Cory Clayton (acc2ds@virginia.edu)
DS5001
6 May 2021

Exploratory Analysis of Movie Scripts from the Disney Franchise

Introduction:

Movies have had a central role in modern culture since they arrived in the early 1900s. They allow people to be transported all around the world to both real and fictional places, and help shape the modern culture as well as reflect its values. The goal of this project is to use exploratory text analysis techniques and apply them to movie scripts from the Disney franchise. This is not an exhaustive study, as the acquiring of source material was quite labor intensive, but more of a proof of concept that these techniques, many of which were designed to be applied to more traditional texts, can be used to provide insights into these movies.

Acquiring and Processing:

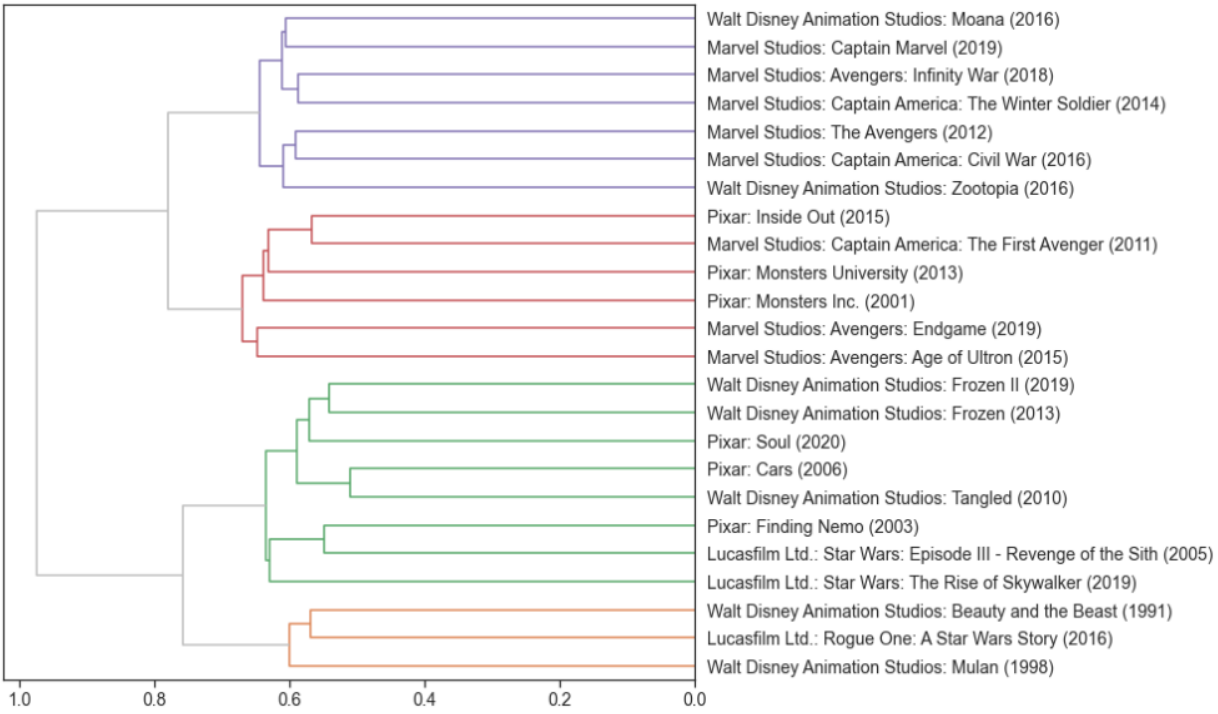
The corpus includes 24 movie scripts from this Disney franchise, 3 Star Wars movies, 8 Marvel Comic movies, 6 Pixar movies, and 7 Disney Animated movies. The movies can also be split into the live action movies from Marvel and Lucasfilms, and the animated movies from Pixar and Disney Animated Studios. These scripts were webscrapped from the webpage https://transcripts.fandom.com/wiki/Transcripts_Wiki. All of these scripts were transcribed by fans using inconsistent methods which meant every new script needed to be inspected and processed individually, as well as quality checked at the end to fix any character speaking errors.

The ordered hierarchy of content objects (OHCO) used for this project had Movie at the top, then action, grouping all the dialogue after a given action description, followed by dialogue for each character that spoke, and sentence number and token finished off the OHCO. Ideally scene level data would be encoded as it is a good analog to chapter, where action has an inconsistent level often only grouping 2-5 lines of dialogue. I believe that the scene data would have been valuable in allowing the movie to be segmented in a size between dialogue lines and whole movies and some insights will be missed without that data, but this is a hazard of working with imperfect sources.

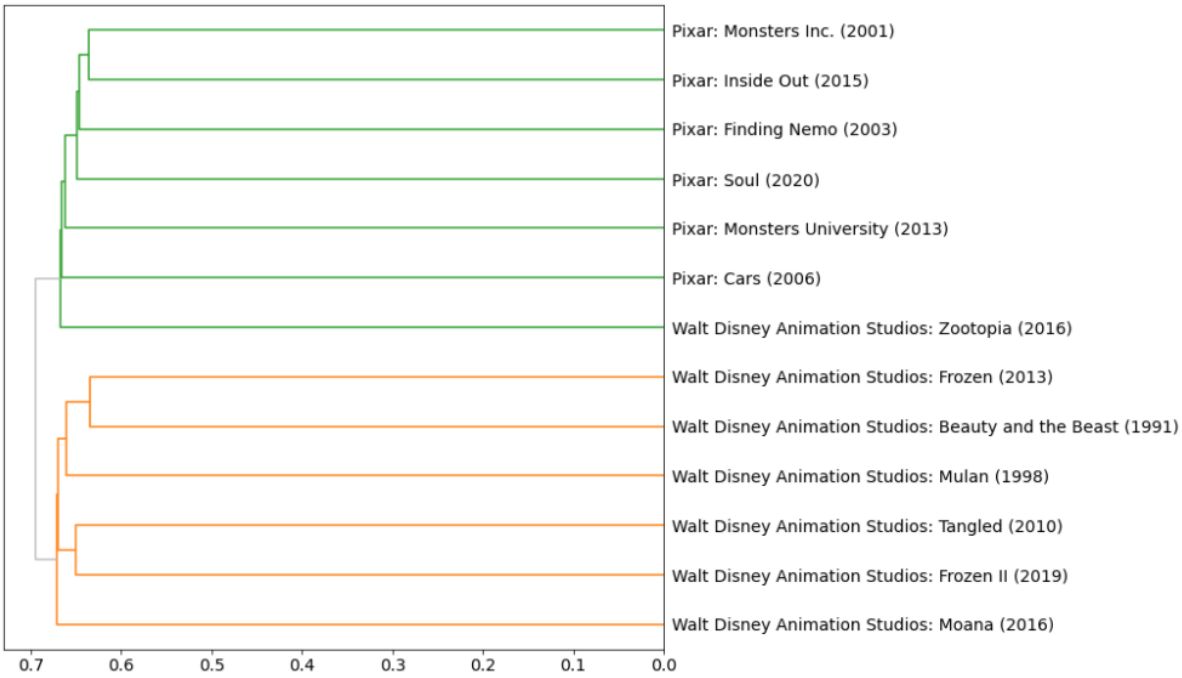
Clustering and Similarity:

Term frequency - inverse document frequency (TF-IDF) is a feature vectorization method widely used in text mining to reflect the importance of a term to a document in a corpus. Using the TF-IDF the movies can then be clustered using a variety of different clustering algorithms. The clusters can be visualized using dendrograms that represent the distance between clusters as the height of the connection.

This dendrogram is created using cosine similarity on all the movies. On the top half is all the Marvel movies, and the bottom half has the Lucasfilm movies, but the Pixar and Disney Animated split. Moana and Zootopia could be considered more action based animated movies which would make sense for them to be clustered together with the Marvel movies, but Mulan being in the other cluster questions that theory.



When the clustering algorithm is applied to only the animated movies using Jiccard's similarity the following dendrogram is produced. It groups all of the Pixar movies and Zootopia on one side, and the rest of the Disney Animated movies on the other. This makes sense as a human grouping and a non human grouping.



Principal Component Analysis:

The TF-IDF can also be used to try and reduce the feature space of a movie (or any other grouping) using Principal Component Analysis (PCA). This technique invented in the early 1900s is a surprisingly effective data science tool where each component captures as much variance in the data as possible that was not accounted for by the previous components. Looking at a graph of the movies plotted on the first two principal components it is clear that the first component is separating live action from animated movies. The second component is able to distinguish between Marvel and Lucasfilm as well as distinguish Pixar and Disney Animated, making it group the movies by Studio using only two principal components. Zootopia continues to be grouped with the rest of the Pixar movies as before.

I tried to find a clear distinction or genre groupings using the second or third principal components, but none were apparent.

Topic Modelling:

Using tokens and ngrams, Latent Dirichlet Allocation is able to group words that appear to be similar and important, and these groups can be thought of as a topic. When the topic model is applied to the whole corpus you can then segment the corpus to investigate which topics appear more prominent to the different subsets. I found that 40 topics was the optimal number of topics for this corpus using unigrams and bigrams as the features.

Here are the most prominent topics when the corpus is subsetted into studios.

Lucasfilms top topic is very spaceship themed, with other military language in the other top topics.

studio	Lucasfilm Ltd.	Marvel Studios	Pixar	Walt Disney Animation Studios	label
topic_id					
26	0.054242	0.030177	0.017305	0.016819	26 guy, ship, fleet, shield, right, space, yeah, gate, everybody, plans
14	0.037424	0.035538	0.042236	0.021845	14 sir, yes, yes sir, body, nemo, im, trade, youre, men, gonna
22	0.034989	0.025152	0.030525	0.029017	22 youre, minute, time, morning, droids, fine, oh, way, right, position
27	0.032916	0.024152	0.024003	0.019539	27 time, lady, memory, way, brother, isnt, dont, youre, car, core
32	0.032659	0.025353	0.023868	0.024696	32 power, wait, uh, people, wait wait, come, weapons, jump, cube, things
11	0.032451	0.028648	0.031558	0.023984	11 way, home, point, view, daughter, time, thats, turn, sorry, right

In the Marvel's top topics also military themes can be seen, but also words like power, beast, spark and weapon, that make it seem more super hero.

studio	Lucasfilm Ltd.	Marvel Studios	Pixar	Walt Disney Animation Studios	label
topic_id					
8	0.020669	0.036182	0.017306	0.018231	8 moment, hmm, thor, kind, target, dont, hes, way, eyes, boss
19	0.020650	0.035546	0.021615	0.018796	19 yeah, hell, dude, time, man, years, hours, sorry, yeah yeah, way
14	0.037424	0.035538	0.042236	0.021845	14 sir, yes, yes sir, body, nemo, im, trade, youre, men, gonna
25	0.019884	0.034772	0.020092	0.017835	25 gonna, kids, talk, powers, mission, fear, time, fight, area, people
21	0.021122	0.034035	0.048336	0.035318	21 okay, hey, hey hey, yeah, thing, spark, train, okay okay, mask, left
1	0.024417	0.032765	0.022527	0.028476	1 sorry, course, time, day, shh, stark, weapon, drop, captain, yeah
18	0.028018	0.031012	0.022548	0.043879	18 girl, time, father, head, plan, guys, sorry, youre, beast, village

Disney Animated's top topics have a lot more familiar word, like father, sister, love but also adventure words.

studio	Lucasfilm Ltd.	Marvel Studios	Pixar	Walt Disney Animation Studios	label
topic_id					
18	0.028018	0.031012	0.022548	0.043879	18 girl, time, father, head, plan, guys, sorry, youre, beast, village
13	0.020844	0.019057	0.016701	0.043597	13 hook, way, bit, maui, love, ooh, time, thing, deal, beast
15	0.023415	0.025509	0.027675	0.036900	15 yeah, time, dream, school, today, gonna, day, life, attack, hands
2	0.019009	0.020580	0.015458	0.036001	2 heart, gonna, future, people, congratulations, right, island, honor, way, time
21	0.021122	0.034035	0.048336	0.035318	21 okay, hey, hey hey, yeah, thing, spark, train, okay okay, mask, left
23	0.022185	0.017990	0.016742	0.031624	23 elsa, sister, love, way, uh, doesnt, years, kingdom, world, books
28	0.024787	0.019856	0.021256	0.030322	28 ah, son, blood, man, right, bye, hold, drive, death, day
24	0.026194	0.023150	0.022713	0.029493	24 right, looks, strength, hand, thing, cold, birthday, head, storm, way
22	0.034989	0.025152	0.030525	0.029017	22 youre, minute, time, morning, droids, fine, oh, way, right, position

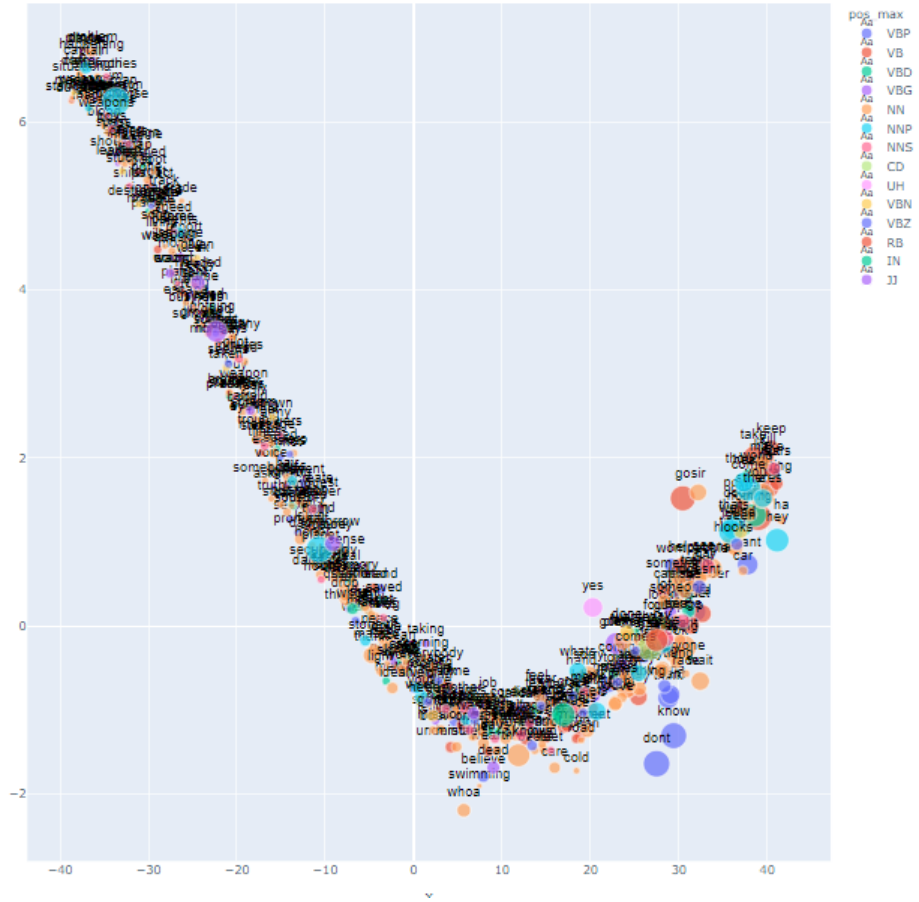
Pixar's top topics seem less unique, but you can see Monsters Inc in topic 33.

studio	Lucasfilm Ltd.	Marvel Studios	Pixar	Walt Disney Animation Studios	label
topic_id					
21	0.021122	0.034035	0.048336	0.035318	21 okay, hey, hey hey, yeah, thing, spark, train, okay okay, mask, left
14	0.037424	0.035538	0.042236	0.021845	14 sir, yes, yes sir, body, nemo, im, trade, youre, men, gonna
11	0.032451	0.028648	0.031558	0.023984	11 way, home, point, view, daughter, time, thats, turn, sorry, right
22	0.034989	0.025152	0.030525	0.029017	22 youre, minute, time, morning, droids, fine, oh, way, right, position
33	0.020748	0.023117	0.030481	0.020026	33 door, door door, youre, monsters, hey, chance, guys, tower, ships, field
7	0.021557	0.022586	0.030313	0.022933	7 whoa, energy, whoa whoa, luck, wow, life, attention, rule, place, people

Word Embedding:

Word embedding tries to generate word vectors and learn the meaning of words by their context. There are different approaches to embedding words, some use PCA like above, but this analysis used word2vec which is a simple two layer neural network.

This corpus produced interesting, and I would consider atypical word embedding results. As seen in the tSNE graph below.

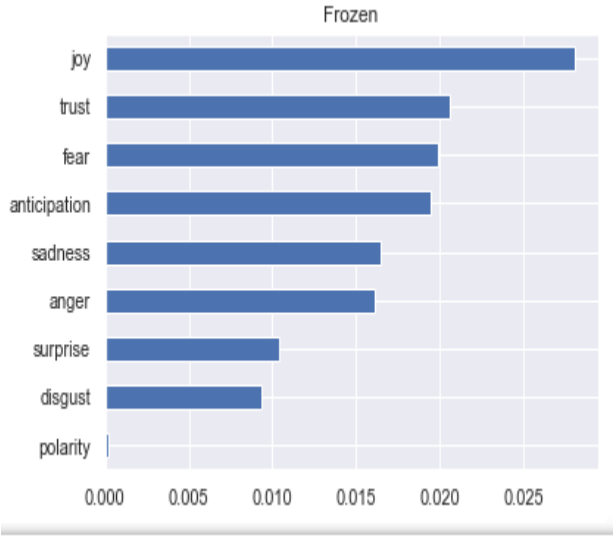


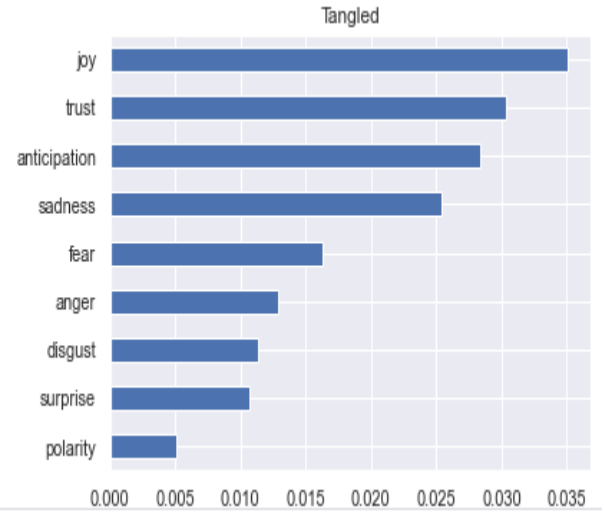
I found a similar linear structure when I grouped by live action and animated as well. Some theories I have that might contribute to this are the fragmented nature of using only speech in a word embedding. Having a larger corpus might also help expand word2vec's ability to find meaning.

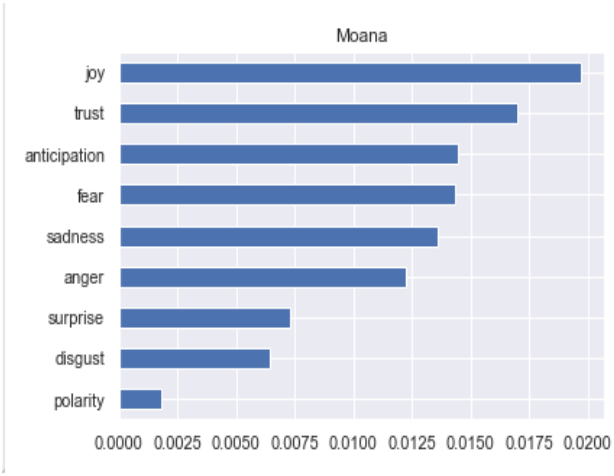
Sentiment Analysis:

Sentiment Analysis is an attempt to describe the emotional valence of a text. Sometimes texts can be binary classified as positive or negative, or it can be expanded to categories of emotion typically using eight basic emotions: joy, sadness, anger, fear, trust, disgust, surprise, anticipation. There are several different strategies for sentiment analysis all of which have failings. This analysis will be using the NRC Emotional Lexicon.

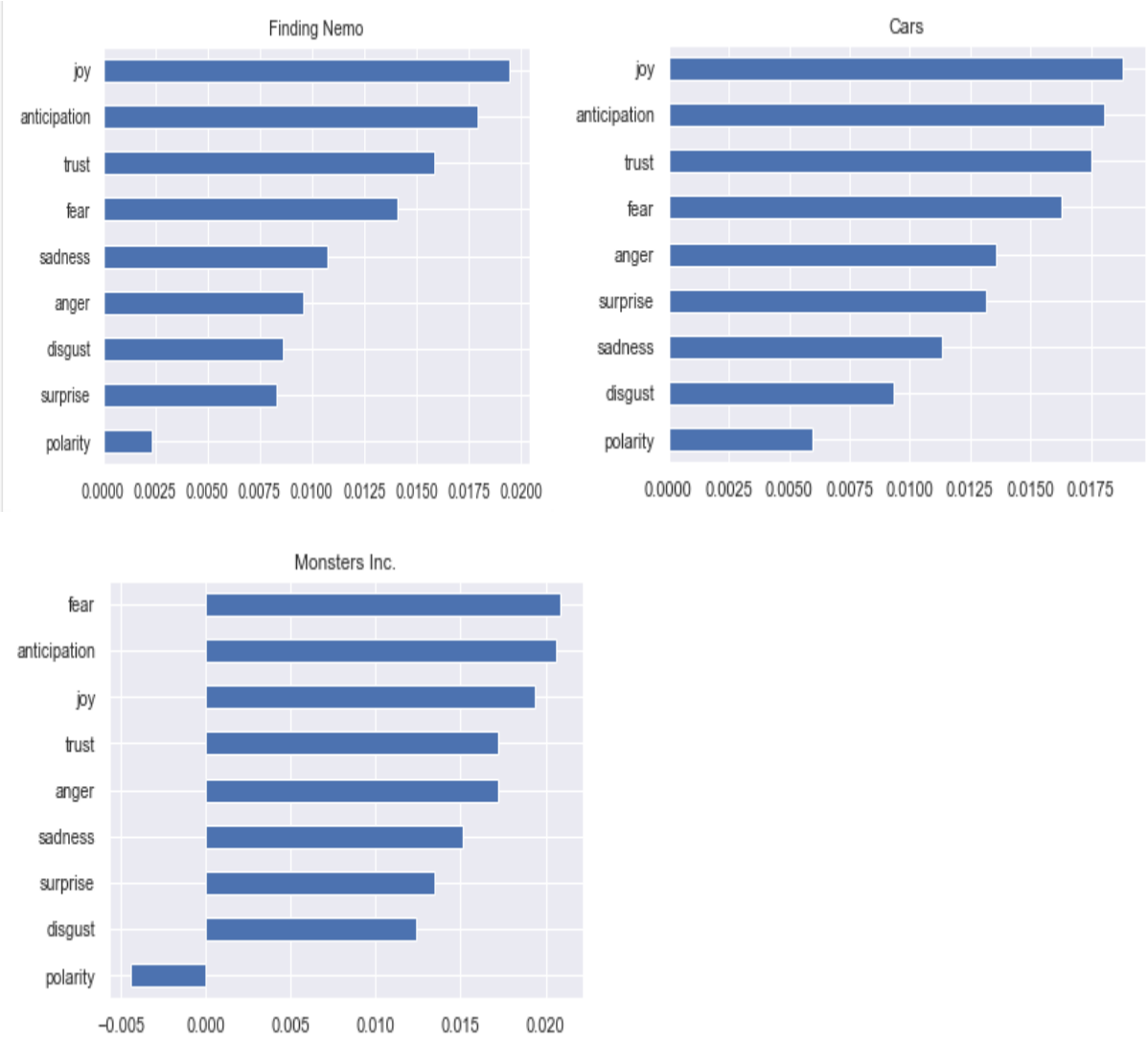
Some Disney princess, we see joy and trust are the top emotions, but fear plays a much bigger role in Frozen, of both Elsa being afraid of her powers, and the town being afraid of her.



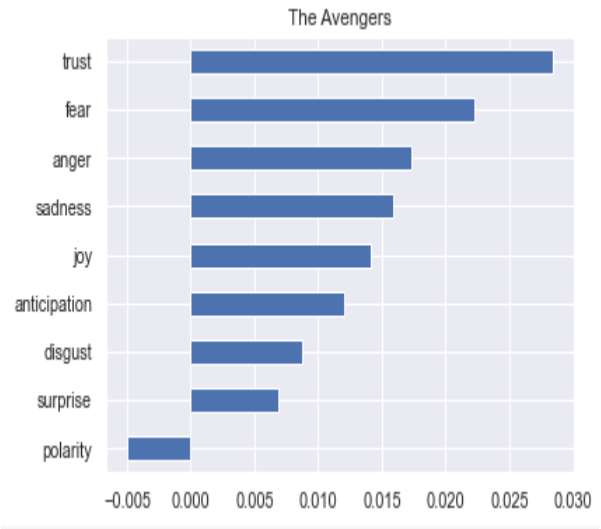


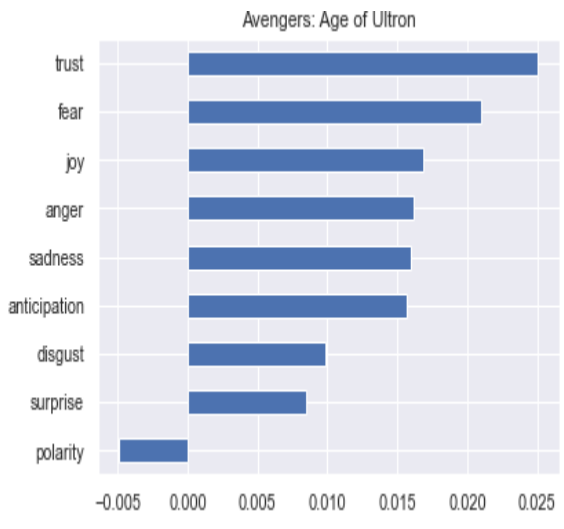


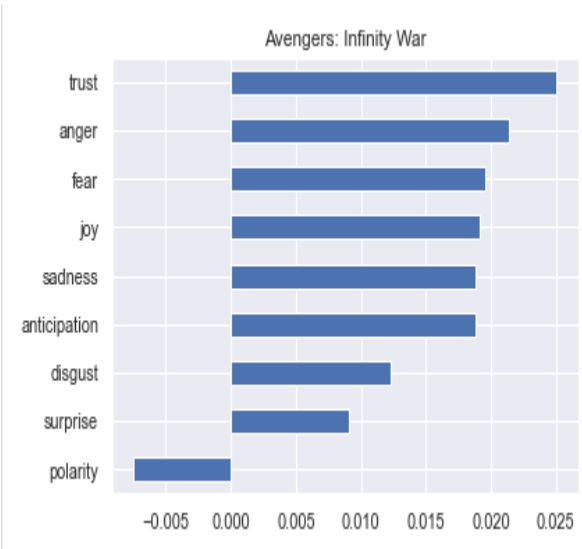
In Pixar, Monsters Inc. has an extremely high fear considering it's an animated movie for kids. This makes more sense when understanding the premise of the movie is they scare kids to produce energy, and evolving to realize there are better ways.

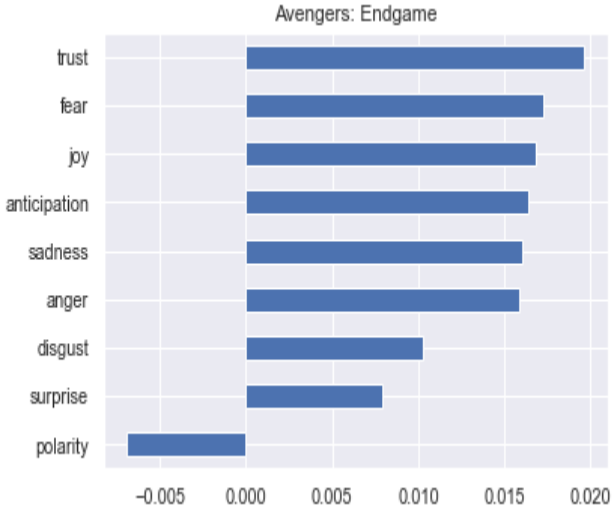


In contrast to the animated movies that had joy at the top for all the films, in the Avenger movies joy is between third and fifth, where trust and fear are always at the top. For a team based action movie fighting the world's extension this also makes sense.









Conclusion:

Overall applying these exploratory text analysis techniques to this selection of movie scripts provided some interesting ideas and helped guide where further exploration is needed. It is clear that movie scripts are a good area to apply text analysis, I believe it would be beneficial to have scene level data, as well as a larger and possibly wider corpus to analyze. And additional investigation is needed to provide useful word embedding.

Corpus compiling and some analysis done in collaboration with Siddharth Surapaneni

Appendix Library Table

movie_id	name	studio	director	genre	year
m1	Rogue One: A Star Wars Story	Lucasfilm Ltd.	Gareth Edwards	Action Adventure Science Fiction	2016
m2	Star Wars: Episode III - Revenge of the Sith	Lucasfilm Ltd.	George Lucas	Action Science Fiction Adventure	2005
m3	Star Wars: The Rise of Skywalker	Lucasfilm Ltd.	J.J. Abrams	Science Fiction Adventure Action	2019
m4	Captain America: The First Avenger	Marvel Studios	Joe Johnston	Science Fiction Action Adventure	2011
m5	Captain America: The Winter Soldier	Marvel Studios	Anthony Russo Joe Russo	Action Adventure Science Fiction	2014
m6	Captain America: Civil War	Marvel Studios	Anthony Russo Joe Russo	Action Science Fiction Adventure	2016
m7	The Avengers	Marvel Studios	Joss Whedon	Science Fiction Adventure Action	2012
m8	Avengers: Age of Ultron	Marvel Studios	Joss Whedon	Science Fiction Action Adventure	2015
m9	Avengers: Infinity War	Marvel Studios	Anthony Russo Joe Russo	Adventure Action Science Fiction	2018
m10	Avengers: Endgame	Marvel Studios	Anthony Russo Joe Russo	Adventure Science Fiction Action	2019
m11	Captain Marvel	Marvel Studios	Ryan Fleck Anna Boden	Science Fiction Action Adventure	2019
m12	Beauty and the Beast	Walt Disney Animation Studios	Gary Trousdale Kirk Wise	Animation Family Fantasy Romance	1991
m13	Inside Out	Pixar	Pete Docter	Adventure Comedy Family Drama Animation	2015
m14	Monsters University	Pixar	Dan Scanlon	Animation Family	2013
m15	Monsters Inc.	Pixar	Pete Docter	Family Comedy Animation	2001
m16	Moana	Walt Disney Animation Studios	Ron Clements John Musker	Comedy Adventure Animation Family	2016
m17	Zootopia	Walt Disney Animation Studios	Byron Howard Rich Moore	Comedy Family Adventure Animation	2016
m18	Mulan	Walt Disney Animation Studios	Tony Bancroft Barry Cook	Adventure Family Animation	1998
m19	Soul	Pixar	Pete Docter	Family Drama Music Comedy Animation Fantasy	2020
m20	Frozen	Walt Disney Animation Studios	Chris Buck Jennifer Lee	Family Adventure Animation	2013
m21	Frozen II	Walt Disney Animation Studios	Chris Buck Jennifer Lee	Fantasy Music Family Adventure Comedy Animation	2019
m22	Tangled	Walt Disney Animation Studios	Byron Howard Nathan Greno	Family Animation	2010
m23	Cars	Pixar	John Lasseter	Animation Adventure Family Comedy	2006
m24	Finding Nemo	Pixar	Andrew Stanton	Family Animation	2003