Interactive Tool for Exploring Word Embeddings

As with many of the concepts we have covered in this class, I have been equally intrigued and confused when we learned about word embeddings. After the completing the assignment relating to word embeddings, I felt that I had a better grasp on the topic, but I still feel that it is one of the more complex concepts in computational linguistics. Because word embeddings are used in numerous NLP tasks, I thought it would be interesting to create a tool that would help people explore word embeddings, including how they are formed, how they change when the vocabulary is expanded, and how they can be used in other settings. My main focus would be creating this “embedding exploration” tool for the English language, but I would be interested to expand it to other languages if time permitted.

To complete this project, I would first need to become more familiar with word embeddings. I used Polyglot and GloVe for the word embedding assignment, so I would need to deepen my understanding of these tools as well as experiment with others in order to find the best engine to use for the exploration tool. Depending on the scope of the project (which I’m still trying to map out), I might end up writing my own wrapper program for one of the aforementioned embedding tools. As for “embedding exploration”, I want to give the user the ability to either copy-and-paste in a vocabulary file, upload a vocabulary file, or enter custom vocabulary. A custom vocabulary wouldn’t be ideal for an actual language model, but I think this approach would facilitate a greater understanding of word embeddings. For example, I could enter the word “tree”, and the embedding would be essentially useless. However, if I entered “tall tree”, it would be interesting to see how the original embedding for “tree” changed as the vocabulary expanded. Other GUI features would allow the user to quickly explore word embeddings.

This exploration tool would require a couple different technologies. The main focus would obviously be a system for creating word embeddings, which as I mentioned already, would probably be chosen as I conducted my research. Central to the tool would be some sort of GUI, so that users could enter their own vocabulary and watch the word embeddings form. For this to happen, I would need some sort of GUI framework. I might use a Python or Java GUI framework, or maybe even use HTML/CSS with a Python/Java backend. I’ll need to do more research on the available options, but I know that I could use Polyglot or Gensim with Python, and I believe GloVe is implemented in Java, so I should be able to find a way to interact with the embedding tools, and create a tool for exploring word embeddings interactively.

Throughout the project, I will probably use some pre-made language models that come with Polyglot or GloVe, especially during the testing phases. Other “knowledge sources” might not be necessary, since one key feature of the exploration tool will be observing how embeddings are formed from user input. At this point, I also don’t know exactly how system will be evaluated. Most of the “evaluation” will be simply whether or not the system works. Since I won’t necessarily be creating a new tool, I won’t need to check whether the output is accurate, rather I will just check if it is consistent with the output from the existing system.

I think this project will be very interesting, and I’m excited to start working on it. As I brainstormed all the different concepts we covered in class, word embedding is the one that stuck out to me the most. As I considered it more, I came up with the idea of making it more accessible to people by creating the “embedding explorer.” This project might be different than others in the past, so I’m open to any feedback on my proposal. I want to meet the requirements of the assignment, while doing something I find interesting. There are a lot of questions and details that I still don’t have the answer to, but I think that will make the project rewarding.