Using Natural Language Sentiment Analysis to Monitor Changes in Edgar Allan Poe's Mental Health Through his Prose

Problem Statement

Efforts to programatically determine sentiments in written text have become an increasingly prevalent fixture of machine learning study. While this technology has wide ranging applications, the insights into public health and psychology are perhaps its most beneficial use. Using the works of Edgar Allan Poe as a test case, I hope to demonstrate the viability of natural language processing and sentiment analysis in tracking mental health decline. By looking at someone whose life is well studied, we can measure the accuracy of our sentiment analysis against known events in Poe's life.

Data

- I used eapoe.org as my source for Poe's stories.
  - This website contains detailed accounts of the publication history of each of Poe's stories and poems.
  - I focused on Poe's stories because I was concerned that the stylization of Poe's Poems might throw off the sentiment analysis.
- I manually copied all of his stories into individual text documents and minorly edited the documents to fix any minor formatting issues.
- The text files were named with an abbreviated title and the three letter month and year.

Methodology

- I zeroed in on Poe's wife Virginia first showing symptoms of tuberculosis as a likely inflection point in Poe's life.
- As a proof of concept, I first used a four layer RNN model from the Google Tensorflow website. This model was fairly overfit and produced wildly inconsistent results with the Poe stories between runs, so I moved to using TextBlob.
  - TextBlob is a simplified subset of the larger and more in-depth Natural Language Toolkit library. TextBlob greatly reduces the complexity of setting up a model and training embeddings at the cost of low visibility on how it does so.
- To process the data, I went through each text file of Poe's stories and split the name along the dash to keep the title in dataset while converting the month and year into a datetime object. The contents of each story then had both their sentiment polarity and sentiment subjectivity recorded. In each of these graphs I plotted the raw data in blue and the average per year in orange.
- I also plotted sentiment polarity against sentiment subjectivity and used a color spectrum starting at blue and ending at red to see if there was any combined relationship between the two variables.

Results and Challenges

The TextBlob model showed no indication that there was a distinct shift in Poe's polarity across his works. My prediction that Virginia developing TB seems to be supported by the graph of the subjectivity of Poe's works, as that date seems to roughly mark a sudden shift to more objectivity. This result could mean that Poe having to face the mortality of his wife influenced him to be more grounded in his writing. There was no apparent relation between polarity and subjectivity

In the future, I should try and move away from TextBlob. Although easy to use, it is opaque and its embeddings are likely not designed towards 19th century text. I should also look at specific words as opposed to total stats.