




Gender in Movie Scripts

Jasmin Dial, Emma Peterson,
Joan Wang, Regina Widjaya



Agenda

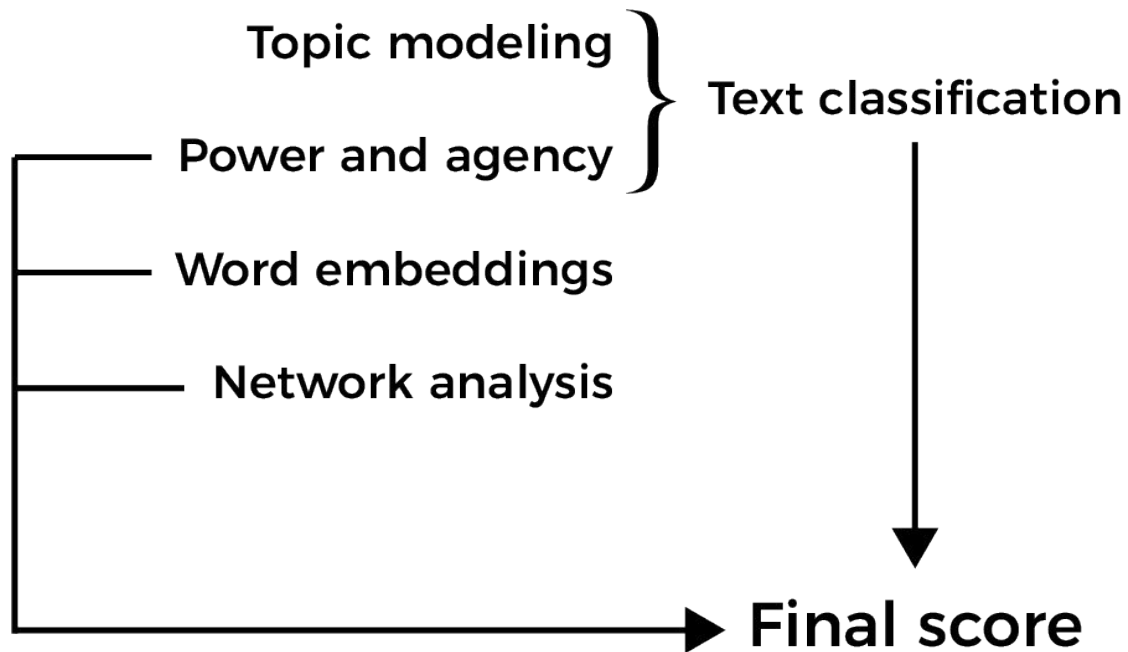
- Introduction
- Project structure
- Sub-projects
 - Topic modeling
 - Power and agency
 - Text classification
 - Network analysis
 - Word embeddings
- Final score
- Exploratory analysis

Introduction

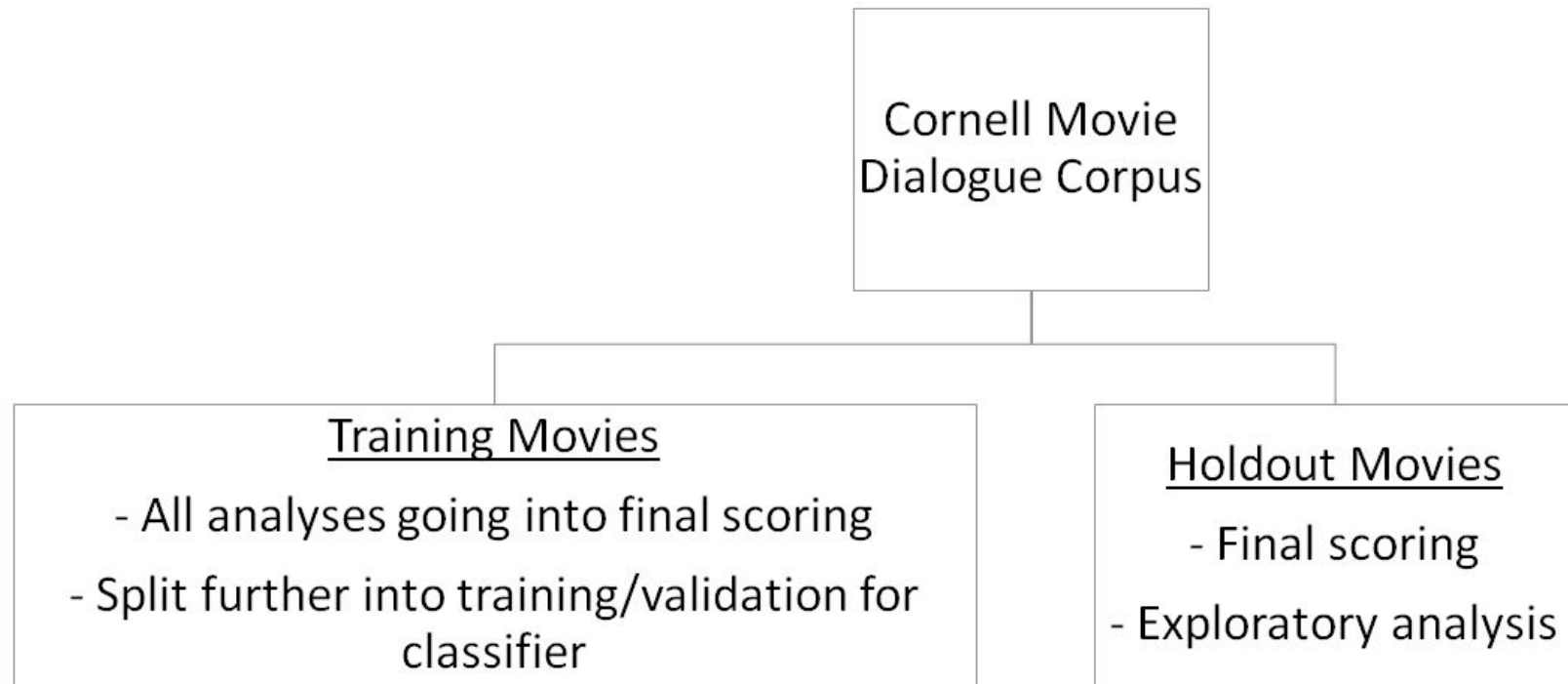
Data: Cornell Movie Dialogs Corpus

- 220,000+ conversations
- 9,000+ characters
- 600+ movies
- Metadata: genre, release year, IMDB rating and number of votes, gender

Project structure



Separation of data



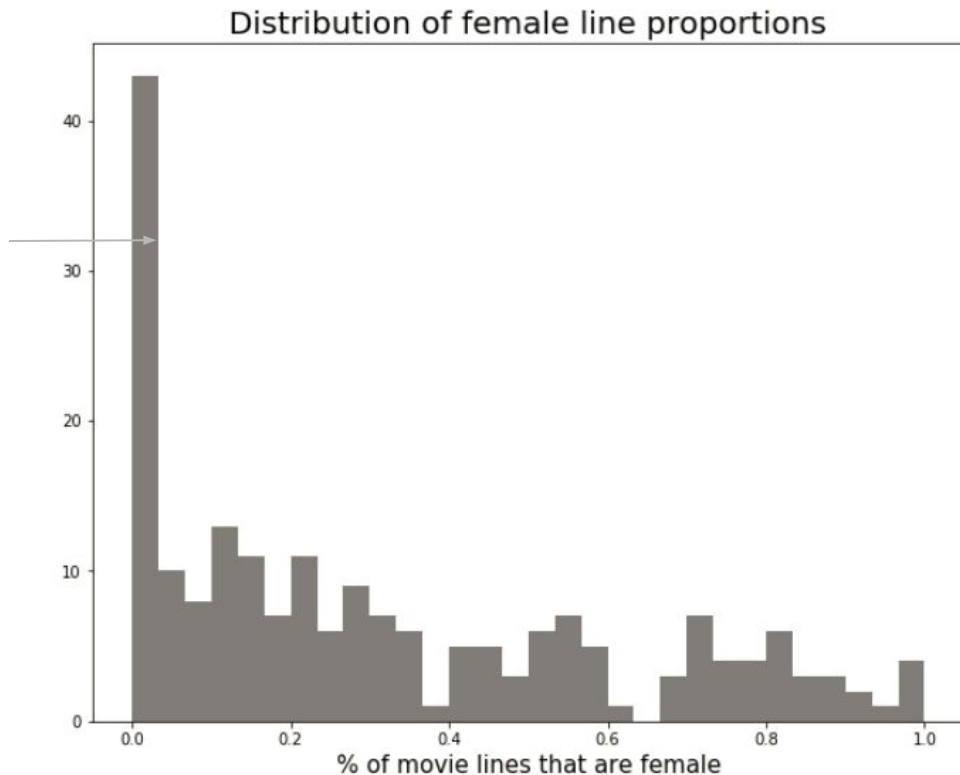


Proportion of Lines by Gender



Proportion of female lines

Many movies
have little to no
female lines



Lowest proportion female lines

	movie_title	movie_year	genre	pct_female
0	invaders from mars	1953	['horror', 'sci-fi']\n	0.0
1	confidence	2003	['crime', 'thriller']\n	0.0
2	mission: impossible ii	2000	['action', 'adventure', 'thriller']\n	0.0
3	mimic	1997	['drama', 'horror', 'sci-fi']\n	0.0
4	halloween iii: season of the witch	1982	['horror', 'mystery', 'sci-fi']\n	0.0
5	star trek vi: the undiscovered country	1991	['action', 'mystery', 'sci-fi', 'thriller']\n	0.0
6	leviathan	1989	['adventure', 'horror', 'mystery', 'sci-fi', 'thriller']\n	0.0
7	glengarry glen ross	1992	['drama']\n	0.0
8	frequency	2000	['crime', 'drama', 'sci-fi', 'thriller']\n	0.0
9	the french connection	1971	['action', 'crime', 'thriller']\n	0.0

Highest proportion female lines

	movie_title	movie_year	genre	pct_female
3	agnes of god	1985	['drama', 'mystery', 'thriller']\n	0.979695
4	the horse whisperer	1998	['drama', 'romance', 'western']\n	0.949054
5	a nightmare on elm street: the dream child	1989	['fantasy', 'horror', 'thriller']\n	0.929412
6	playback	1996	['thriller']\n	0.926027
7	white angel	1994	['drama', 'thriller']\n	0.897959
8	cruel intentions	1999	['drama', 'romance', 'thriller']\n	0.889665
9	someone to watch over me	1987	['action', 'crime', 'drama', 'romance', 'thriller']\n	0.880000

Topic Modeling

Topic Modeling

- Process
 - Remove infrequent words and non-meaningful POS tags
 - Create bigrams and trigrams
 - Separate training data into gender, genre-specific lines
 - Fit topics for each subset of training data
 - Choose model based on coherence score

Sample of Topics

Trained on action, female

```
[ (7,
  '0.059*"mr" + 0.054*"well" + 0.043*"look" + 0.035*"sulu" + 0.031*"order" + 0.028*"thruster" + 0.028*"aft" + 0.017*"miss_teschmacher" + 0.014*"job" + 0.012*"lieutenant"'),
  (9,
    '0.097*"think" + 0.056*"way" + 0.034*"find" + 0.033*"always" + 0.032*"mr_peel" + 0.018*"meet" + 0.018*"starfleet" + 0.017*"worry" + 0.015*"land" + 0.014*"bit"'),
  (6,
    '0.048*"something" + 0.040*"please" + 0.031*"day" + 0.030*"life" + 0.021*"first" + 0.020*"run" + 0.020*"real" + 0.019*"maybe" + 0.018*"year" + 0.015*"stop"'),
  (11,
    '0.040*"space" + 0.036*"honor" + 0.031*"really" + 0.030*"help" + 0.027*"jim" + 0.017*"send" + 0.017*"show" + 0.016*"new" + 0.014*"four" + 0.014*"car"'),
  (8,
    '0.094*"sir" + 0.037*"father" + 0.037*"yes" + 0.024*"still" + 0.019*"anything" + 0.017*"power" + 0.014*"thank" + 0.014*"without" + 0.014*"break" + 0.012*"great"'),
```

Sample of Topics

Trained on action, male

```
[(7,
  '0.186*"light" + 0.096*"yes" + 0.032*"put" + 0.031*"hear" + 0.028*"course" + 0.023*"mr" + 0.016*"real" + 0.014*"dinner" + 0.014*"together" + 0.009*"truth"'),
 (4,
  '0.088*"first" + 0.075*"time" + 0.045*"kill" + 0.042*"way" + 0.042*"people" + 0.038*"work" + 0.021*"everything" + 0.020*"name" + 0.018*"wait" + 0.018*"big"'),
 (3,
  '0.053*"little" + 0.037*"find" + 0.037*"john" + 0.031*"day" + 0.022*"girl" + 0.019*"stupid" + 0.018*"always" + 0.016*"someone" + 0.015*"true" + 0.014*"lose"'),
 (2,
  '0.087*"would" + 0.046*"believe" + 0.029*"life" + 0.021*"sex" + 0.020*"mother" + 0.016*"cynthia" + 0.015*"long" + 0.014*"sick" + 0.013*"money" + 0.013*"die"'),
 (8,
  '0.126*"mean" + 0.071*"place" + 0.035*"house" + 0.022*"woman" + 0.020*"man" + 0.018*"car" + 0.017*"suppose" + 0.016*"must" + 0.016*"many" + 0.015*"guy"'),
```


Features for Classification

- Used score for probability of topic given a document/line

	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 & Agency



Power & Agency Analysis

- Verbs are rated as positive/negative agency and positive/negative power
 - “abolishes” → + agency, + power
 - “awaits” → – agency, – power

	+ 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

Power & Agency Analysis

	gender_from_f	genre	power_pos_prop_f	gender_from_m	power_pos_prop_m	diff
8	f	sci-fi	0.292281	m	0.242271	-0.050010
1	f	adventure	0.248956	m	0.233058	-0.015897
2	f	biography	0.256775	m	0.254805	-0.001970
7	f	horror	0.256447	m	0.271171	0.014724
5	f	drama	0.242133	m	0.256999	0.014866
3	f	comedy	0.252291	m	0.269248	0.016958
0	f	action	0.274512	m	0.293866	0.019354
6	f	fantasy	0.246974	m	0.270436	0.023462
9	f	thriller	0.260652	m	0.314656	0.054003
4	f	crime	0.230964	m	0.289269	0.058305

Text Classification

Model specs

- Classifier: Multinomial Naive Bayes
 - Alpha = 0.5
 - Bag of words
- Outcome variable: speaker gender (“gender_from”)
- Other specifications we tried:
 - Separate models by genre
 - Outcome variable as speaker pair (“gender_from” & “gender_to”):
MM, MF, FM, FF

Model specs (cont.)

- Additional non-text features
 - Probability-based score of top three topics by gender
 - Proportion of verbs that are positive/negative power and agency

```
X_holdout[FEATURE_COLS].head()
```

	words	agency_pos_prop	power_pos_prop	agency_neg_prop	power_neg_prop	MT1	MT2	MT3	WT1	WT2	WT3
0	thanks miss	0.0	0.0	1.0	1.0	6.131145	6.841707	6.666505	4.747797	4.686748	4.755125
1	you kind i amanda	0.0	0.0	0.0	0.0	6.211220	6.929817	6.751709	4.744825	4.684112	4.752109
2	right well thanks drink stuff amanda reason me stick around part anymore	1.0	0.5	0.0	0.0	6.347796	7.081794	6.892825	7.110352	4.904115	4.974433
3	glum hawk night still young fill plenty compensatory possibility	0.0	0.0	0.0	0.0	7.570643	8.195575	6.987399	5.852607	5.792455	6.897112
4	huh	0.0	0.0	0.0	0.0	6.131144	6.841705	7.666502	4.662018	4.601227	4.668652

Performance on holdout set

Accuracy: 67%

Precision: 42%

Recall: 25%

Confusion matrix:

	Pred M	Pred F
Actual M	30,530	5,529
Actual F	11,861	4,021

... But what's the baseline?

The most “male” lines

	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..."	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...	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.	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.	1997	0.994079	['drama']\n
4	grand hotel	I'm Baron von Gaigern.	1932	0.993760	['drama', 'romance']\n

The most “female” lines

	movie_title	text	movie_year	female_prob	genre
0	leviathan	Your suit, Becky!\n	1989	0.987255	['adventure', 'horror', 'mystery', 'sci-fi', 'thriller']\n
1	ghost ship	Maureen?\n	2002	0.985189	['horror', 'mystery', 'thriller']\n
2	ghost ship	Maureen.\n	2002	0.985189	['horror', 'mystery', 'thriller']\n
3	ghost ship	Maureen.\n	2002	0.985189	['horror', 'mystery', 'thriller']\n
4	ghost ship	Let's not be too hasty.\n	2002	0.975697	['horror', 'mystery', 'thriller']\n
5	frances	Claire?\n	1982	0.985081	['biography', 'drama']\n
6	a nightmare on elm street 3: dream warriors	You like gymnastics?\n	1987	0.982592	['fantasy', 'horror', 'thriller']\n
7	back to the future	It's polyester.\n	1985	0.982228	['adventure', 'family', 'sci- fi']\n
8	the time machine	She's gotten into your equations.\n	2002	0.981690	['sci-fi', 'adventure', 'action']\n

The most “male” movies

	avg_prop_female_lines	avg_male_prob	movie_title	movie_year	genre
0	0.034921	0.937486	true believer	1989	['drama', 'crime']\n
1	0.068337	0.936244	apocalypse now	1979	['drama', 'war']\n
2	0.000000	0.933414	glengarry glen ross	1992	['drama']\n
3	0.152695	0.933175	do the right thing	1989	['drama']\n
4	0.322251	0.932354	the majestic	2001	['drama', 'romance']\n
5	0.421941	0.931982	one flew over the cuckoo's nest	1975	['drama']\n
6	0.101617	0.931890	nothing but a man	1964	['drama', 'romance']\n
7	0.436293	0.931512	affliction	1997	['drama', 'mystery', 'thriller']\n
8	0.024691	0.931115	the verdict	1982	['drama']\n
9	0.346705	0.930441	the day the earth stood still	2008	['drama', 'sci-fi', 'thriller']\n

The most “female” movies

	avg_prop_female_lines	avg_female_prob	movie_title	movie_year	genre
0	1.000000	0.860538	i walked with a zombie	1943	['horror']\n
1	0.183036	0.845237	le grand bleu	1988	['adventure', 'drama', 'romance']\n
2	0.333333	0.842549	friday the 13th part iii	1982	['horror']\n
3	0.800752	0.841393	i still know what you did last summer	1998	['horror', 'mystery', 'thriller']\n
4	0.569519	0.840736	cherry falls	2000	['horror', 'mystery', 'thriller']\n
5	0.667470	0.839264	the wizard of oz	1939	['adventure', 'family', 'fantasy', 'musical']\n
6	0.450746	0.839212	ghostbusters	1986	['animation', 'comedy', 'fantasy', 'sci-fi', 'horror']\n
7	0.202247	0.838664	back to the future	1985	['adventure', 'family', 'sci-fi']\n
8	0.241379	0.838109	the beach	2000/1	['adventure', 'drama', 'romance', 'thriller']\n
9	0.804800	0.837933	the curse	1987	['sci-fi', 'horror']\n

Network Analysis

Network Analysis

- **Degree centrality:** fraction of nodes to which a given node is connected
- **Betweenness centrality:** measure of how often a given node acts as a bridge along the shortest path between two other nodes

	degree	betweenness
female	0.240	0.105
male	0.289	0.161

Network Analysis

By genre

	genre	degree_f	degree_m	diff		genre	betweenness_f	betweenness_m	diff
8	sci-fi	0.571970	0.244444	-0.327525	8	sci-fi	0.454798	0.070343	-0.384455
6	fantasy	0.313065	0.251511	-0.061554	6	fantasy	0.199763	0.110678	-0.089085
7	horror	0.298368	0.282734	-0.015634	7	horror	0.142633	0.120228	-0.022404
1	adventure	0.281241	0.311114	0.029873	9	thriller	0.122226	0.137093	0.014866
2	biography	0.158539	0.191420	0.032881	5	drama	0.115159	0.143230	0.028071
9	thriller	0.214631	0.248342	0.033711	3	comedy	0.089080	0.152570	0.063490
5	drama	0.238794	0.284000	0.045206	1	adventure	0.085731	0.153880	0.068148
3	comedy	0.206780	0.255542	0.048762	2	biography	0.064815	0.139327	0.074512
0	action	0.247743	0.332621	0.084878	0	action	0.092217	0.200529	0.108313
4	crime	0.187315	0.304975	0.117660	4	crime	0.048507	0.205693	0.157186

Network Analysis

By movie: degree centrality

	genre	degree_f	degree_m	diff	movie_title_f
182	drama	0.928571	0.112245	-0.816327	contact
135	sci-fi	1.000000	0.250000	-0.750000	arcade
98	drama	0.857143	0.309524	-0.547619	mimic

	genre	degree_f	degree_m	diff	movie_title_f
159	crime	0.138889	0.888889	0.750000	vertigo
177	drama	0.250000	0.750000	0.500000	solaris
13	crime	0.250000	0.625000	0.375000	crash

Network Analysis

By movie: betweenness centrality

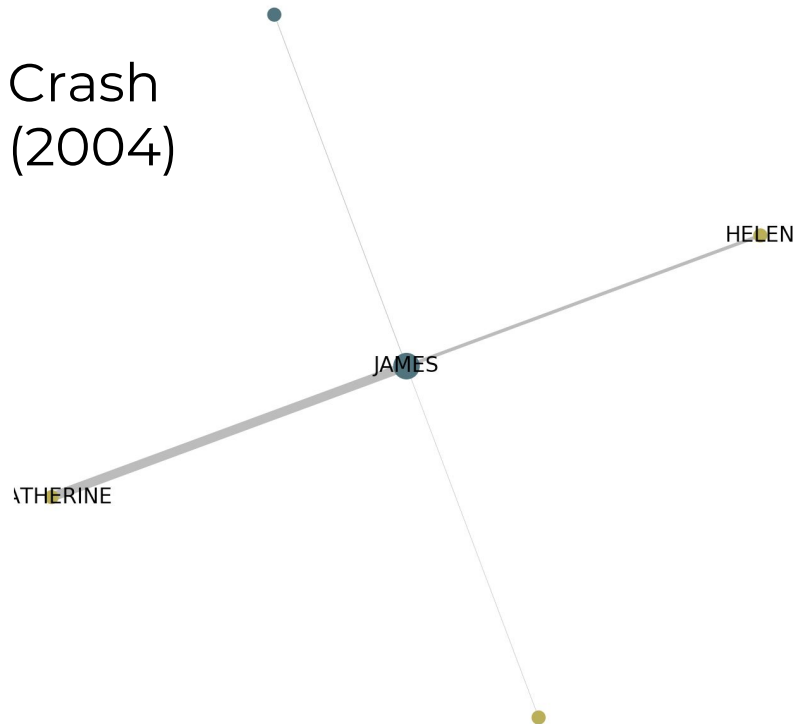
	genre	betweenness_f	betweenness_m	diff	movie_title_f
182	drama	0.945055	0.021193	-0.923862	contact
135	sci-fi	0.875000	0.004464	-0.870536	arcade
145	drama	0.750000	0.025000	-0.725000	white angel

	genre	betweenness_f	betweenness_m	diff	movie_title_f
159	crime	0.055556	0.972222	0.916667	vertigo
13	crime	0.000000	0.500000	0.500000	crash
36	action	0.000000	0.500000	0.500000	the rock

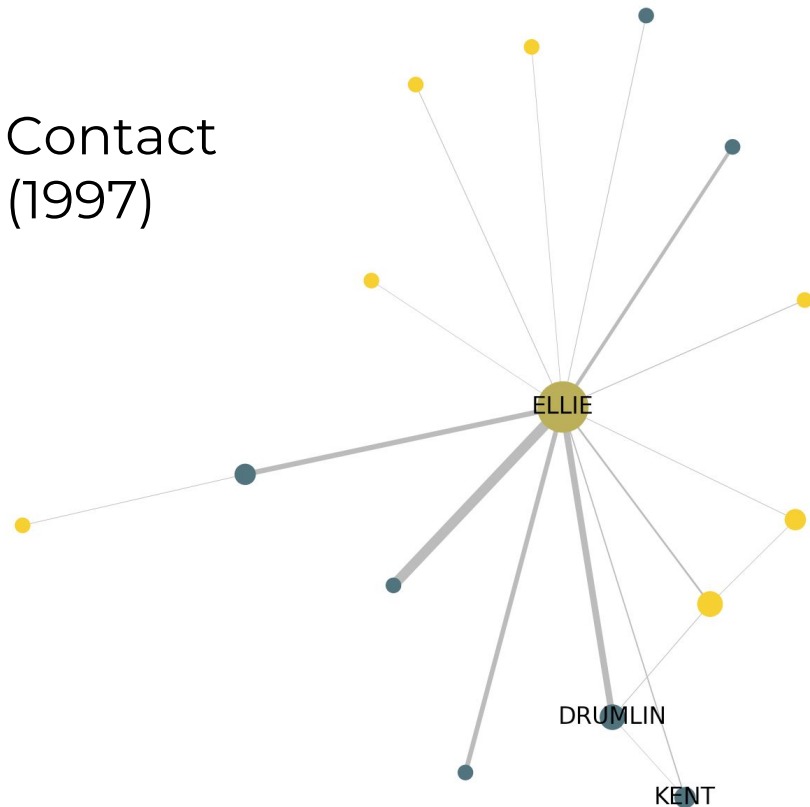
Network Analysis

- Female
- Male
- Unknown

Crash
(2004)



Contact
(1997)

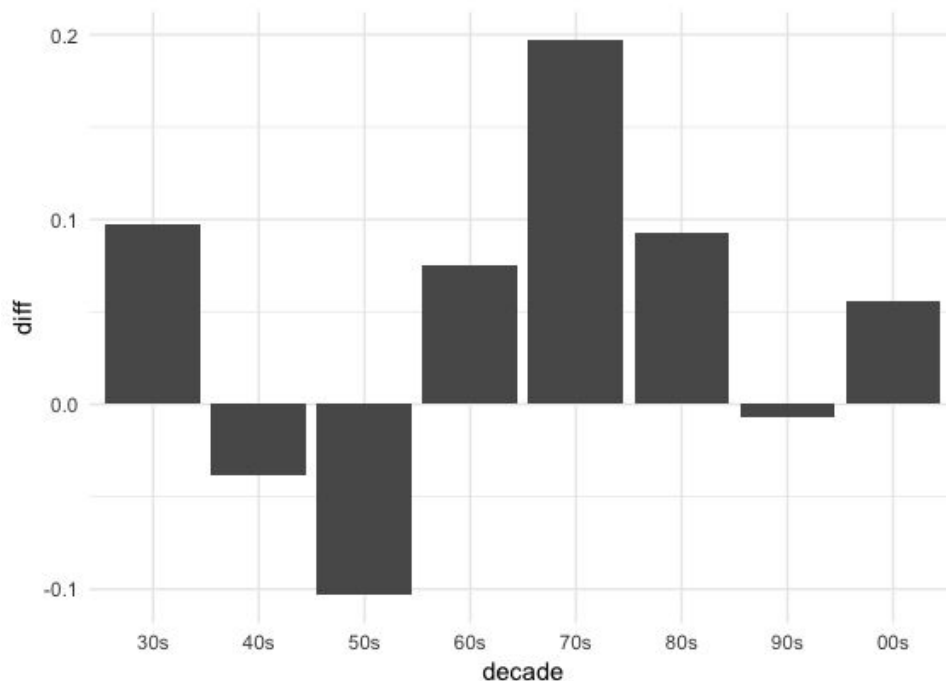


Network Analysis

IMDB ratings

- **Top 20 movies for women: 6.27**
- **Bottom 20 movies for women: 6.92**

Year



Word Embeddings

Word Embeddings

TF-IDF Vectorizer

- **TF: Term Frequency**, which measures how frequently a term occurs in a document.

$$TF(t) = \frac{\text{(Number of times term } t \text{ appears in a document)}}{\text{(Total number of terms in the document)}}$$

- **IDF: Inverse Document Frequency**, which measures how important a term is.

$$IDF(t) = \log_e \left(\frac{\text{Total number of documents}}{\text{Number of documents with term } t \text{ in it}} \right)$$

Cosine Similarity Score

Measuring the **degree of similarity** between two documents, without orders and context.

	count	mean	std	min	25%	50%	75%	max
genre								
action	151.0	0.568041	0.341886	0.000000	0.360406	0.715904	0.831122	1.000000
adventure	27.0	0.576467	0.290256	0.000000	0.482959	0.698670	0.778962	0.889845
animation	10.0	0.492413	0.390296	0.000000	0.046342	0.713862	0.814783	0.846943
biography	23.0	0.588373	0.350862	0.000000	0.378852	0.763138	0.877579	0.945393
comedy	118.0	0.679655	0.314736	0.000000	0.653381	0.829528	0.882248	0.956988
crime	67.0	0.605394	0.358645	0.000000	0.427927	0.783201	0.869661	0.942188
documentary	3.0	0.532347	0.461164	0.000000	0.393635	0.787270	0.798521	0.809772
drama	137.0	0.672491	0.272052	0.000000	0.580605	0.774413	0.858866	0.950163
family	1.0	0.748310	NaN	0.748310	0.748310	0.748310	0.748310	0.748310
fantasy	14.0	0.478423	0.310591	0.000000	0.296773	0.505159	0.737766	0.852766
film-noir	1.0	0.851644	NaN	0.851644	0.851644	0.851644	0.851644	0.851644
horror	38.0	0.611660	0.282075	0.000000	0.584903	0.716222	0.802254	0.889793
mystery	5.0	0.549305	0.407814	0.000000	0.225655	0.796920	0.830960	0.892990
romance	2.0	0.892056	0.007455	0.886785	0.889421	0.892056	0.894692	0.897328
sci-fi	5.0	0.292383	0.400476	0.000000	0.000000	0.000000	0.717454	0.744463
short	4.0	0.460928	0.420272	0.000000	0.164568	0.476412	0.772772	0.890888
thriller	10.0	0.758324	0.137617	0.539559	0.685775	0.777634	0.874783	0.912970

Cosine Similarity Score

Highest Scores

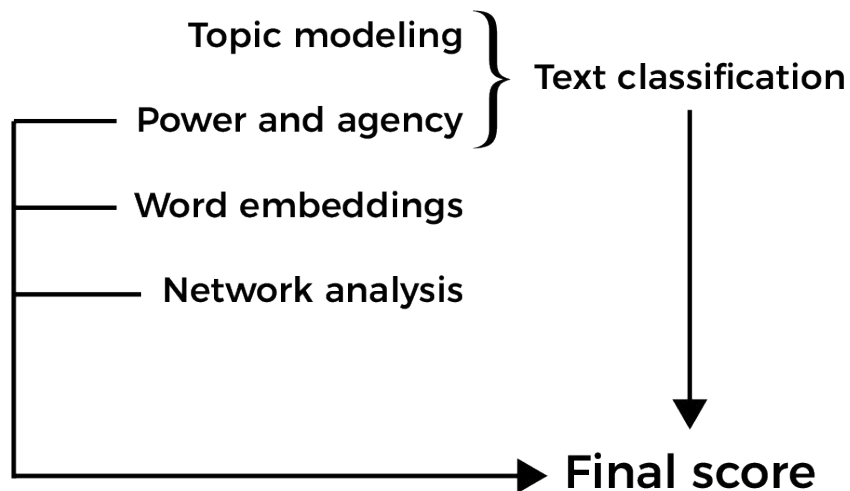
	movie_id	genre	year	gender_cosim	movie_title
410	m469	comedy	1986	0.956988	peggy sue got married
542	m588	drama	1973	0.950163	u-turn
154	m238	drama	1950	0.948640	all about eve
312	m380	comedy	1986	0.946351	hannah and her sisters
206	m285	comedy	1987	0.946135	broadcast news
166	m249	comedy	1997	0.945672	as good as it gets
210	m289	biography	1995	0.945393	casino
161	m244	drama	2001	0.944203	the anniversary party
460	m513	drama	1975	0.943513	shampoo
8	m105	crime	1997	0.942188	jackie brown

Lowest Scores

	movie_id	genre	year	gender_cosim	movie_title
7	m104	biography	1991	0.001068	jfk
523	m570	action	1999	0.001263	three kings
389	m45	crime	2003	0.001552	confidence
21	m118	fantasy	1985	0.002015	legend
75	m167	crime	1954	0.002308	rear window
448	m502	action	1998	0.002344	saving private ryan
558	m601	drama	2000	0.002365	what lies beneath
467	m52	comedy	1933	0.002829	duck soup
241	m316	comedy	1993	0.002965	dave
485	m536	comedy	1964	0.003140	dr. strangelove or: how i learned to stop worr...

Final Scoring

Final Scoring Method



- Final score = an approximate measure of degree of “genderedness” is a given movie
- High → more gender disparity, low → less gender disparity
- Relative ranking is more meaningful than raw score
- Each component is normalized by training mean and sd for scale similarity

Final Score Components

```
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
```

1. Abs diff in proportion of lines that are M vs F



Final Score Components

```
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
```

1. Abs diff in proportion of lines that are M vs F

2. Reversed cosine similarity between M & F word embeddings

Final Score Components

```
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
```

1. Abs diff in proportion of lines that are M vs F

2. Reversed cosine similarity between M & F word embeddings

3. Avg M class probability among M lines



Final Score Components

```
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
```

1. Abs diff in proportion of lines that are M vs F
2. Reversed cosine similarity between M & F word embeddings
3. Avg M class probability among M lines
4. Avg F class probability among F lines

Final Score Components

```
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
```

1. Abs diff in proportion of lines that are M vs F
2. Reversed cosine similarity between M & F word embeddings
3. Avg M class probability among M lines
4. Avg F class probability among F lines
5. Abs diff in avg network degree centrality of M & F characters

Final Score Components

```
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
```

1. Abs diff in proportion of lines that are M vs F
2. Reversed cosine similarity between M & F word embeddings
3. Avg M class probability among M lines
4. Avg F class probability among F lines
5. Abs diff in avg network degree centrality of M & F characters
6. Abs diff in avg network betweenness centrality of M & F characters

Final Score Components

```
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
```

```
#####
```

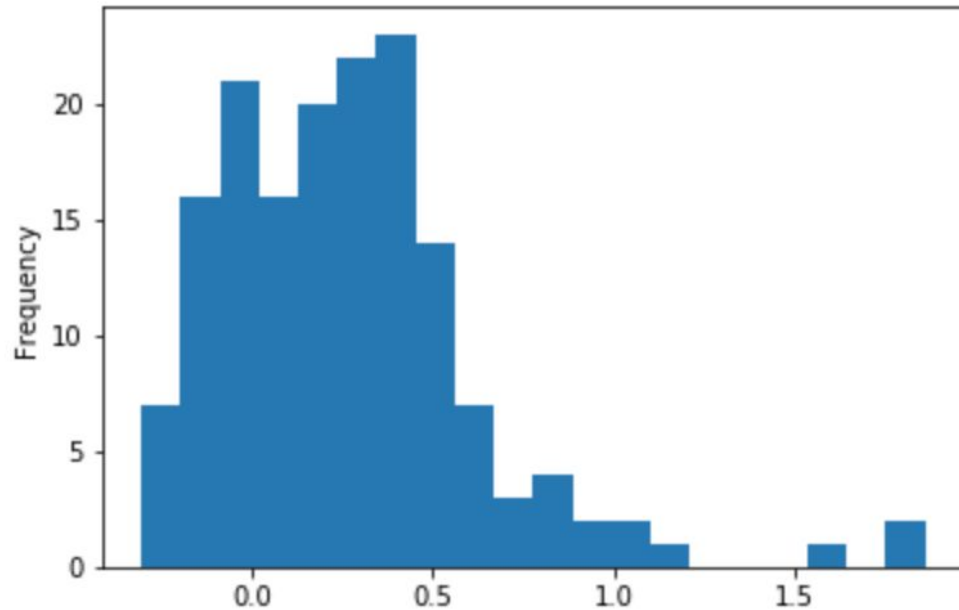
```
Final score: -0.34
```

1. Abs diff in proportion of lines that are M vs F
2. Reversed cosine similarity between M & F word embeddings
3. Avg M class probability among M lines
4. Avg F class probability among F lines
5. Abs diff in avg network degree centrality of M & F characters
6. Abs diff in avg network betweenness centrality of M & F characters

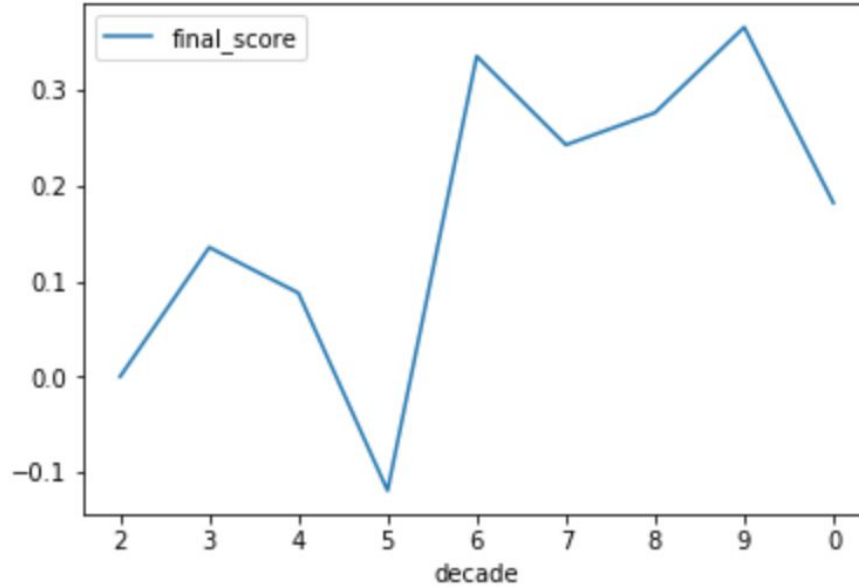
Final score: equality-weighted average of the six components

Exploratory Insights

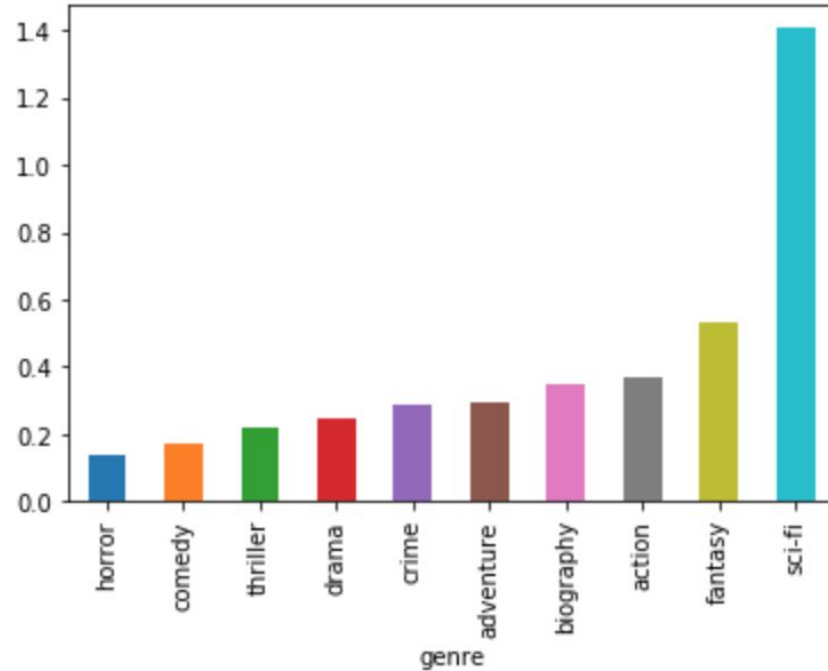
Distribution of final scores



Final score over time



Final score by genre



Thanks!