CAPP 30255 Final Project

Gender in Movie Scripts

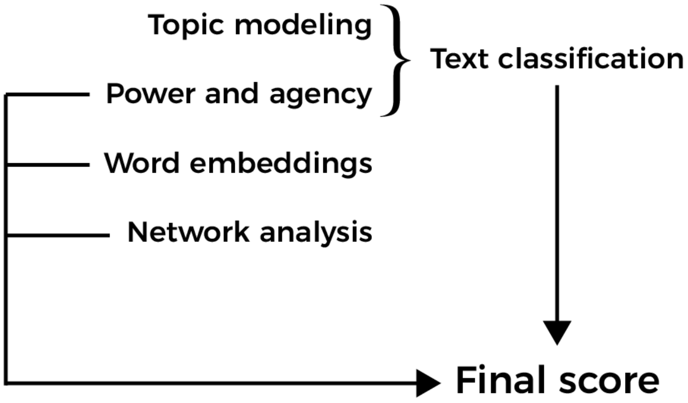
Emma Peterson | *Jasmin Dial* | *Joan Wang* | *Regina Widjaya*

Introduction

The following paper is structured as follows:

Project overview

Our project utilizes a combination of machine learning, natural language processing, and network analysis tools to better understand the level of gender imbalance in movie scripts. We developed several methods of measuring gender imbalance, each of which feeds into a final score for each movie. The structure of our process is outlined below, and each piece is discussed in detail in the following sections.



Background

Some previous work has addressed gender differences in movies. We aim to build on this prior work by combining multiple measures of gender imbalance into a single score.

Key Female Characters in Film Have More to Talk About Besides Men: Automating the Bechdel Test (2015) attempted to automate the process of determining whether or not a movie passes the Bechdel Test. Using linguistic features and network analysis, they found that about 40% of the movies tested failed to include at least two women speaking to each other about something other than a man. They also found that failing movies tended to portray women as less important or peripheral characters.

Connotation Frames of Power and Agency in Modern Films (2017) used crowdsourcing to obtain scores for a variety of verbs based on power and agency. Power was defined as a measure of a character's authority over another, while agency was defined as a measure a character's ability to push their own storyline forward. The authors found that men are portrayed to have more authority and a higher level of agency. The verbs and their scores are available for download, and we plan to incorporate these into our analysis.

Finally, *A Quantitative Analysis of Gender Differences in Movies Using Psycholinguistic Normatives* (2015) examined "gender ladenness", or the degree of perceived femininity/masculinity on a numerical scale ranging from very masculine (-4) to very feminine (+4). The authors started with 925 words and their scores from a previous paper and utilized word similarity techniques to estimate ratings for other words. Their findings were somewhat predictable, including that romantic movies tend to include more feminine language, and action movies tend to include more masculine language.

Dataset

For this project, we used the Cornell Movie Dialog Corpus dataset created by Cristian Danescu-Niculescu-Mizil and Lillian Lee from Cornell University. The dataset is a large metadata-rich collection of fictional conversations extracted from raw movie scripts. It contains 220,579 conversational exchanges between 10,292 pairs of movie characters. The data records a total of 304,713 lines spoken by 9,035 characters from 617 movies released between the years 1927 and 2010. The metadata available included movie genres, IMDB rating, number of IMDB votes, as well as gender information and position on movie credit for 3,744 and 3,321 characters respectively.

The dataset included the following tables and information:

- Movie Characters : character id, names, gender, movie id, movie title, credit position
- Movie Lines : character id, names, movie id, line/text id, lines/text
- Movie Title : movie id, movie title, genre, imdb rating, imdb votes, year
- Movie Conversation : speaker character id, receiver character id, movie id, turns id

Pre-processing

The original movie dataset was relatively clean and well-structured; most of the data required only minor adjustments, and the majority of pre-processing work was performed on the conversation text data.

Adjustments were made to the characters dataset to standardize and populate gender classifications. We infer missing gender classifications using the NLTK names corpus and successfully classify up to 2,229 additional characters, which adds 25,256 more lines of text to our analysis. Characters with unisex names are classified as male characters. The final un-gendered characters comprised only 1% of the dialogue lines in the conversation text data, which we then excluded from our analysis.

Minor adjustments were made to the title and and conversation dataset in order to separate multiple movie genres and speaking turns into individual data entries.

The text (conversation/dialog lines) dataset went through multiple stages of pre-processing: tokenization, stopword removal, and lemmatization. We tokenize dialog corpus using the NLTK word-punctuation tokenizer to separate lines of dialog into series of words separated by space and punctuations. We chose this method of tokenization to ensure better lemmatization on abbreviated words. Stopwords are removed from the corpus using the NLTK stopword database. Pronouns are kept in the corpus for further exploration in topic modeling and power and agency sub-analysis. In the lemmatization process, we tagged tokenized words using the nltk part-of-speech (POS) package and infer the base form of tokenized words given it's positional (grammatical) context and reduce informal word variations that might occur in the conversational nature of movie dialogs. We did not use stemming in pre-processing our data as it might be too reductive for some of our analysis that require context and variations in vocabulary (i.e., topic modeling, power and agency, and word embeddings).

Sub-Projects

Proportion of Lines by Gender

We calculated the distribution of lines between genders in a given movie. This metric will be used as one of the weight factors in the final movie scoring.

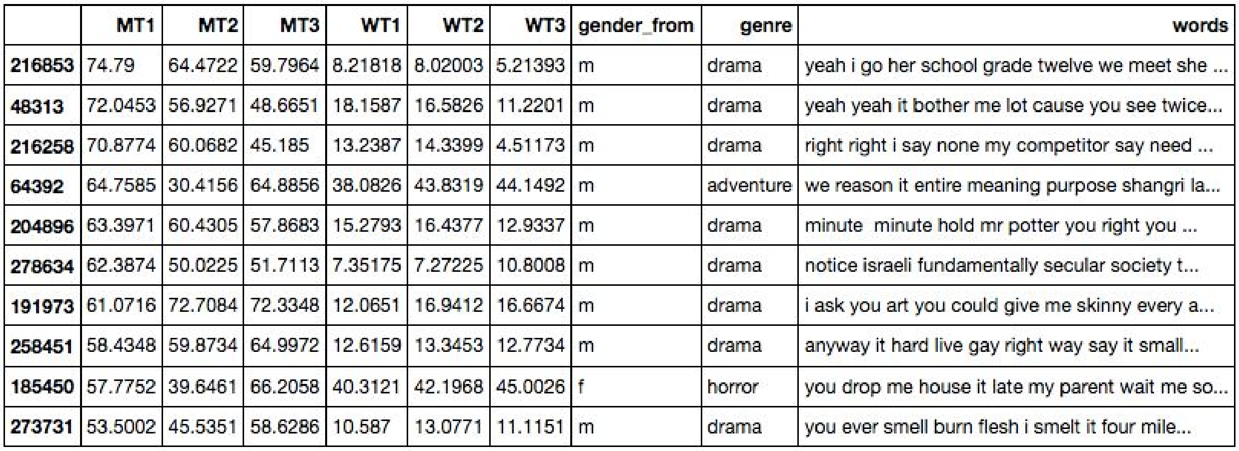
Topic Modeling

To get a sense of the difference in lines spoken by men and women, we calculated the most frequent n-grams for each group, after removing stop words. These frequent terms did not vary much by gender, as can be expected since movie lines largely represent actual speech. We then turned to topic modeling, the process of extracting important topics from a large document, to detect differences in how language is written for different speakers. For this process, we used the Latent Dirichlet Allocation (LDA) calculation from the Python module Gensim. We started by using the LDA implementation in sklearn, but documentation suggested that Gensim may perform better on our relatively large data set. The model is trained on a set of words, and then determines topics matching the number of topics specified, in addition to a specified number of words associated with each topic. The words associated with each topic also include a weight for their contribution to a topic.

For our purposes, we separated the training data into lines spoken by men and lines spoken by women. Models with multiple numbers of topics were fit, and the optimal model was chosen by the coherence score, which is calculated through the Gensim LDA module. To additionally detect differences in gendered language that varies across genres, we fit a model for each of the most common movie genres (action, comedy, drama, and crime) by gender. In total, we stored 10 sets of topic models - one per gender and one per genre gender per genre.

Before the models were fit, we removed part-of-speech tags ‘IN’, ‘CD’, and ‘MD’, which act only as supporting words, and would not contribute to detecting topics. We created bigrams and trigrams from the individual words of the text, which could then be detected as an element of a topic in addition to individual words. Lastly, we removed words in our data set that were in the bottom 10% of overall frequency, hoping that this would avoid using terms that are very specific to one movie but do not appear overall. Because the various models varied in coherence - the action, female model had coherence of 0.52 while the crime, female model had coherence of only 0.31, we might benefit from additional pre-processing steps, which might include detecting and removing names.

Finally, we used the detected topic models to create features for both the training and test sets to use for text classification portion of the project. The authors of *Topic Modeling Based Classification of Clinical Reports (2013)* report that topic models have been used to select keywords as classification, as features to compare to bag of words, or as features in addition to bag of words. We used features in addition to BOW, and a representation of movie lines with the topic model features MT1 through WT3 is shown below. On top of outputting topics for a text, LDA can determine the probability of observing these topic in a given document, or in our case, movie line. For each line, we output a score for this probability, which is given by the Gensim LDA module, depending on the genre of the line. For lines from the most common genres, MT1-MT3 and WT1-WT3 represent the probabilities of observing the top three topics determined by the male lines of that genre and the female line of that genre. For movies from uncommon genres, which represent about ⅙ of our data, the features represent the probabilities for the top three overall topics for men and women. In theory, we would expect lines spoken by men to have higher values for MT1-MT3 than WT1-WT3. Though this is not always the case, especially among lines spoken by women, including these features does slightly improve text classification performance, and we might consider adding features for additional topics.



	MT1	MT2	MT3	WT1	WT2	WT3	gender_from	genre	words
216853	74.79	64.4722	59.7964	8.21818	8.02003	5.21393	m	drama	yeah i go her school grade twelve we meet she ...
48313	72.0453	56.9271	48.6651	18.1587	16.5826	11.2201	m	drama	yeah yeah it bother me lot cause you see twice...
216258	70.8774	60.0682	45.185	13.2387	14.3399	4.51173	m	drama	right right i say none my competitor say need ...
64392	64.7585	30.4156	64.8856	38.0826	43.8319	44.1492	m	adventure	we reason it entire meaning purpose shangri la...
204896	63.3971	60.4305	57.8683	15.2793	16.4377	12.9337	m	drama	minute minute hold mr potter you right you ...
278634	62.3874	50.0225	51.7113	7.35175	7.27225	10.8008	m	drama	notice israeli fundamentally secular society t...
191973	61.0716	72.7084	72.3348	12.0651	16.9412	16.6674	m	drama	i ask you art you could give me skinny every a...
258451	58.4348	59.8734	64.9972	12.6159	13.3453	12.7734	m	drama	anyway it hard live gay right way say it small...
185450	57.7752	39.6461	66.2058	40.3121	42.1968	45.0026	f	horror	you drop me house it late my parent wait me so...
273731	53.5002	45.5351	58.6286	10.587	13.0771	11.1151	m	drama	you ever smell burn flesh i smelt it four mile...

Power and Agency

The use of language conveying power and agency was analyzed with data from *Connotation Frames of Power and Agency in Modern Films* (2017). This paper used crowdsourcing techniques to obtain power and agency scores for a variety of verbs. Power was defined as a measure of a character's authority over another, while agency was defined as a measure a character's ability to push their own storyline forward. We were able to download these verbs, each of which is associated with a positive, equal, or negative score for both power and agency. For example, "abolishes" is rated as both positive agency and positive power, while "awaits" is rated as both negative agency and negative power.

We first pre-process the verbs using the same methods that were used on the movie lines themselves. Next, we obtain a count of the total verbs in each line of dialogue, as well as a count of the number of verbs falling into a positive or negative category for either power or agency. Finally, the raw counts are converted into proportions representing the fraction of verbs in a given line that falls into one of the following categories: negative agency, positive agency, negative power, or positive power.

Averaging these proportions by gender, we found that male characters used slightly higher proportions of positive power and agency verbs and slightly lower proportions of negative power and agency verbs. We also found that when analyzing the proportions by speaker pairs, female to female interactions included the lowest proportion of positive power verbs, while male to male interactions included the highest proportion of positive power verbs. All of these findings confirmed our initial expectations, though the differences were quite small (see below).

	+ agency	**− agency**	**+ power**	**− power**
female	0.374	0.164	0.251	0.078
male	0.380	0.154	0.272	0.067

	+ power
f → f	0.239
f → m	0.253
m → f	0.258
m → m	0.280

Before adding the proportions to the text classification model, we wanted to ensure that there was a sufficient amount of variation across movies. After an assessment of the distribution of the proportions, we determined that it appeared possible that they encoded additional information about the text, which the classifier might find useful.

Text Classification

The purpose of this project component is to produce a machine learning model that accurately distinguishes between dialogue that has a male or female leaning, at least by Hollywood script-writing standards. We built a Multinomial Naive Bayes (MNB) text classification model to classify the speaker of movie lines as either male (outcome = 0) or female (outcome = 1).

To build this model, we followed the following steps:

- Inputted “lines” of preprocessed dialogue (one speaker’s turn in a script) as input observations
- Turned them into frequency counts of a bag of words
- Transformed the counts using TFIDF (Term Frequency times Inverse Document Frequency)
- Combined the TFIDF matrix with two other groups of non-text features calculated in previous sections: probabilities of top three topics by gender and percentage of verbs in positive/negative power and agency lists
- Trained a MNB classifier on the resulting features matrix

When this model is applied to the test set, it achieves the following performance:

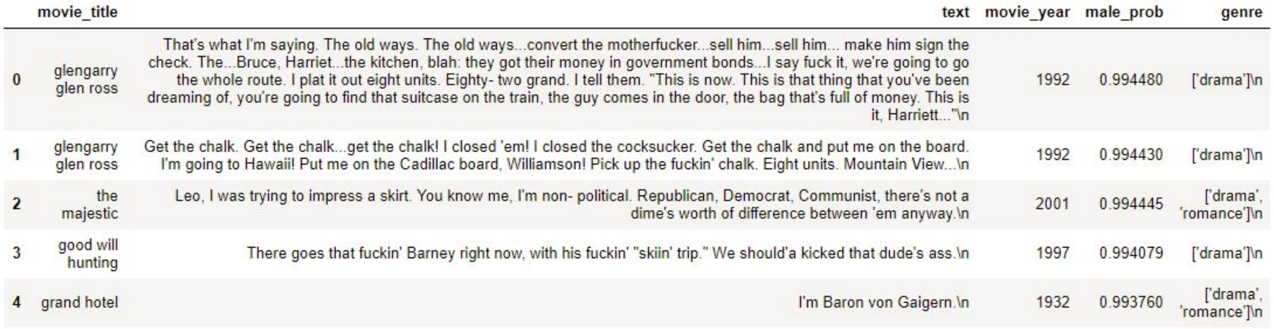
- Accuracy: 67%
- Precision: 42%
- Recall: 25%
- Confusion matrix:

30,530	5,529
11,861	4,021

Prior to arriving at this model, we iterated through a number of different classification specifications and model types. We first performed a grid search of a number of different classifiers and hyperparameter combinations. Classifiers we tried were Multinomial Naive Bayes, Logistic Regression, Random Forest, Decision Tree, and Gradient Boost. Logistic regression models were a close runner up to MNB, but we chose MNB because of their popularity in Natural Language Processing and their interpretation as a generative model.

Once we arrived at using MNB, we built separate classifiers by movie genre to see if models trained on movies belonging to a specific genre performed better than the overall model. We also tried predicting the speaker pair for each movie line, i.e. M to M, M to F, F to M, or F to F. None of these models performed much better than the overall model. We also tried processing the text as bigrams and trigrams instead of bag of words, which produced much worse results.

We used this model to quantify the degree to which an unseen movie line fits the Hollywood mold of male and female. Highly "masculine” lines are expected to receive high probability scores for the male class, and we see evidence of this in movie lines from the holdout set. The following lines received the highest probability score for the male class:



	movie_title	text	movie_year	male_prob	genre
0	glengarry glen ross	That's what I'm saying. The old ways. The old ways...convert the motherfucker...sell him...sell him... make him sign the check. The...Bruce, Harriet...the kitchen, blah: they got their money in government bonds...I say fuck it, we're going to go the whole route. I plat it out eight units. Eighty- two grand. I tell them. "This is now. This is that thing that you've been dreaming of, you're going to find that suitcase on the train, the guy comes in the door, the bag that's full of money. This is it, Harriett..."\n	1992	0.994480	['drama']\n
1	glengarry glen ross	Get the chalk. Get the chalk...get the chalk! I closed 'em! I closed the cocksucker. Get the chalk and put me on the board. I'm going to Hawaii! Put me on the Cadillac board, Williamson! Pick up the fuckin' chalk. Eight units. Mountain View...\n	1992	0.994430	['drama']\n
2	the majestic	Leo, I was trying to impress a skirt. You know me, I'm non- political. Republican, Democrat, Communist, there's not a dime's worth of difference between 'em anyway.\n	2001	0.994445	['drama', 'romance']\n
3	good will hunting	There goes that fuckin' Barney right now, with his fuckin' "skiin' trip." We should'a kicked that dude's ass.\n	1997	0.994079	['drama']\n
4	grand hotel	I'm Baron von Gaigern.\n	1932	0.993760	['drama', 'romance']\n

Network Analysis

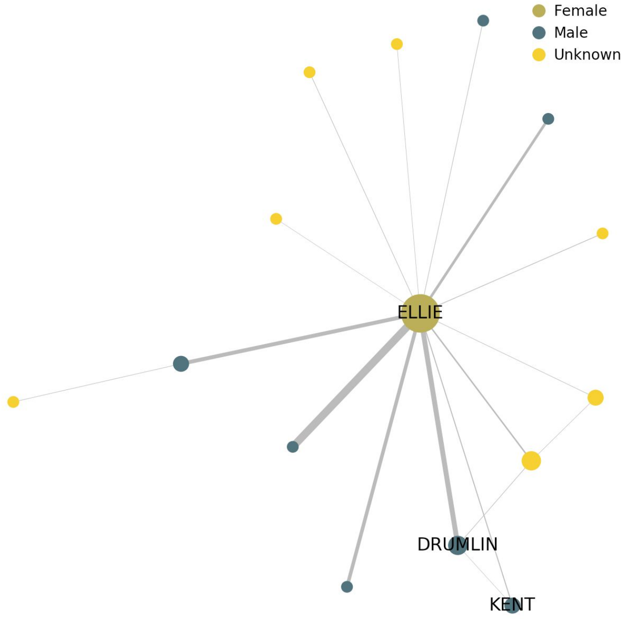
We used the networkx package to perform network analyses of each movie in order to assess the centrality of characters by gender. For each movie, we built an undirected network in which each character is a node and each edge connects two characters who interacted with one another. Among the various definitions of node centrality, we selected degree centrality and betweenness centrality as two measures of a character's importance.

For a given character, degree centrality represents the fraction of other characters with which they interacted. Betweenness centrality is a measure of how often the character is located on the shortest path between two other characters. We selected these two measures of centrality because we could imagine a scenario in which a character had very few direct connections (i.e. low degree centrality), but was nonetheless important in connecting other characters to each other (i.e. high betweenness centrality) – or vice versa.

Averaging these centrality measures by gender across all movies, we found that male characters had a higher level of both degree and betweenness centrality:

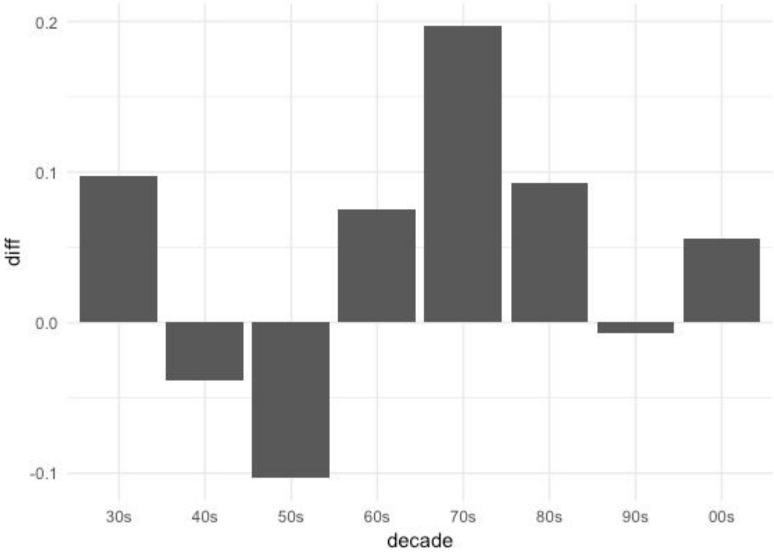
	degree	**betweenness**
female	0.240	0.105
male	0.289	0.161

In assessing these measures by genre, we found that sci-fi, fantasy, and horror were the only three in which female characters are portrayed with greater centrality than male characters on average. We created graph images in order to visually inspect some movies to ensure that the centrality measures made intuitive sense based on the structure of their networks. For example, Contact (1997) was the movie in which female centrality most strongly outweighed male centrality. Here, the nodes are sized by each character's number of lines, and the edges are weighted by the number of interactions between two characters:



We also looked at the centrality measures in comparison to IMDB ratings. There is a 0.65 gap in average IMDB ratings between the top 20 movies for female centrality and the top 20 movies for male centrality (with the latter group receiving higher ratings on average). We did not assess whether this difference is significant, and it represents an area of possible further research. If the difference is significant, it would be interesting to determine the cause. Perhaps people rate female-centric movies poorly because they are female-centric, or perhaps female-centric movies are generally of lower quality because there is a smaller market for such movies.

The trend in centrality scores over time does not immediately suggest any particular pattern. The graph below shows the average difference between male and female centrality by decade, with negative values suggesting higher female centrality.


0.2
0.1
0.0
-0.1
diff
30s
40s
50s
60s
70s
80s
90s
00s
decade

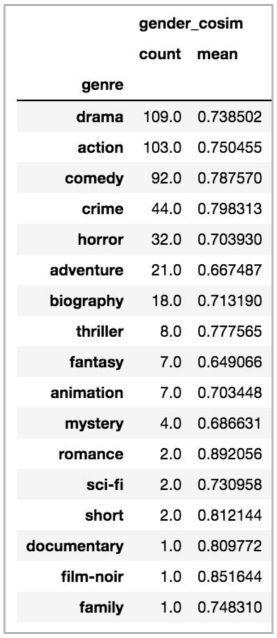
Word Embeddings

The goal of this sub-project is straightforward: to quantify and calculate how similar male and female dialogs are in a given movie.

First we embed scoring into tokenized words in a document using the TF-IDF (Term Frequency-Inverse Document Frequency) measures and vectorize them. TF calculation measures how many times a term occured in a document, while IDF calculation measures how important a word is in the context of a document. We then take the vectorized document and calculate the 'physical distance' of words between two compared documents in a vector space using the cosine distance measurement. The cosine distance 'score' measure the degree of similarity between two compared documents.

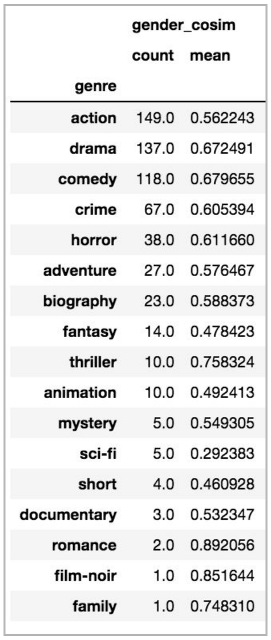
Similarity scoring between two documents ranged from 0 being not similar at all to 1 being almost identical. The table below shows cosine similarity scoring on the movie genre level. The table on the left shows similarity scoring between male and female when both genders have dialogs in the movies (i.e., movies with dialogs from one gender only, which have an automatic similarity score of zero, are excluded) the sample similarity score ranges between .64 (fantasy) to .89 (romance), with movie genres with most samples in dataset such as drama, action, comedy, and crime all have similarity scores at around .70.

Dialog Similarity Score in Movie Genres
(movies with both gender)



	gender_cosim	
	count	mean
genre		
drama	109.0	0.738502
action	103.0	0.750455
comedy	92.0	0.787570
crime	44.0	0.798313
horror	32.0	0.703930
adventure	21.0	0.667487
biography	18.0	0.713190
thriller	8.0	0.777565
fantasy	7.0	0.649066
animation	7.0	0.703448
mystery	4.0	0.686631
romance	2.0	0.892056
sci-fi	2.0	0.730958
short	2.0	0.812144
documentary	1.0	0.809772
film-noir	1.0	0.851644
family	1.0	0.748310

Dialog Similarity Score In Movie Genres
(all movies)

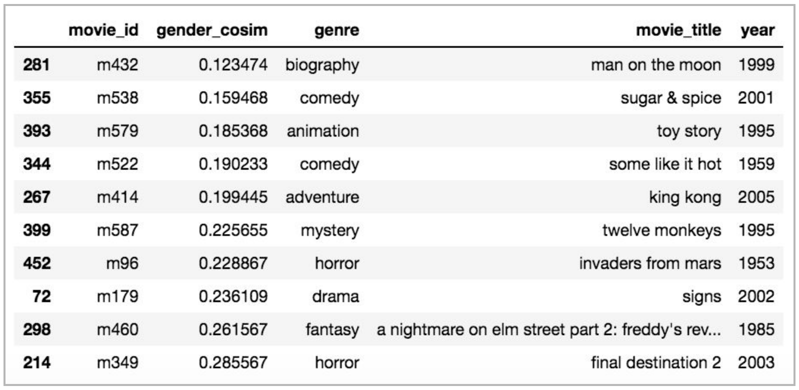


	gender_cosim	
	count	mean
genre		
action	149.0	0.562243
drama	137.0	0.672491
comedy	118.0	0.679655
crime	67.0	0.605394
horror	38.0	0.611660
adventure	27.0	0.576467
biography	23.0	0.588373
fantasy	14.0	0.478423
thriller	10.0	0.758324
animation	10.0	0.492413
mystery	5.0	0.549305
sci-fi	5.0	0.292383
short	4.0	0.460928
documentary	3.0	0.532347
romance	2.0	0.892056
film-noir	1.0	0.851644
family	1.0	0.748310

When movies with only one gender dialogs are introduced to the analysis we started to see larger variations in the scores, as we can see from the scores on the right table above. From all movie samples in the dataset with only one gender dialogs, around 92% (105 out of 113) is missing female dialogs/characters. This proportion then acted as a weighting factor. Looking back to the four most populated movie genres in the sample, we can start to see that in genres that are traditionally male oriented such as action lost almost a whole .20 similarity score to .56.

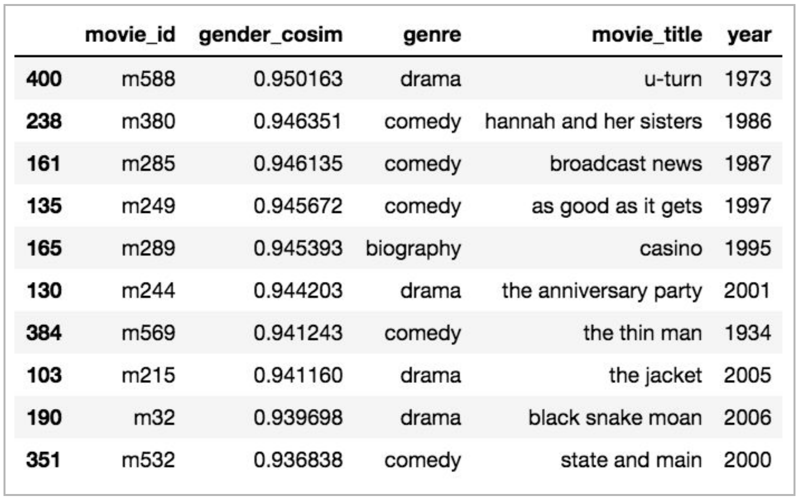
On the movie level, movies with highest similarity scores came mostly from the drama and comedy genres, while those with the lowest scores are more varied.

10 Movies With Least Similar Dialogs



	movie_id	gender_cosim	genre	movie_title	year
281	m432	0.123474	biography	man on the moon	1999
355	m538	0.159468	comedy	sugar & spice	2001
393	m579	0.185368	animation	toy story	1995
344	m522	0.190233	comedy	some like it hot	1959
267	m414	0.199445	adventure	king kong	2005
399	m587	0.225655	mystery	twelve monkeys	1995
452	m96	0.228867	horror	invaders from mars	1953
72	m179	0.236109	drama	signs	2002
298	m460	0.261567	fantasy	a nightmare on elm street part 2: freddy's rev...	1985
214	m349	0.285567	horror	final destination 2	2003

10 Movies With Most Similar Dialogs



	movie_id	gender_cosim	genre	movie_title	year
400	m588	0.950163	drama	u-turn	1973
238	m380	0.946351	comedy	hannah and her sisters	1986
161	m285	0.946135	comedy	broadcast news	1987
135	m249	0.945672	comedy	as good as it gets	1997
165	m289	0.945393	biography	casino	1995
130	m244	0.944203	drama	the anniversary party	2001
384	m569	0.941243	comedy	the thin man	1934
103	m215	0.941160	drama	the jacket	2005
190	m32	0.939698	drama	black snake moan	2006
351	m532	0.936838	comedy	state and main	2000

Movie scoring

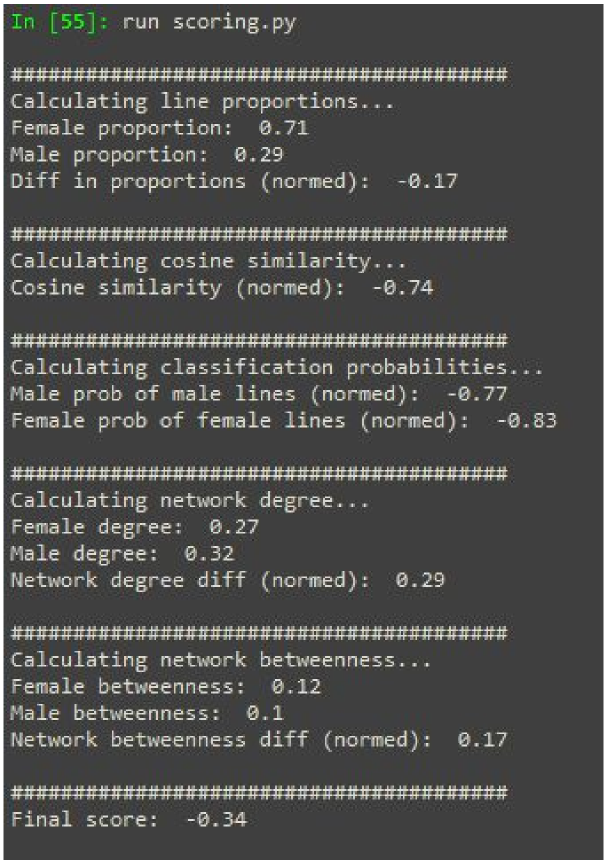
Process

With the exception of topic modeling, each of the above sub-projects feeds into an overall score for each movie. This final score is an approximate measure of the degree of “genderedness” in a movie, where high scores represent greater gender disparity and low scores represent lower gender disparity.

The score is a combination of:

1. The proportion of male versus female dialogue (represented as an absolute difference between the two proportions)
2. Reversed cosine similarity between male and female dialogue
3. Average male class probability among male lines
4. Average female class probability among female lines
5. Absolute difference between male and female degree centrality
6. Absolute difference between male and female betweenness centrality

Each of these individual components is normalized by the corresponding training mean and standard deviation for the sake of scale similarity. The final score is an equally-weighted combination of the six components. Below is an example of the output from scoring a single movie.

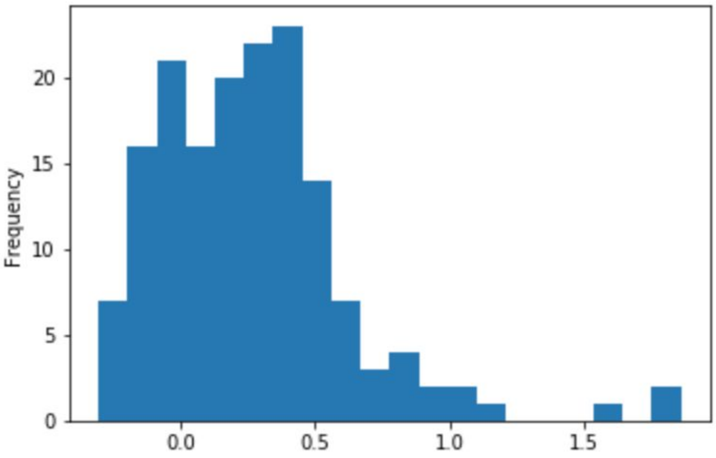


```
In [55]: run scoring.py

########################################
Calculating line proportions...
Female proportion:  0.71
Male proportion:  0.29
Diff in proportions (normed):  -0.17

########################################
Calculating cosine similarity...
Cosine similarity (normed):  -0.74

########################################
Calculating classification probabilities...
Male prob of male lines (normed):  -0.77
Female prob of female lines (normed):  -0.83

########################################
Calculating network degree...
Female degree:  0.27
Male degree:  0.32
Network degree diff (normed):  0.29

########################################
Calculating network betweenness...
Female betweenness:  0.12
Male betweenness:  0.1
Network betweenness diff (normed):  0.17

########################################
Final score:  -0.34
```

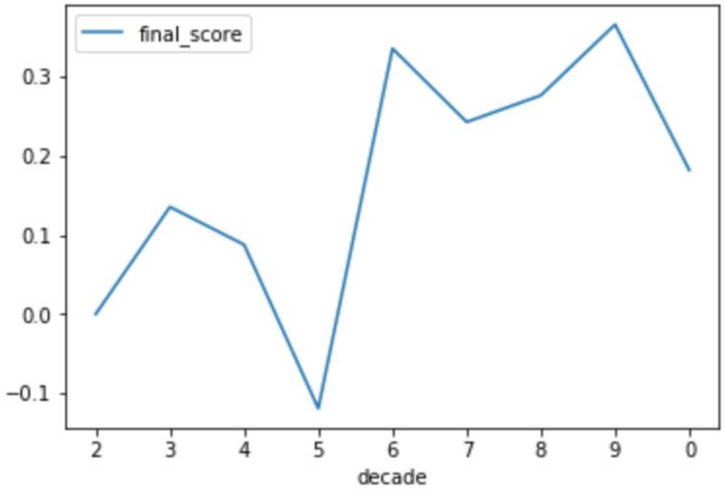
These scores do not suggest any particular direction of genderdness – for example, a high score for a given movie may suggest that the genders in that movie are very different in ways that favor women, or in ways that favor men. In addition, the differencing and normalization of the values that comprise the scores means that the scores in and of themselves are not highly meaningful, and are instead useful as relative rankings. As such, exploration of the metrics that feed into each score is necessary in order to better understand why particular movies receive particular scores.

Exploratory Analysis

We first looked at the overall distribution of the final scores. The plot below demonstrates that the scores are fairly right skewed, with a majority of movies receiving low scores (representing low levels of gender imbalance).

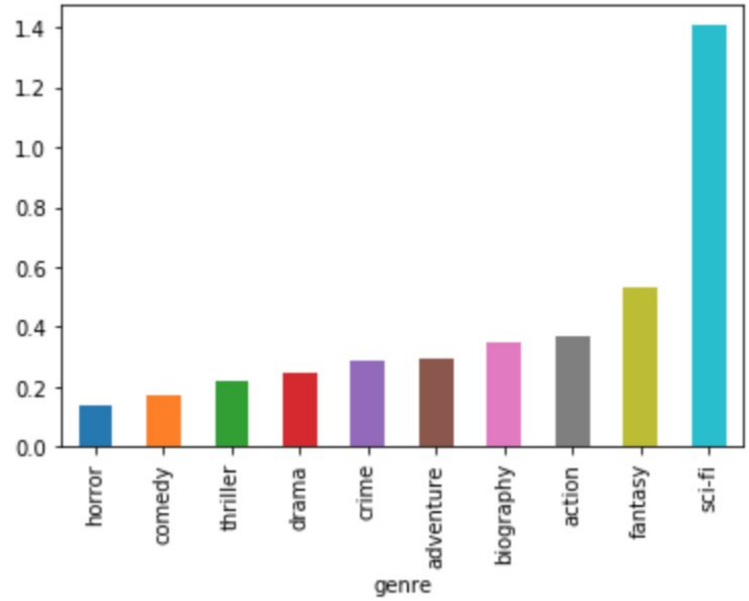


Next, we assessed average scores over time. The plot below suggests no obvious pattern, except perhaps that scores tend to be higher (suggesting greater imbalance) after the 1960s than pre-1960s. However, it is important to note that the original dataset features far more movies from the post-1960s period than the pre-1960s period. As such, it is not clear whether this finding represents a true difference, or whether it is simply a result of differences in sample size.



Finally, we reviewed average scores by genre. Sci-fi receives significantly higher scores on average than any other genre. Based on the results from the power and agency analysis, as well as the network analysis, it is likely that this imbalance falls in favor of female characters. Further research is necessary to better understand the ways in which the portrayal of gender in sci-fi differs from that of other genres – as well as the reasons for this difference.

It is also important to note that horror receives the lowest scores on average. The fact that it receives such a low score – suggesting relative gender balance – aligns well with previous work that has shown that horror is the only genre in which women appear and speak as often as men, and it is also the genre that is most likely to pass the Bechdel Test.


horror
comedy
thriller
drama
crime
adventure
biography
action
fantasy
sci-fi
genre