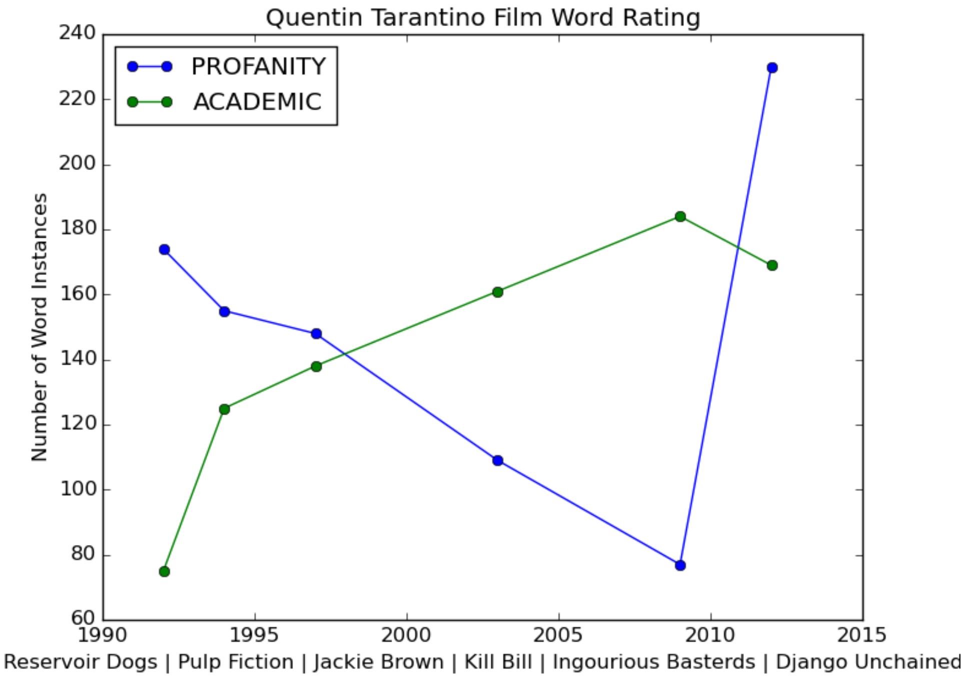
Keenan Zucker -- Text Mining Project -- Software Design 2015

Project Overview: For my project, I decided to look at scripts written by Quentin Tarantino, my favorite director. Tarantino is known for films that are often very violent and R-rated. I wanted to examine certain elements of his films using the scripts of the films. I used 6 different Tarantino films, looking at profane words, academic words, and average word length for each film, to try to determine trends of his style over the last twenty years.

Implementation: I had to start by finding the texts. I found scripts online, then parsed the HTML code for the script text, and saved each one as a .txt file. Once I had the files, I combined all the scripts into a single list, with six elements, one for each film script. I then have functions for each of the categories of analysis. I used a 'for' loop to go through the scripts and find swear words and academic words using a wordlist from pattern.en. I also calculated the average length of each word using simple math in a separate function.

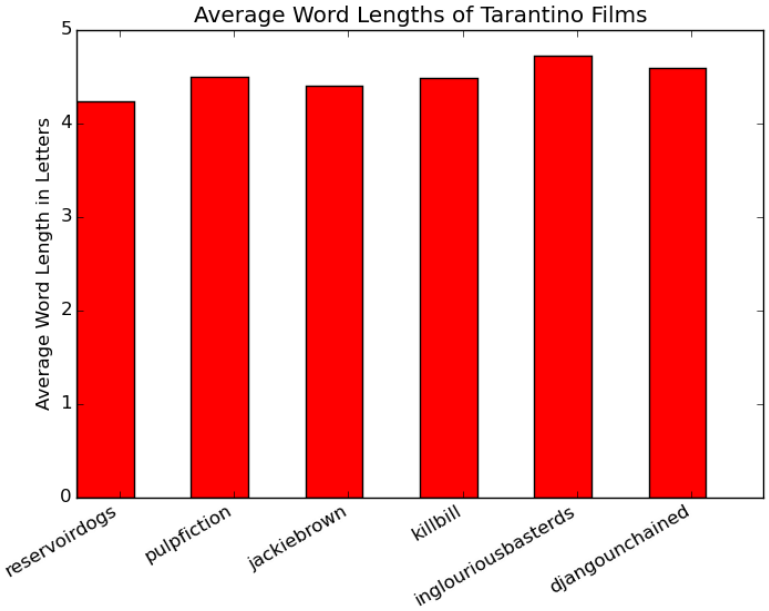
I considered creating a dictionary to store a value of the number of swear words for each script. However, I decided against this choice because I subsequently wanted to find more information about the words, like academic quality and average length, so I thought it would make more sense to create a list for each of those counters as well. I also have a better grasp on manipulating lists than I do with dictionaries.

Results: The results I achieved were pretty interesting, in my opinion. These movies were created over a twenty year period, so it was fascinating to see how his films have changed with regards to language choice over the years. This first graph shows the number of profane words and the number of academic words over the course of six films, and about 20 years. It is interesting because the trend for the first five films is downward in profanity and upwards in academic words, signifying and increased integrity of films, perhaps. However, for his most recent film,



Django Unchained, the opposite is true, as the films is has significantly more profanity than any of this other films, with a whopping 230 swear words!

My third study looked at the average word length of each of his films. As shown, the general trend of the word length is increasing, possibly correlating to the average increase in academic language in his films. However, this test doesn't give much insight into the specifics of the films and their merits, it simply examine the scripts. It cannot account for the cinematic quality of the films.



Reflection: I wrote this project in an interesting way. I first wrote it all as a single script, before breaking it up into functions. I think it would have been easier if I had been more conscious about splitting it up earlier, to avoid doing extra work. I think the implementation of the code worked pretty well. The only problem I had was with working with the 'sentiment' part from pattern.en. It would throw back errors because of unicode conversion flaws. Some of the scripts ran fine, but others gave errors, even when I used 'plaintext()' so I decided to give up on doing that part of the analysis, and worked on average word length instead. I learned a lot about plotting and the pattern toolbox, which was very useful and I think will be useful going forward. In the scope of only one week, I thought this project took an appropriate amount of work for me, and I had fun doing it!