Project Design

My initial project design was a little too ambitious it seems. I think my fatal flaw was that I didn't do enough exploratory analysis of my data before diving in. The reason the translator failed was because my data didn't provide clean translations line for line like I assumed it would. I only realized this until I had built a sequence to sequence model that started outputting garbage.

Tools

Gensim - text preparation and word2vec model

Pandas - data manipulation

Sklearn - Clustering algorithms

Data

I scraped a large amount of data from No Fear Shakespeare that I ended up not using. My final datasets came from the Kaggle open data sets and Project Gutenberg's text archive of Shakespeare's sonnets. All in all, I had about 120,000 lines of text from ~30 plays and ~120 sonnets.

Algorithm

I created a word2vec model that created word embeddings based on all of Shakespeare's works. This embedding helped me get context around old english words and the characters of his plays. I used a skip-gram model and using a window of 10 to create 100 new features for every word.

I used those features to cluster the characters together using k-means. I considered some other clustering algorithms briefly, but ended up sticking with K means. I didn't use DBSCAN because

I wasn’t interested in identifying outliers and I didnt use hierarchical clustering because the features were impossible to understand. Even if there were hierarchical relationships between the clusters, they wouldn’t have been interpretable.