NEO4J in Linguistics

Project Abstract: Julie Blackburn

Virtually everyone attracted to computer science has an idea of what IBM Watson is – a breakthrough question and answering system which extensively uses Resource Dedication Framework (RDF) in its programming. Other IBM projects access similarly conceived “linked data,” albeit not always through graph databases. There have been some notable projects where the AI will train from social media circles fueled by genuine human interaction, (such as Microsoft’s Tay) and the reflection of ourselves can be quite frightening in how the AI trends to hateful speech – a phenomenon so prevalent on sites like Twitter and Facebook.

Let’s put a twist on this concept. As someone who has a Masters Degree in English, the prospect of being able to harness the power of Big Data and use computer science to expeditiously analyze bodies of literature - and in turn, draw conclusions about our culture or raise other hypothesis with these powerful technologies, is absolutely irresistible.

Starting with a universally well-known author, with a large and familiar body of work - William Shakespeare. Shakespeare’s entire recorded life’s texts were available from the Gutenberg Corpus. Going forward, one could combine entire bodies of literature for a given year, or for a selected genre over a range of time, or compare and contrast genres or decades of literary works - making the Big Data implications even more palpable. (For example, I would have found it interesting to see how a concept like violence trends in an author like George R.R. Martins works as he has a popular series which becomes progressively more gruesome with each novel; and followed its popularity by a scale like Amazon ratings and reviews. However, these texts were not available in something like the Gutenberg Corpus.)

In my project, I am combining two technologies in a way that had not been covered in our Big Data class, NEO4J and Word2Vec. I was able to train data with Python scripts and then create a very dynamic visualization, linking sentence structure off of a root word. The idea was the follow a “Subject-Predicate-Object” (SVO) linguistics sentence structure – something that could readily translate to a database about known facts and relationships using the Resource Description Framework (RDF).

The challenge to my research, was being that the subject matter I chose was so figurative and poetic, rather than literal, the connections were not always as clear (even in a graph database) as they could be. The result was beautiful, and powerful in that the eyes don’t just glaze over something without any idea of the relationship between the words, (made possible with training from Word2Vec, and visualized by NEO4J) as they would in a spreadsheet though. I believe the ability to use these tools in tandem with each other could be very powerful for future research, including literature, product reviews, and of course social media trends.

<https://youtu.be/HXZZWQcvhOY> 2 min <https://youtu.be/cfmB_y_I5Dg> 15 min