Rachel Yang

Software Design

25 February 2016

Text Mining Analysis

Film Word Clouds

Project Overview:

I utilized movie subtitles and images from Google Searches to construct a word cloud for a specific film. Showcasing the most common words used in its script, I also overlapped the word cloud onto an iconic image and matched the font colors to the film's color scheme. I had hoped to create a minimalist but engaging movie poster that spoke to devoted fans and followers of these films.

Implementation:

The first thing I had to do was download the movie subtitles. I originally wanted to be able to format this project so that you could do this with any input movie title, using google search or some movie script database api, but to get started I decided to download the movies I wanted to create a word cloud for and, given more time, would expand. Then, putting those movie subtitles into a list, I parsed through the list and used the file options to read through the old file and save a new file of the edited movie script without the time stamps from the subtitles. After that, I parsed through the words to reduce the script into the simplest dictionary of valuable and necessary words.

From there, I used an iconic image from the movie to generate the proper font colors for the word cloud and the approximate shape for the word cloud to take. Using matplotlib and PIL's Image, I generated and displayed a word cloud for each edited film's subtitles.

I had wanted to implement more doctests and unit tests, but wasn't sure how to go about it with most of the functions in this project, so I chose to not include them in those functions. For the ones that I could expect a certain result, such as with the edited file names, I included them.

I also originally wanted to search for the images using the Flickr api or Google Image Search api given a certain keyword that included the film's title, but I got roadblocked by trying to figure out API license key issues and decided that I would simply download the images myself and work on expanding my project to a larger scope at a later time.

Finally, I had wanted to create the word cloud using a histogram function, finding the most frequent words found in the subtitles and then using that to determine word size, so on and so forth. But given the time constraints, I chose to implement the WordCloud package.

Results:

Movies have many identifying features that people who have watched them or who have seen their marketing are familiar with. Color schemes, like with Kill Bill's characteristic yellow and black (from Uma Thurman's suit) and Frozen's ice blue and blonde yellow help people recognize the film. But also, iconic symbols from the movie help too: for example, Finding Nemo's word cloud is shaped into its leading characters Marlin and Dory and Lord of the Ring's word cloud is shaped into a ring. The Matrix's word cloud has its iconic red and blue pill. Star Wars's is shaped into a stormtrooper helmet. But most of all, the best thing that helps to identify films is the words, namely character names. Some of the word clouds didn't take to the image's shape as well as the others, but one could still tell what movie it was based on the names like Doc and Marty for Back to the Future and Harry and Dumbledore for Harry Potter.

Movies do a fantastic and surprising job at marketing their films as icons. The fact that we can recognize a film of two or more hours worth of spoken words by a couple of buzzwords scattered over some characteristic image or color scheme is crazy.

On top of this, the word clouds help us to see what the key themes are in the film. One of Mulan's bigget words is family. Frozen's is love. High School Musical's is team. The Godfather's is business and family. We can summarize what the movie represents and speaks for, as simply as looking at the most common words in a script.

Color Scheme:

Okinawa
Hey
good
now
never
name
may
Tanaka
friend
know
girl
gonna
Black
go
want
one
big
Keep
Yeah
will
Okay
year
head
time
bitch
tell
shit
say
said
Japanese
right
sword
baby
thing
room
God
Uh
Bill
Four
Sofie
Oh
kill
Ren
Charlie
come
yes
sake
Wiggletoe
American
Look
half
got
still
give
even
Boss

sister
one
Arendelle
Wait
Anna
go
Let
Now
look
open
Queen
way
Olaf
forever
Princess
Oh
know
mean
Hans
want
true
Yeah
Elsa
see
ice
summer
got
Please
time
back
love
will
heart
right
Sven
Come
Okay
Kristoff
need

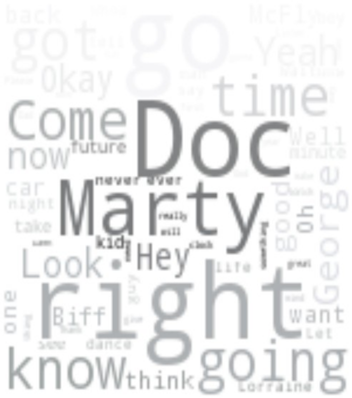
Symbols:

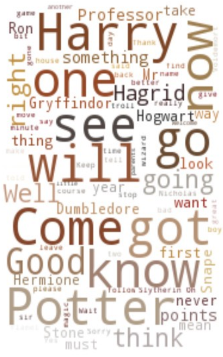
good
now
need
mean
want
right
us
know
going
Morpheus
Matrix
world
Neo
believe
one
go
time
tell
got
Mr

now
ring
must
see
Gandalf
Bilbo
Baggins
go
us
one
come
know
Frodo
Mordor
never
will

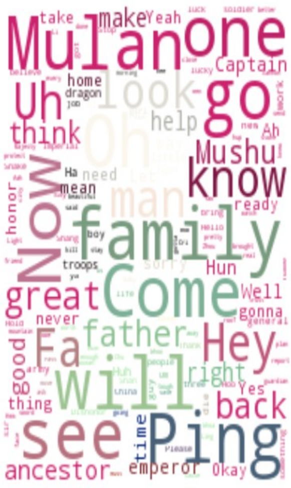
need
back
Come
see
got
going
droid
go
will
us
know
First

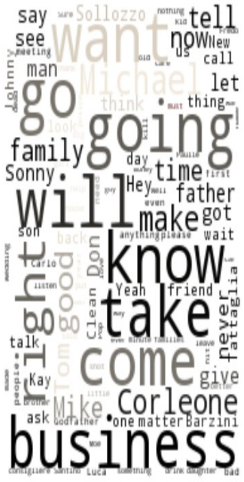
Character Names:





Theme words:





Reflection:

The editing of the movie subtitles went very well. By saving it as a new file, I was able to see how well it reduced the original script into the necessary words I needed to create my word cloud, and every time it did its job. At the same time, there are many things I would do to improve this project given the time. I would first of all make the word cloud mask more accurate to the image, fitting close to the lines to make the iconic image more clear and defined even with the words laid on top. I also would expand my code to be applicable for any movie and to search for any corresponding image. It's very limited as it stands now, only analyzing specific movie subtitles and images downloaded by the user. And I would utilize more docstring tests/unit tests.

Going forward, I will choose a less ambitious project. I was so focused on creating something impressive that I got in over my head and ended up with a project that didn't exactly fit into what was assigned. I also spent too much time trying to pick a solid idea that I didn't have enough time to fully complete my project. With that, I wish I knew that image search api's would be difficult to implement into my project the way that I wanted it to; I wish I knew to pick an idea earlier to give myself more time; I wish I knew that it would have made it easier to talk to professors or ninjas more throughout the process.