IR – Text Indexing

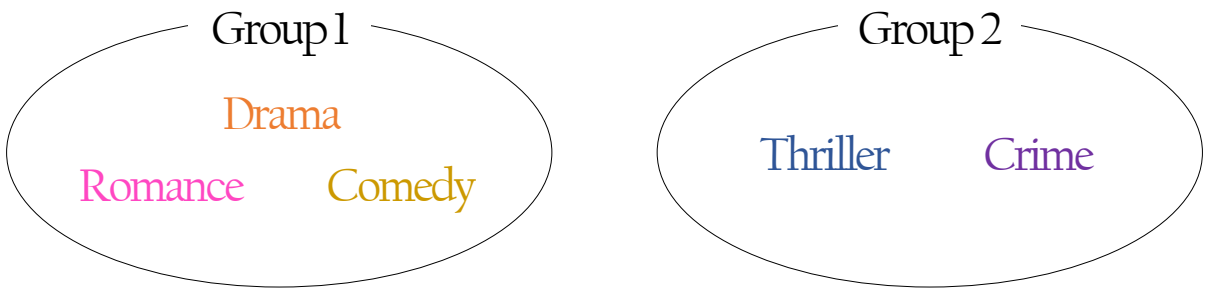
NLP and Information Retrieval Class

20191181 Seunguk Yu

1. Assumption
2. Data Crawling & Preprocessing
3. Indexing
4. Result & Conclusion

1. Assumption

1) The word distribution will be similar within each group according to the category.



2) If the frequency is high within each category, the tf-idf value will generally be high.

2. Data Crawling & Preprocessing

Drama Movie Scripts

12 and Holding (2004-04 Draft)
Written by Anthony Cipriano

12 Monkeys (1994-06 Draft)
Written by David Peoples,Janet Peoples

12 Years a Slave (Undated Draft)
Written by John Ridley

127 Hours (Undated Draft)
Written by Simon Beaufoy,Danny Boyle

1492: Conquest of Paradise (1991-09 Draft)
Written by Roslyne Bosch

17 Again (2007-10 Draft)
Written by Jason Filardi

187 (1996-11 Draft)
Written by Scott Yagemann

2012 (2008-02 Second draft)
Written by Roland Emmerich,Harald Kloser

25th Hour (2001-04 Draft)
Written by David Benioff

28 Days Later (Undated Draft)
Written by Alex Garland

42 (2012-07 Revised draft)
Written by Brian Helgeland

44 Inch Chest (Undated Draft)
Written by Louis Mellis,David Scinto

50-50 (2008-07 Draft)
Written by Will Reiser

500 Days of Summer (Undated First draft)
Written by Scott Neustadter,Michael H. Weber

8 Mile (2001-04 Draft)
Written by Scott Silver

Romance Movie Scripts

10 Things I Hate About You (1997-11 Draft)
Written by Karen McCullah Lutz,Kirsten Smith,William Shakespeai

17 Again (2007-10 Draft)
Written by Jason Filardi

500 Days of Summer (Undated First draft)
Written by Scott Neustadter,Michael H. Weber

Adjustment Bureau, The (Undated Draft)
Written by George Nolfi,Philip K Dick

Adventures of Buckaroo Banzai Across the Eighth Dimension, The
Written by Earl Mac Rauch

Airplane (1979-06 Shooting draft)
Written by Jim Abrahams, David Zucker,Jerry Zucker

Almost Famous (1998-12 Draft)
Written by Cameron Crowe

American President, The (Undated Draft)
Written by Aaron Sorkin

American Werewolf in London (Undated Draft)
Written by John Landis

Amour (Undated Draft)
Written by Michael Haneke

Angel Eyes (1999-10 Draft)
Written by Gerald DiPego

Annie Hall (Undated Draft)
Written by Woody Allen,Marshall Brickman

Artist, The (Undated Draft)
Written by Michel Hazanavicius

As Good As It Gets (Undated Draft)
Written by Mark Andrus,James L. Brooks

Autumn in New York (Undated Shooting draft)
Written by Allison Burnett

Comedy Movie Scripts

10 Things I Hate About You (1997-11 Draft)
Written by Karen McCullah Lutz,Kirsten Smith,William Shak

12 (Undated Draft)
Written by Lawrence Bridges

17 Again (2007-10 Draft)
Written by Jason Filardi

30 Minutes or Less (2009-12 Draft)
Written by Michael Diliberti,Matthew Sullivan

48 Hrs. (Undated Draft)
Written by Steven E. De Souza,Walter Hill,Roger Spottiswo

50-50 (2008-07 Draft)
Written by Will Reiser

500 Days of Summer (Undated First draft)
Written by Scott Neustadter,Michael H. Weber

A Serious Man (2007-06 Draft)
Written by Joel Coen,Ethan Coen

Ace Ventura: Pet Detective (Undated Draft)
Written by Jack Bernstein,Tom Shadyac,Jim Carrey

Adaptation (2000-11 Draft)
Written by Charlie Kaufman,Donald Kaufman

Addams Family, The (1991-04 Shooting Script)
Written by Charles Addams,Caroline Thompson

Adventures of Buckaroo Banzai Across the Eighth Dimensi
Written by Earl Mac Rauch

After School Special (2000-01 Draft)
Written by David H. Steinberg

Airplane (1979-06 Shooting draft)
Written by Jim Abrahams, David Zucker,Jerry Zucker

Airplane 2: The Sequel (1982-02 Draft)
Written by Ken Finkleman

Thriller Movie Scripts

12 Monkeys (1994-06 Draft)
Written by David Peoples,Janet Peoples

127 Hours (Undated Draft)
Written by Simon Beaufoy,Danny Boyle

15 Minutes (Undated Draft)
Written by John Hertzfield

2012 (2008-02 Second draft)
Written by Roland Emmerich,Harald Kloser

48 Hrs. (Undated Draft)
Written by Steven E. De Souza,Walter Hill,Roger Spottiswo

8MM (1997-05 Draft)
Written by Andrew Kevin Walker

A Few Good Men (1991-07 Revised draft)
Written by Aaron Sorkin

Absolute Power (1996-05 Draft)
Written by David Baldacci,William Goldman

Abyss, The (1988-08 Draft)
Written by James Cameron

Adjustment Bureau, The (Undated Draft)
Written by George Nolfi,Philip K Dick

After.Life (2008-07 Draft)
Written by Agnieszka Wojtowicz-Vosloo

Air Force One (Undated Draft)
Written by Andrew W. Marlowe

Alien (1978-06 Draft)
Written by Walter Hill,David Giler

Alien 3 (1991-01 Draft)
Written by Rex Pickett

Alien vs. Predator (Undated Draft)
Written by Peter Briggs

Crime Movie Scripts

15 Minutes (Undated Draft)
Written by John Hertzfield

25th Hour (2001-04 Draft)
Written by David Benioff

44 Inch Chest (Undated Draft)
Written by Louis Mellis,David Scinto

A Few Good Men (1991-07 Revised draft)
Written by Aaron Sorkin

A Most Violent Year (Undated Draft)
Written by J.C. Chandor

A Prayer Before Dawn (2016-02 Draft)
Written by Jonathan Hirschbein

A Scanner Darkly (Undated Draft)
Written by Charlie Kaufman

Absolute Power (1996-05 Draft)
Written by David Baldacci,William Goldman

Alien Nation (1987-10 Draft)
Written by Rockne O'Bannon,James Cameron

American Gangster (2006-07 Shooting draft)
Written by Steven Zaillian,Mark Jacobson

American History X (1997-02 Draft)
Written by David McKenna

American Hustle (Undated Draft)
Written by Eric Warren Singer,David O. Russell

American, The (2009-05 Second draft)
Written by Rowan Joffe

Analyze That (2002-06 Draft)
Written by Peter Steinfeld,Harold Ramis,Peter Tolan,Kenneth Longeran

Analyze This (1998-07 Draft)
Written by Peter Tolan,Harold Ramis,Kenneth Lonergan

Drama
Movie Counts: 652
Crawlable Scripts: 637

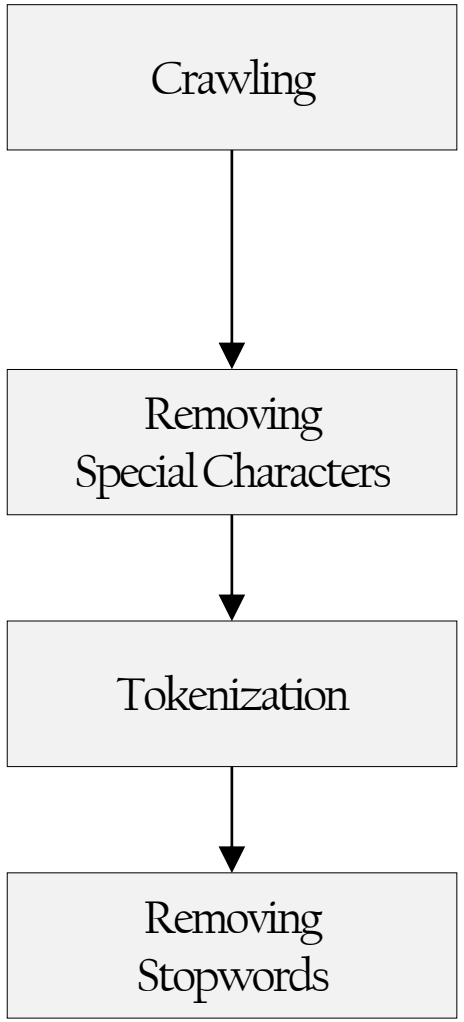
Romance
Movie Counts: 211
Crawlable Scripts: 207

Comedy
Movie Counts: 384
Crawlable Scripts: 372

Thriller
Movie Counts: 406
Crawlable Scripts: 389

Crime
Movie Counts: 231
Crawlable Scripts: 218

2. Data Crawling & Preprocessing



Except special characters and lowercase characters

```
special_characters = ['!', '?', '"','#','$','%','&','(',')','*','+',
                      '/',':',';','<','=','>','@','[','₩₩',']','^',
                      '`','{','|','}','~','₩t', '₩n', '-', ',', '.']
```

Word tokenization by python split() function

Removing english stopwords
from https://www.textfixer.com/tutorials/common-English-words.txt

```
['a', 'able', 'about', 'across', 'after', 'all', 'almost', 'also', 'am', 'among', 'an', 'and', 'any', 'are', 'as', 'at', 'be', 'becaus
e', 'been', 'but', 'by', 'can', 'cannot', 'could', 'dear', 'did', 'do', 'does', 'either', 'else', 'ever', 'every', 'for', 'from', 'ge
t', 'got', 'had', 'has', 'have', 'he', 'her', 'hers', 'him', 'his', 'how', 'however', 'i', 'if', 'in', 'into', 'is', 'it', 'its', 'jus
t', 'least', 'let', 'like', 'likely', 'may', 'me', 'might', 'most', 'must', 'my', 'neither', 'no', 'nor', 'not', 'of', 'off', 'often',
'on', 'only', 'or', 'other', 'our', 'own', 'rather', 'said', 'say', 'says', 'she', 'should', 'since', 'so', 'some', 'than', 'that', 'th
e', 'their', 'them', 'then', 'there', 'these', 'they', 'this', 'tis', 'to', 'too', 'twas', 'us', 'wants', 'was', 'we', 'were', 'what',
'when', 'where', 'which', 'while', 'who', 'whom', 'why', 'will', 'with', 'would', 'yet', 'you', 'your']
```

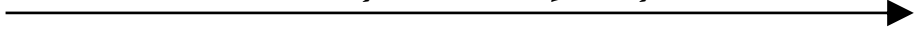
3. Indexing

Build Indexing Terms (example of 'Romance' category)

romance_indexing_terms

```
{'ten': 814,
 'things': 1837,
 'hate': 513,
 'written': 353,
 'karen': 710,
 'mccullah': 3,
 'lutz': 3,
 'kirsten': 20,
 'smith': 214,
 'based': 91,
 "'taming": 2,
 'shrew': 8,
 'william': 1334,
 'shakespeare': 64,
 'revision': 6,
 'november': 159,
 '12': 253,
 '1997': 16,
 'padua': 7,
```

Sort by word frequency



romance_indexing_terms

```
[('up', 22434),
 ('out', 20178),
 ("i'm", 14564),
 ('int', 13512),
 ("it's", 13438),
 ('back', 13369),
 ("don't", 13240),
 ('know', 12444),
 ('looks', 11234),
 ('down', 11201),
 ('one', 10594),
 ('day', 10368),
 ('over', 9704),
 ('now', 9114),
 ('night', 9113),
 ('see', 8917),
 ('room', 8683),
 ('door', 7938),
 ("you're", 7840),
```

3. Indexing

Build Postings Lists (example of 'Romance' category)

```
romance_indexing_terms
```

```
[('up', 22434),
 ('out', 20178),
 ("i'm", 14564),
 ('int', 13512),
 ("it's", 13438),
 ('back', 13369),
 ("don't", 13240),
 ('know', 12444),
 ('looks', 11234),
 ('down', 11201),
 ('one', 10594),
 ('day', 10368),
 ('over', 9704),
 ('now', 9114),
 ('night', 9113),
 ('see', 8917),
 ('room', 8683),
 ('door', 7938),
 ("you're", 7840),
```

Sort by word frequency

```
print(romance_postings_lists, end='')
```

```
{'up': [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 29, 30, 31, 32, 33, 34, 35,
36, 37, 38, 39, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 7
1, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 101, 102, 103, 104, 10
5, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 117, 118, 119, 120, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133,
134, 135, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 156, 157, 158, 159, 160, 161, 16
2, 163, 164, 165, 166, 167, 168, 169, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189,
190, 192, 193, 194, 195, 196, 197, 198, 200, 201, 202, 203, 204, 205, 206], 'out': [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,
15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 5
1, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 8
5, 86, 87, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 11
7, 118, 119, 120, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 137, 138, 139, 140, 141, 142, 143, 144, 145,
146, 147, 148, 149, 150, 151, 152, 153, 154, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 171, 172, 173, 17
4, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 192, 193, 194, 195, 196, 197, 198, 200, 201, 202,
203, 204, 205, 206], "i'm": [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 21, 22, 23, 24, 25, 26, 27, 29, 30, 3
1, 32, 33, 34, 35, 36, 37, 38, 39, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 6
5, 66, 67, 68, 69, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 10
1, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 117, 118, 119, 120, 122, 123, 124, 125, 126, 127, 128, 129,
130, 131, 132, 133, 134, 135, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 156, 157, 15
8, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185,
```

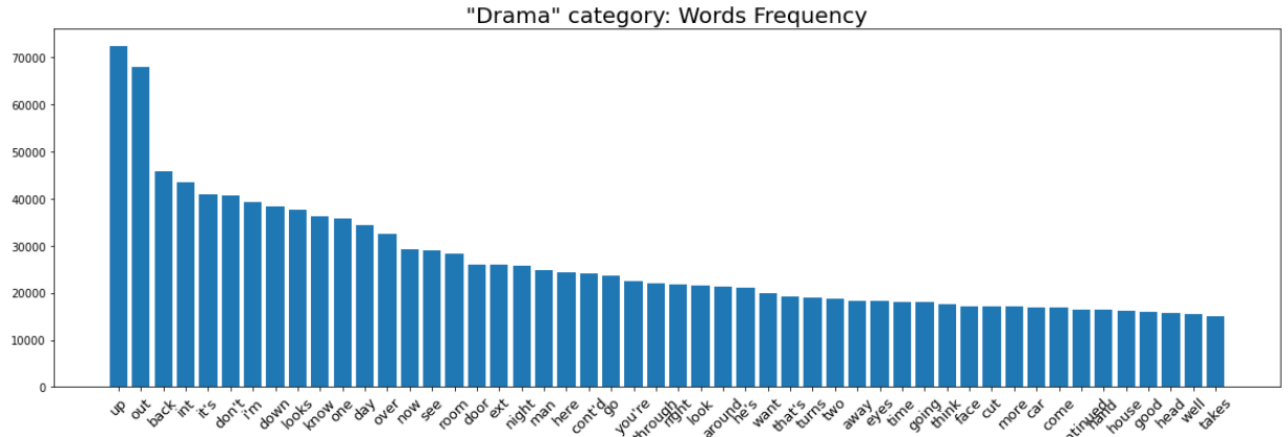
Compute Tf-Idf by using Indexing Terms & Postings List

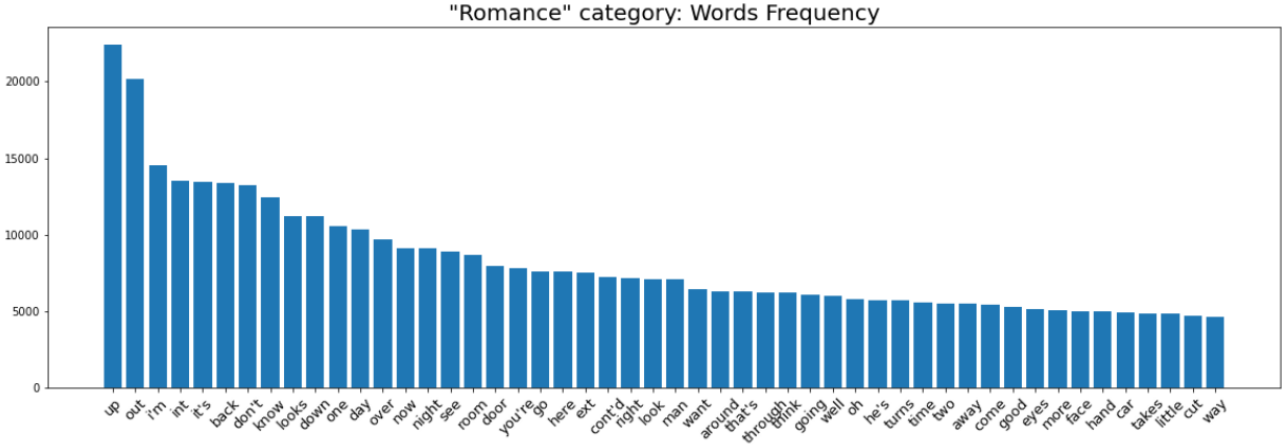
4. Result & Conclusion

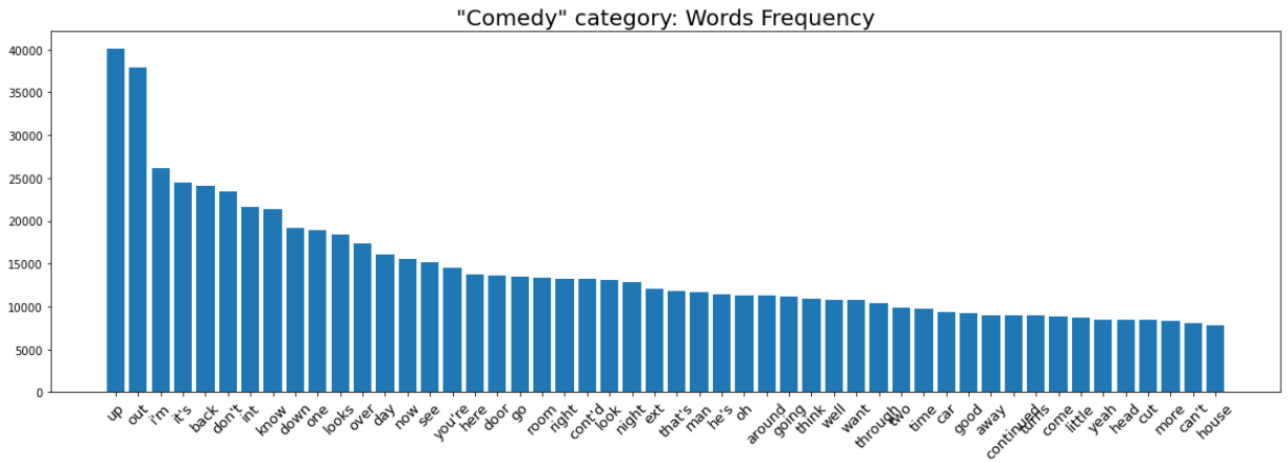
Assumption 1

Group 1: Drama, Romance, Comedy

Mutual Words	Individual Words
up, out, back, it's, don't, looks, day, over, now, see, room, door, cont'd, through, ...	Drama face, takes Romance way, face, Takes Comedy yeah, can't, head, continued





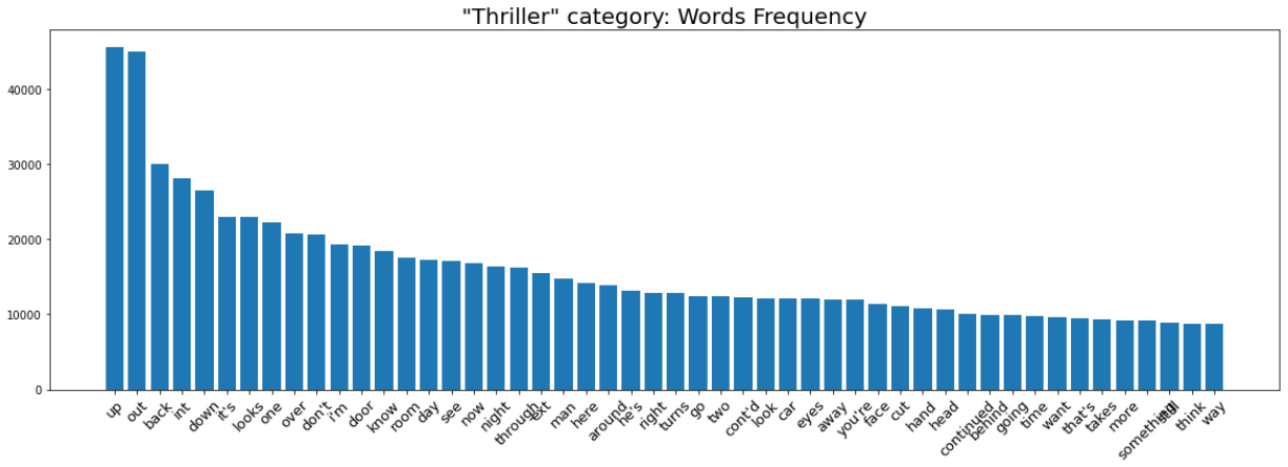


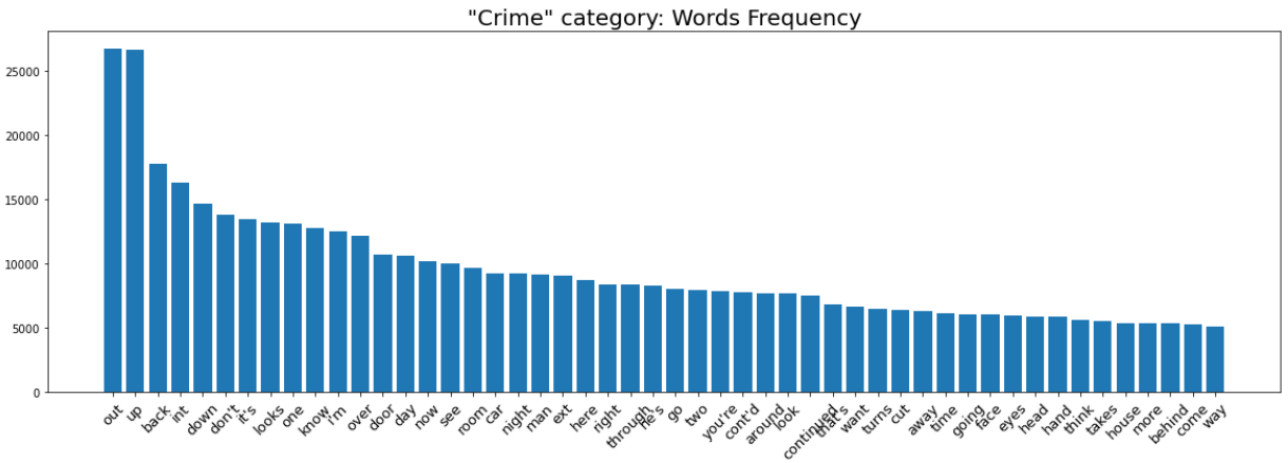
4. Result & Conclusion

Assumption 1

Group 2: Thriller, Crime

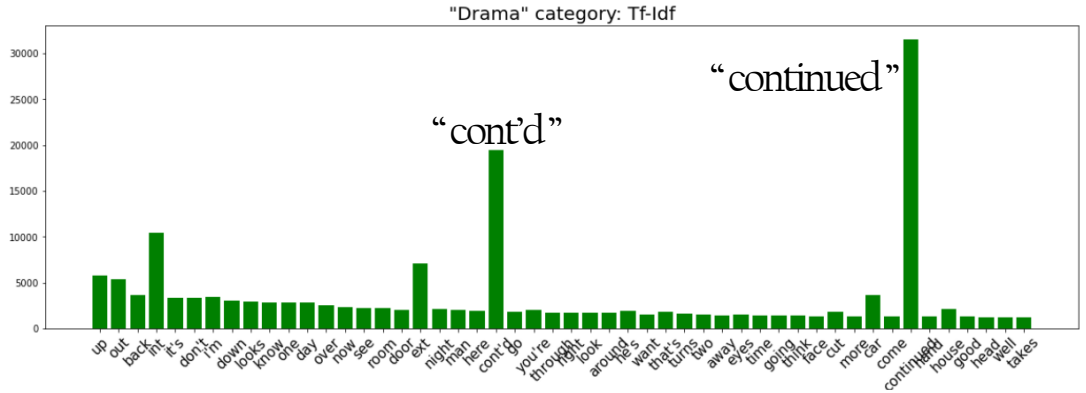
Mutual Words	Individual Words
up, out, back, int, don't, looks, day, over, now, see, room, door, cont'd, through, ...	Thriller still, something Crime house, come

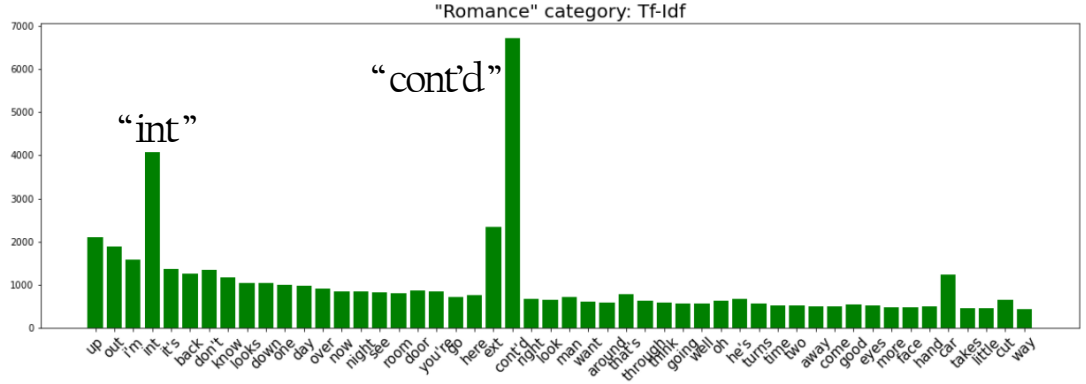


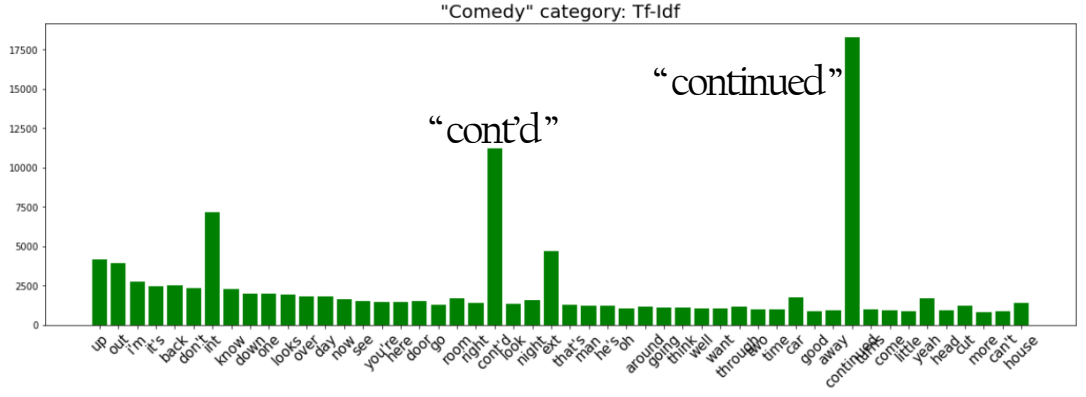


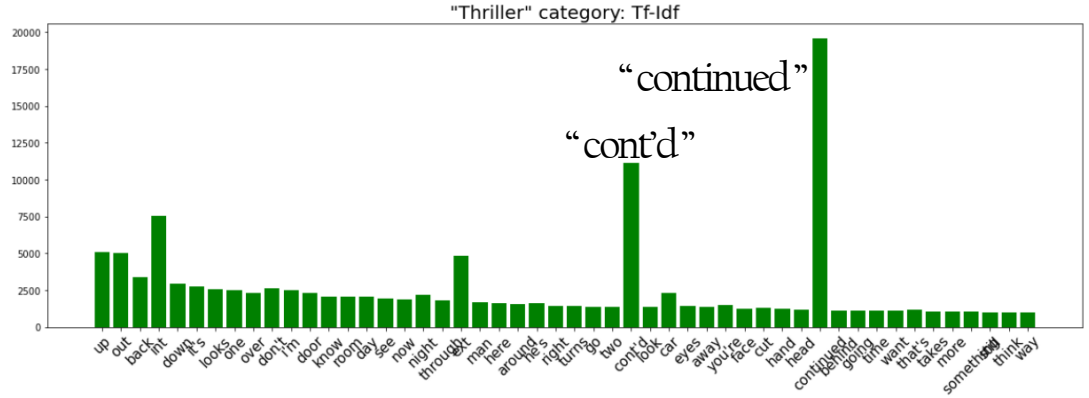
4. Result & Conclusion

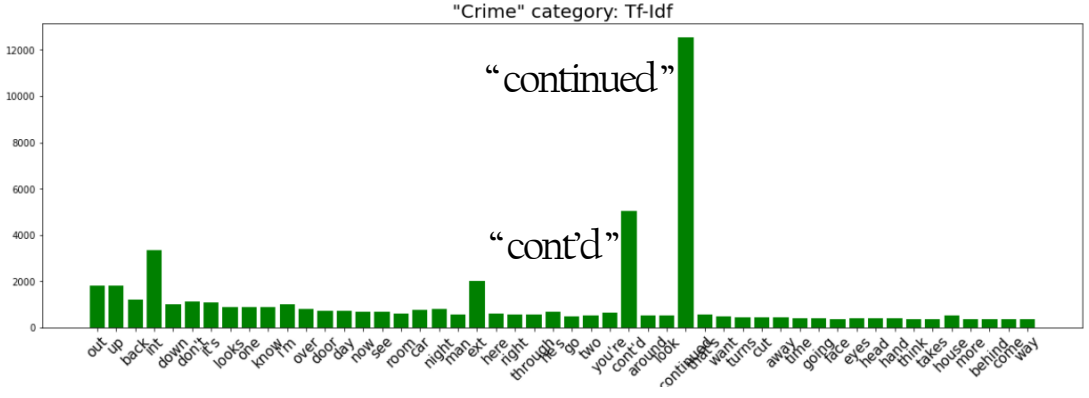
Assumption 2











4. Result & Conclusion

1) The word distribution will be similar within each group according to the category.

Answer for Assumption: In fact, the word distribution was similar in each group.

Problems: However, almost all words did not represent the characteristics of each category, such as prepositions or words that are 'excessively casual' like "day", "now", and so on.

Solutions: If we use methods such as increasing the number of stopwords and not including specific 'parts of speech' like preposition, the word distribution of each category would have shown better aspects.

2) If the frequency is high within each category, the tf-idf value will generally be high.

Answer for Assumption: In fact, Tf-idf was simply not proportional to the word frequency.

Problems & Solutions: Same as Assumption 1