mapper-like techniques.If interested, we can chat more after class.

8 Evaluation

Since this area is rather commonly explored, specifically with T-SNE, I will explore what other people have done to visualize word vectors, particularly in research papers. I will also see if there any useful insights I can see that would be helpful in my thesis, as well as compare similarities in my visualization with those provided by WordNet.

9 Proposed Methods

All of the methods discussed in the HD component of the class, and if I can create structures using WordNet, then I could try using some of the methods discussed in the TOPO component.

10 Software

I plan on using Python, specifically all of the libraries provided in Anaconda (NumPy, SciPy, Pandas, etc.) as well as kepler-mapper from the first project. For presentation, I would like to take any insights I get from Python and either set up a server to present these in a website, or export the data to visualize with D3 in a simple HTTP server on Github.

11 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

12 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.