Analyzing Movies Using Phrase Mining

Daniel Lee Huilai Miao Yuxuan Fan

March 7, 2021

Abstract

Movies are a rich source of human culture from which we can derive insight. Previous work addresses either a textual analysis of movie plots or the use of phrase mining for natural language processing, but not both. Here, we propose a novel analysis of movies by extracting key phrases from movie plot summaries using AutoPhrase, a phrase mining framework. Using these phrases, we analyze movies through 1) an exploratory data analysis that examines the progression of human culture over time, 2) the development and interpretation of a classification model that predicts movie genre, and 3) the development and interpretation of a clustering model that clusters movies. We see that this application of phrase mining to movie plots provides a unique and valuable insight into human culture while remaining accessible to a general audience, e.g., history and anthropology non-experts.

1 Introduction

Movies are a rich source of human culture from which we can derive insight through a comprehensive textual analysis of movie plot summaries.

Here, we propose an analysis of movies by extracting key phrases corresponding to discrete entities of human culture. Such an analysis can help us better understand popular topics, public attitudes, and the overall progression and themes of human culture throughout history.

This analysis is novel since we extract human culture from movies, an unconventional source, instead of relying on traditional sources, e.g., historical texts; we expect such a study to provide a unique perspective as a result. In addition, we expect such a study to be especially useful for history and anthropology non-experts, as we extract accessible key phrases that serve as relevant keywords for further research by the reader.

Previous work addresses either a textual analysis of movie plots or the use of phrase mining for natural language processing, but not both. Previous analyses of movies are limited as they tend to simply extract n-grams using raw term frequencies, which often leads to incomplete or spurious phrases, instead of using a dedicated and sophisticated phrase mining framework that is capable of extracting complete and coherent key phrases, such as the AutoPhrase framework [6].

Here, we use AutoPhrase to explore a novel approach by applying phrase mining to the analysis of movies.

2 Data

Our dataset comes from the CMU Movie Summary Corpus [2] and consists of movie plot summaries extracted from Wikipedia and movie metadata extracted from Freebase.

The dataset contains around 42,000 movies from 1893 to 2014 as seen in Figure 1, a sizable dataset for our study. Table 1 describes the variables of the processed dataset.

Although movies can come from different countries and may be in different languages, all of the movie plot summaries are in English (as the dataset is extracted from English Wikipedia).

The variable of focus here is `summary`, from which we extract key phrases to drive our analysis.

3 Methods

We first extract key phrases from `summary` using AutoPhrase [6], a phrase mining framework that extracts high-quality phrases from a given text. We use AutoPhrase for its ability to extract high-quality phrases more effectively than traditional, rudimentary, phrase mining techniques, while maintaining minimal human effort during training.

We first train AutoPhrase on all movie plot summaries in the dataset, then extract the key phrases from each individual summary and add these phrases as a variable in our dataset as seen in Table 2. Figure 2 gives an example of phrases extracted from a movie plot summary by AutoPhrase.

Figure 1: Number of movies in the dataset

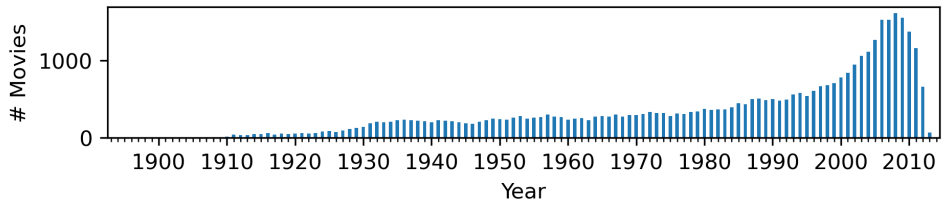


Table 1: Dataset variables

Variable	Description	Example
`name`	Movie name	*Star Wars Episode IV: A New Hope*
`date`	Movie release date	*1977-05-25*
`revenue`	Movie box office revenue (USD)	*775,398,007*
`runtime`	Movie runtime (minutes)	*122*
`languages`	Movie languages	*{English}*
`countries`	Movie countries	*{United States of America}*
`genres`	Movie genres	*{Action, Adventure, Coming-of-age, Family, Fantasy, Science Fiction, Space western}*
`summary`	Movie plot summary	*The film begins with an opening crawl explaining that the galaxy is in a state of civil war and that spies for the Rebel Alliance have . . .*

Table 2: Added variables

Variable	Description	Example
`phrases`	Phrases extracted from `summary`	*{alderaan, anakin skywalker, assault, aunt and uncle, bay, c-3po, chewbacca, civil war, commanding officer, darth vader, . . . }*

Figure 2: Example of phrases extracted by AutoPhrase

The film begins with an opening crawl explaining that the galaxy is in a state of civil war and that spies for the Rebel Alliance have stolen plans to the Galactic Empire's Death Star, a heavily armed and armored space station capable of annihilating an entire planet. Rebel leader Princess Leia is in possession of the plans, but her ship is captured by Imperial forces under the command of the evil lord Darth Vader. Before she is captured, Leia hides the plans in the memory of an astromech droid called R2-D2 , along with a holographic recording. The small droid flees to the surface of the desert planet Tatooine with fellow protocol droid C-3PO . The droids are quickly captured by . . .

Nearly all extracted phrases are indeed high-quality and encapsulate ideas, events, objects, and characters in the movie plot well.

Our analysis consists of three parts that build off of these extracted phrases:

1. An exploratory data analysis (EDA) that examines the progression of human culture over time.

2. The development and interpretation of a classification model that predicts movie genre.

3. The development and interpretation of a clustering model that clusters movies.

We expect the combination of these three methods to give us valuable insight into human culture.

3.1 EDA

We perform an EDA to discover how human culture and events have progressed over time. Here, we use statistical signals, i.e., tf-idf, to identify when certain phrases (corresponding to discrete entities of human culture and events) are popular and relevant.

We use tf-idf to measure a phrase's relevance in time where each phrase is a term and each period of time (e.g. a year or decade) is a document. We use sublinear term frequency (tf) scaling to reduce the significance of very common phrases (e.g. "film", "life") that are uninformative in our analysis.

Here, the tf-idf with sublinear tf scaling for a term t of a document d is given by

$$\text{tf-idf}(t,d) = (1 + \log \text{tf}(t,d)) \cdot \left(\log \frac{1+n}{1+\text{df}(t)} + 1 \right)$$

where n is the total number of documents and $\text{df}(t)$ is the document frequency of t.

Given a TF-IDF vector for each period of time, we then normalize each TF-IDF vector to have a Euclidean norm of 1.

Finally, we define "top" phrases as the phrases with the highest TF-IDF values for the corresponding period of time.

3.2 Classification

Classifying movie genres using movie plot summaries examines the power of phrase mining and explores its application in information extraction in movie sector. The results from the classification can also disclose some implicit features of movie genres through important phrases. To obtain ideal results, we construct a baseline model:

1) Using TF-IDF vectorizer for words embedding

2) Construct multiple logistic regressions inside OneVsRest Classifier for each genre to predict the label.

3) Evaluating the model use f-1 score. We also use some costumed indicator - percentage of correctly predicted label and percentage of movie that has at least one correctly predicted label

We further improve our model in following ways:

1) We try out different algorithms in OneVsRest Classifier including LinearSVC and Multi-linear Perceptron classifier.

2) We perform GridSearch for hyper-parameters tuning, changing different loss function and different regularization penalty for best performance.

We also construct the text feature in two ways

1) Use all non-stopwords in plot summaries

2) Use only extracted phrases of these summaries by Autophrase.

We can compare the results from these two approaches to explore the contribution of phrases to the model.

Finally, to better evaluate our model, we will interpret the coefficients of words of classifiers to see what words or phrases that our predictions mostly rely on. We then generate plots for the words/phrases rank to evaluate the model using our understanding of genre and see if the result can tell us anything more about the genre that we do not know before.

3.3 Clustering

We build a clustering pipeline on the movie plot summaries to discover relationships between movies in a marginally supervised perspective. As shown in Figure 3, our clustering pipeline contains three major components: (1) build and fine-tune a sentence transformer to acquire document embeddings, (2) pick representative sentences to condense plot summaries, and (3) cluster document embeddings.

3.3.1 Model

We purpose to estimate the document embeddings based on a pre-trained sentence transformer ([5]). Experiments in [1] show that the contextualized embeddings generated by pre-trained language models have the capability to preserve domain information. Compared with vanilla transformer networks, sentence transformers are further tailored to estimate contextualized sentence embeddings for semantic similarity tasks ([5]). Therefore, the sentence transformer is the strongest out-of-the-box model we can utilize.

Figure 3: The clustering pipeline

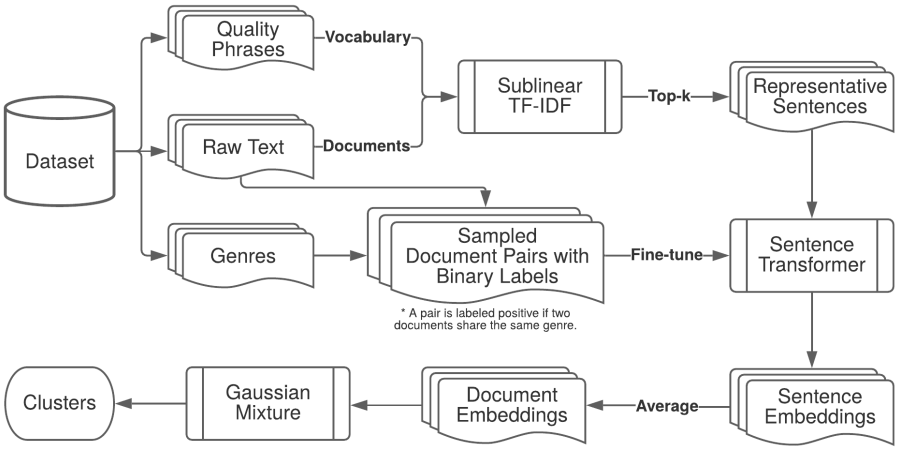
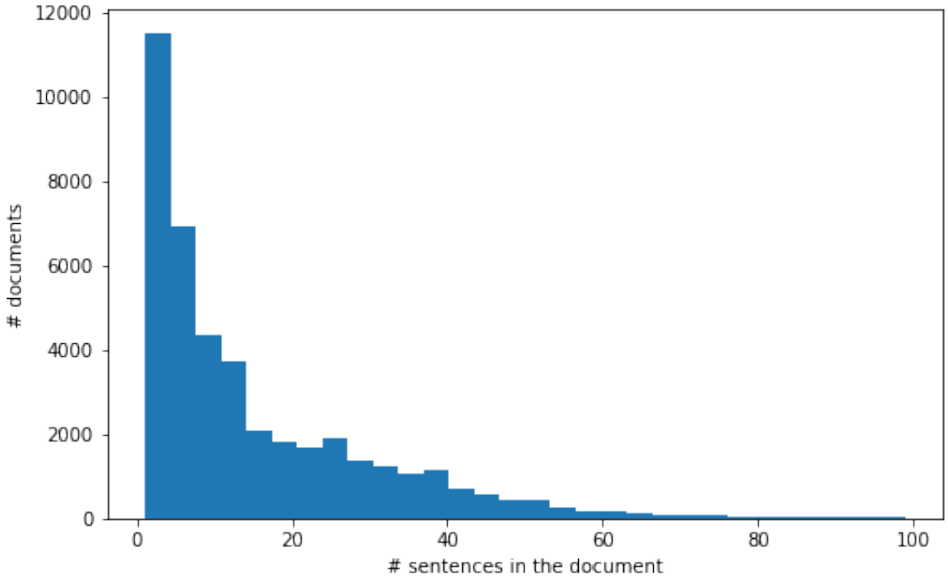


Figure 4: Number of sentences per document (plot summary)



Since the sentence transformer expects sentences as input, we split each summary into sentences, acquire an embedding for each sentence, and average the embeddings to form document embeddings.

3.3.2 Condense plot summaries

As shown in Figure 4, the distribution of the number of sentences in each movie summary is a long-tailed distribution, where 46.05% of summaries have more than 10 sentences. Since the transformer truncates the document to a fixed length, the embeddings acquired from the model cannot fully capture the semantics of long summaries. Therefore, it is necessary to condense the movie summaries while preserving most of the gist.

We purpose to rank the importance of the sentences and choose sentences with top-k importance to represent the entire summary. Here, we define the importance as the average sublinear TF-IDF score, where the dictionary of terms is only the quality phrases. Under this notion of importance, sentences with no quality phrases have an importance score of 0, and sentences with more uncommon quality phrases have a higher importance score. If the document has fewer than k sentences, we keep all sentences. If the document contains no quality phrases, we sample k sentences randomly. To acquire the document embeddings, we only use representative sentences instead of all sentences.

3.3.3 Fine-tuning

Fine-tuning a pre-trained language model with a specific downstream task is a common practice to enhance the model performance. In our clustering pipeline, we use semantic similarity as the downstream task, since we expect documents with similar meanings to have closer embeddings, and vice versa. For two pairs of sentences, we determine the similarity by finding whether the genres associated with these sentences have overlap. To generate the training dataset, we iteratively sample sentence pairs until the training dataset contains at least n pairs, where half of the pairs have positive labels and the other half have negative labels. In each iteration, the algorithm samples two documents from the dataset, generates the binary label based on their genres, and selects k sentence pairs from two documents.

We use the Contrastive Loss ([3]) as the fine-tune target, since it aims to decrease the distance between sentence embeddings of similar sentence pairs and increase the distance between sentence embeddings of dissimilar sentence pairs.

3.3.4 Cluster embeddings

Following the practices in [1], we first reduce the dimensions of embeddings to 50 with PCA, then cluster the dimension-reduced embeddings using Gaussian Mixture Model (GMM). [1] explained that the mixture model is more suitable since we can view each document embedding as from a mixture of different domain distributions, and applying PCA to the embeddings accelerates the training process while marginally damaging the performance.

4 Results

4.1 EDA

Figure 5 shows top phrases by decade, where top phrases are defined earlier as the phrases with the highest tf-idf values for the corresponding decade.

To provide context, we annotate each phrase with the name and year of the highest-grossing movie out of all movies that contain that phrase in that decade. In the case of a tie in gross revenue, we pick the movie with the earlier release date.

We now examine the results by decade:

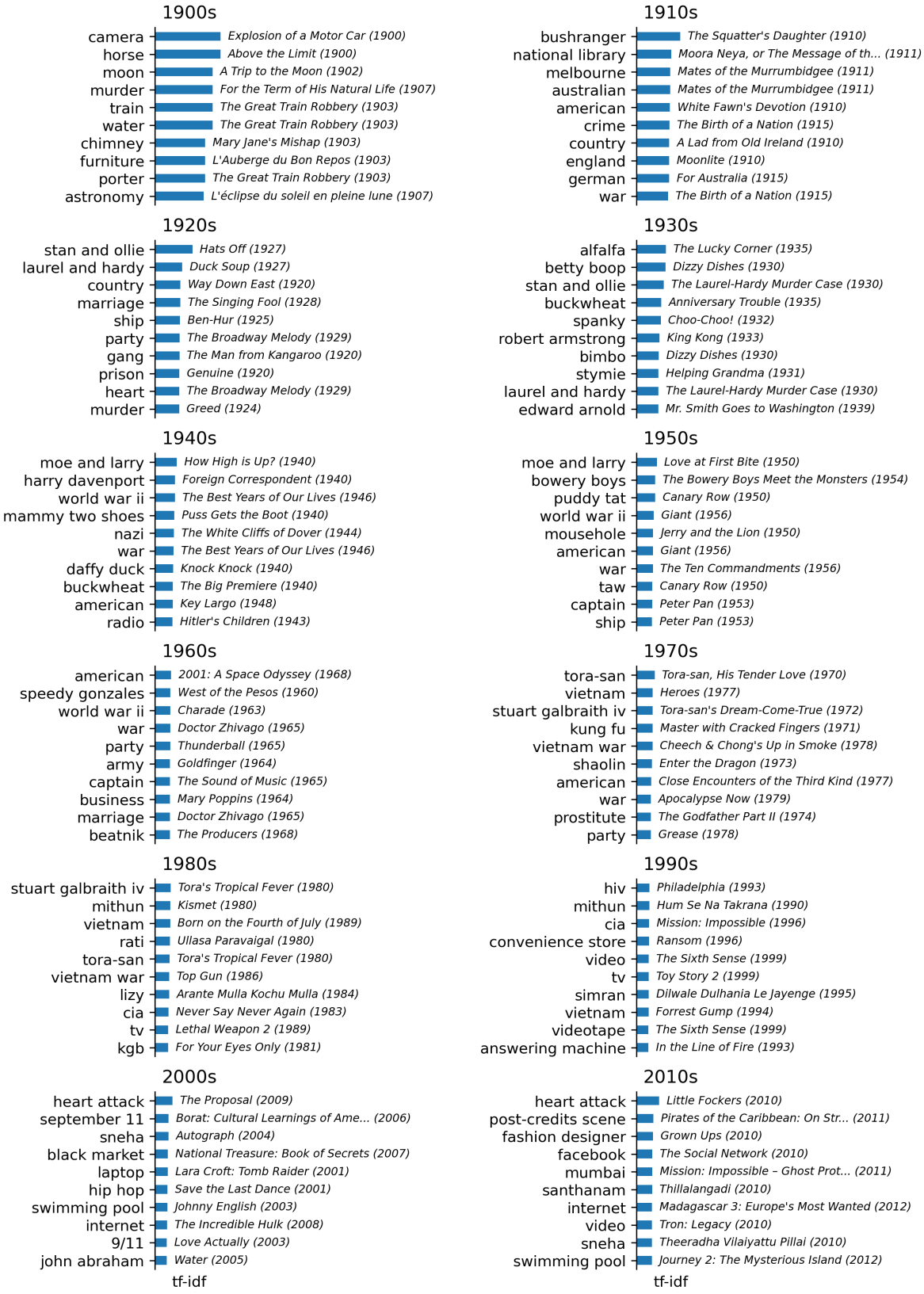
- *1900s*: It is important to note that the earliest commercial movie screening occurred in 1895 and thus movies were still a relatively new phenomenon in the early 1900s. Top phrases in this decade do not appear to be very salient but do reference objects that may be associated with the early 20th century such as "horse" and "train".

- *1910s*: Many of the top phrases and their corresponding top movies in this decade relate to Australia, which is reasonable since the Commonwealth of Australia was established in 1901. This includes the phrases "bushranger" (an outlaw living in the Australian bush), "Melbourne", and "Australian" and the movies *The Squatter's Daughter*, *Moora Neya, or The Message of the Spear*, *Mates from the Murrumbidgee*, *Moonlite*, and *For Australia.*

- *1920s*: The most salient top phrases in this decade, "Stan and Ollie" and "Laurel and Hardy", refer to the internationally famous comedy duo Stan Laurel and Oliver Hardy, whose act was active from the 1920s to 1950s.

- *1930s*: We see the appearance of movie characters Alfafa, Buckwheat, Spanky, and Stymie from the *Our Gang* comedy series, Betty Boop and Bimbo from the *Talkartoon* and *Betty*

Figure 5: Top phrases by decade, with corresponding highest-grossing movie



Boop cartoon series, actors Robert Armstrong in *King Kong* and Edward Arnold in *Mr. Smith Goes to Washington*, and the reappearance of Stan and Ollie.

- *1940s*: Many of the top phrases and their corresponding top movies in this decade relate to World War II, fought from 1939 to 1945, including the phrases "World War II" and "Nazi" and the movies *Foreign Correspondent*, *The Best Years of Our Lives*, *The White Cliffs of Dover*, and *Hitler's Children*. We see the appearance of movie characters Mammy Two Shoes from the the *Tom and Jerry* cartoon series, Daffy Duck from the *Looney Tunes* and *Merrie Melodies* cartoon series, actors Moe and Larry, the two mainstay members of The Three Stooges, a famous comedy team active from the 1920s to 1970s (the third stooge changed multiple times throughout the team's history) and Harry Davenport in *Foreign Correspondent*, and the reappearance of Buckwheat.

- *1950s*: We continue to see phrases and movies relating to World War II, including the movie *Giant*. We see the appearance of movie characters The Bowery Boys and the reappearance of Moe and Larry. We also see the phrase "Puddy Tat" from Tweety's catchphrase "I Taut I Taw a Puddy-Tat" from the Sylvester and Tweety cartoons and the phrase "mousehole" from the *Tom and Jerry* cartoon series.

- *1960s*: We continue to see phrases and movies relating to World War II. We see the use of the phrase "beatnik", a media stereotype prevalent from the 1940s to 1960s associated with the nonconformist Beat Generation literary movement in the post-war era, whose elements were later incorporated into the hippie movement and other counterculture movements. We see the appearance of movie characters Speedy Gonzales from the *Looney Tunes* and *Merrie Melodies* cartoon series.

- *1970s*: Many of the top phrases and their corresponding top movies in this decade relate to the Vietnam War, fought from 1955 to 1975, including the phrases "Vietnam" and "Vietnam War". Other top phrases and their corresponding top movies relate to martial arts, including the phrases "kung fu" and "Shaolin". During this time, martial arts movies rose in popularity, such as those featuring Bruce Lee, which helped lead to an increase in Asian and Asian American representation in cinema. We see the appearance of movie characters Tora-san from the *Otoko wa Tsurai yo* Japanese film series and film historian and critic Stuart Galbraith IV.

- *1980s*: We continue to see phrases and movies relating to the Vietnam War. We see the appearance of America's CIA, founded in 1947, and the Soviet Union's KGB, formed in 1954, two opposing security/intelligence agencies whose appearance in movies may have been popularized by the Cold War during this time period. We see the appearance of Indian cinema in the form of actors Mithun Chakraborty in *Kismet*, considered to be one of the foundational films of Bollywood, Lizy in *Arante Mulla Kochu Mulla*, and Rati Agnihotri in *Ullasa Paravaigal*. We see the reappearance of Tora-San and Stuart Galbraith IV.

- *1990s*: The top phrase in this decade is "HIV", which was first clinically observed in 1981, triggering much of the early HIV/AIDS research in the 1980s. This development may have led to HIV/AIDS being recognized by popular culture in the following years; in fact, the top movie corresponding to "HIV", *Philadelphia*, was one of the first mainstream Hollywood movies to mention HIV/AIDS. We continue to see phrases and movies relating to the CIA and Vietnam War. We see technologies that may have been more readily accessible to consumers in the late 20th century, such as "video", "tv", "videotape", and "answering machine". We see the appearance of movie character Simran Singh in another Bollywood movie, *Dilwale Dulhania Le Jayenge* and the reappearance of Mithun.

- *2000s*: The top phrase in this decade is "heart attack". It is unclear why, but as a continuously leading cause of death since the mid-20th century, cardiovascular disease may have been especially prevalent in popular culture in this time period. We see top phrases relating to the September 11 attacks, namely "September 11" and "9/11". We continue to see technologies that are relevant to consumers in this time period, namely "laptop" and "internet". We see the appearance of "hip hop", which became the top-selling music genre by 1999 and continued to become increasingly popular throughout the 2000s. We see the appearance of actors Sneha in *Autograph* and John Abraham in *Water*.

- *2010s*: We come to the most recent decade in the dataset. "Heart attack" continues to be the top phrase in this decade. We continue to see

technologies that are relevant to consumers in this time period, namely "Facebook" and "internet". Indian cinema continues its presence with the appearance of actor N. Santhanam in *Thillalangadi* and the reappearance of Sneha in *Theeradha Vilaiyattu Pillai*.

We can see that examining and researching top phrases over time gives us valuable insight into human culture and its public attitudes, events, people, ideas, etc. Phrase mining and the use of tf-idf automatically identifies relevant keywords in a way that would be difficult to do manually or by using traditional text mining methods.

Figure 6 in Appendix goes into much finer detail by showing top phrases by year instead of decade and is also available as an animation. There, we observe a similar trend of top phrases across time but also observe phrases that are peculiar to each year.

4.2 Classification

Our baseline had a result F-1 score 0.332 using whole summaries as feature and 0.29 using phrases. Our final model had a result F-1 score of 0.407 using summaries and 0.364 using phrases. l0.9

	F-1 score	label percentage	movie percentage
summary text	AF		
phrases	AX		

We extract the coefficients of words from the classifiers to understand what words matter most for each genre in our classification. It helps to evaluate the model as well as deliver some insights on genre prediction. The plot shows the top 20 words that has the largest coefficients, thus the highest significance to the that genre. Due to lack of ground truth to compare with, we will put forward some interesting findings below from our first 4 most common genres' plots generated by our final model:

- *Drama*: Using summary text, we found some female names like Helene, Carmen, Johann in the plots. Our model seems to place some importance in the Female name when predicting Drama movies. While using phrases, we found more animals in the top phrases, like goa, cattle, guinea pig, etc. There are more negative words/phrases as well, like exorcism, flatulence, local-mob, domestic violence etc.

- *Comedy*: Using summary text, interestingly, we found some male names in the top words, like Elmo, Davey, Jones, Doug etc. Other words mostly human behaviors or status like hurry, ashamed, dig, chop, etc. Using phrase, we find the phrases and words are more region related, like Ethiopian, Baltimore, British Rag etc. There are also lots of hostile words like hostage, homophobia, ensuing battle, blockade, destructive, etc. One hypothesis is that lots of comedies use contradictions to provoke funny scenes.

- *Romance*: Using summary text, we still see lots of people names. There are also words describing relation status like deserted and attracted. Using phrases, there appear to be some locations and community like Havana, Houston, Rojo, death squad, and firing squad. The results here seem quite deviated from what we expected.

- *Thriller*: Using summary text, there are words with uncanny meanings like alien, immortality, bury, deadly. Using phrases, there are words related to criminal scenes like drug cartel, captured and imprisoned.

4.3 Clustering

To examine our clustering pipeline, we cluster all vectors into 200 clusters and perform analysis on the output. We choose this number of clusters as the target since each cluster will roughly contain 200 movies, which is a reasonable size to derive fine-grained semantic units.

Since the original dataset contains more than 300 genres and each document contains multiple genres, assigning the cluster label with the most common label within the cluster ([1]) is not feasible. Instead, we choose to determine the semantics of the cluster using multiple heuristics, including quality phrase distributions and the overlaps with existing genres.

Quality phrases. To preserve the generality, we remove all phrases that occur in less than 100 documents. Then, for each phrase in each cluster, we calculate the document frequency within the cluster relative to its document frequency in the entire dataset. In other words, when a phrase occurs more frequently in this cluster than other clusters, we consider this phrase as a representative of this cluster. We rank the relative document frequencies of quality phrases in each cluster and pick the top 5 phrases to represent the cluster.

Table 3 shows some examples of the representative quality phrases. From the examples, we can observe that many clusters can capture meanings that can be easily identified. We can infer that example 1 contains movies about special forces, example 2 is about martial arts in Asia, and example 3 is about

Table 3: Examples of representative quality phrases in clusters

E.g.	Quality Phrases
1	cia, sniper, helicopter, swat, laser, ...
2	kung fu, martial arts, monks, shanghai, thailand, ...
3	championship, basketball, coach, academic, baseball, ...
4	publisher, tokyo, province, economy, fever, ...
5	santa claus, fairy, frog, daffy, porky, ...

Table 4: Examples of representative genres in clusters

E.g.	Genres
1	Slapstick, Sex comedy, Musical comedy, Comedy of manners, Comedy of Errors, ...
2	Courtroom Drama, Docudrama, Erotic Drama, Melodrama, Erotic thriller, ...
3	Stop motion, Children's Fantasy, Computer Animation, Animation, Family-Oriented Adventure, ...
4	Musical Drama, Ensemble, Experimental, Biography, Gay, ...
5	Comedy of manners, Domestic Comedy, Anime, Sports, Teen, ...

sports. However, we cannot identify the specific topic for some clusters like examples 4 and 5.

Genres. We repeat the same process as examining the quality phrase memberships to examine the overlaps with genres. Table 4 shows some examples of the genre distributions. We can observe that many clusters (like example 1, 2, 3) gather the movies with similar genres together with only a few hints about similarity during the fine-tuning process. However, for clusters like examples 4 and 5, we cannot identify an obvious topic.

We can see that the document embeddings can capture the semantics of the summaries, and feeding them to a clustering pipeline give us insights about the relationships between movies. By defining heuristics and summarize from each cluster, we can observe more connections that are hidden from the existing human crafted features.

5 Discussion

In this paper, we apply phrase mining to movie plots, a novel approach for the textual analysis of movies, for a unique insight into human culture.

In the EDA investigation, we use statistical signals to select the most relevant phrases describing each time period. In the classification investigation, ... In the clustering investigation, we utilize quality phrases and pre-trained language models to discover the relationship between movies in a nearly unsupervised approach.

In each of these investigations, we demonstrate phrase mining's effectiveness in producing valuable insight into human culture.

For future work, we may consider omitting movie character and/or actor names from the top phrases, as movie characters and actors, unless historically significant, are not central to our analysis of human culture.

References

[1] Roee Aharoni and Yoav Goldberg. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online, July 2020. Association for Computational Linguistics.

[2] David Bamman, Brendan T. O'Connor, and Noah A. Smith. Learning latent personas of film characters. In *ACL*, 2013.

[3] Raia Hadsell, Sumit Chopra, and Yann Lecun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006.

[4] Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han. Mining quality phrases from massive text corpora. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, page 1729–1744, New York, NY, USA, 2015. Association for Computing Machinery.

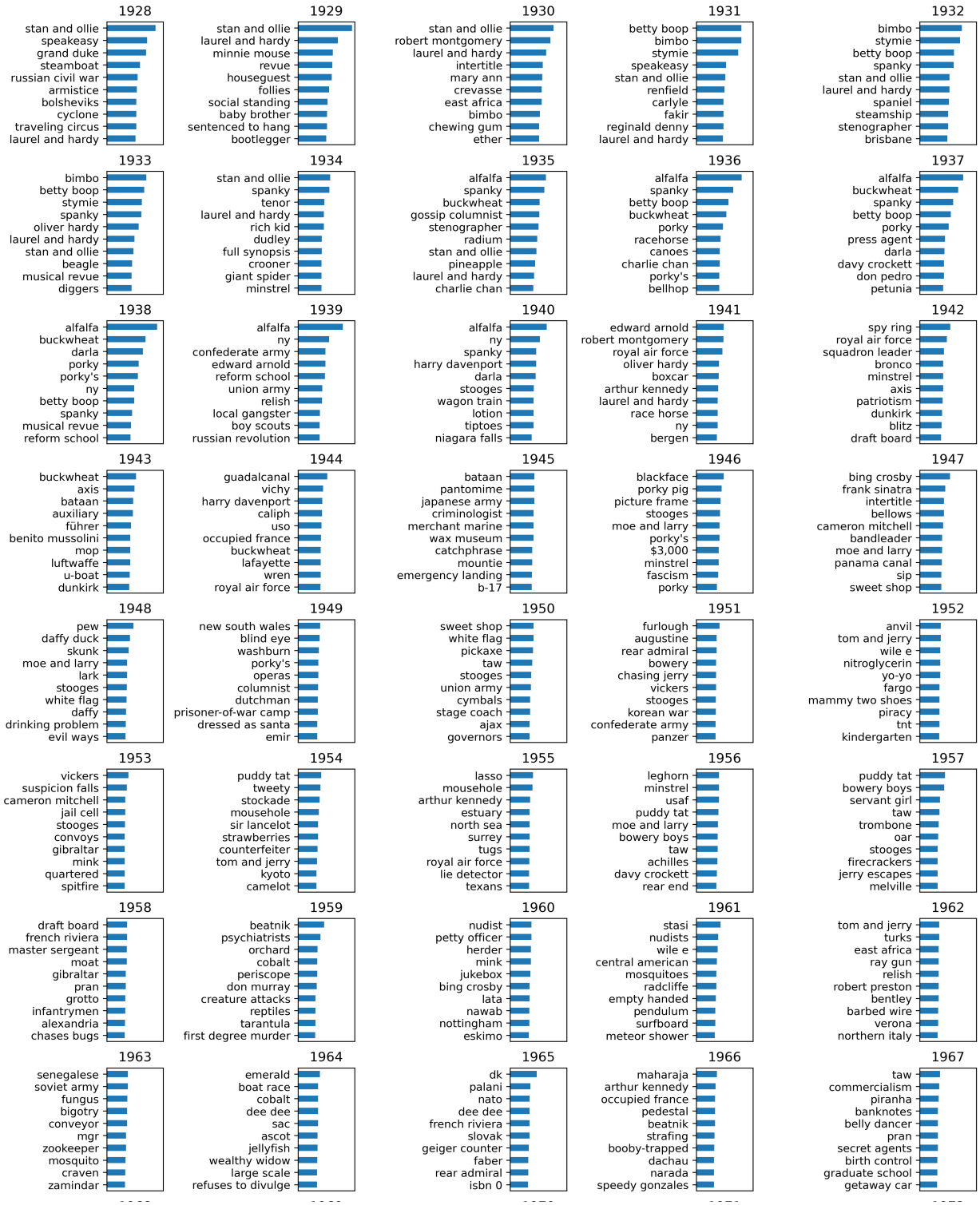
[5] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural*

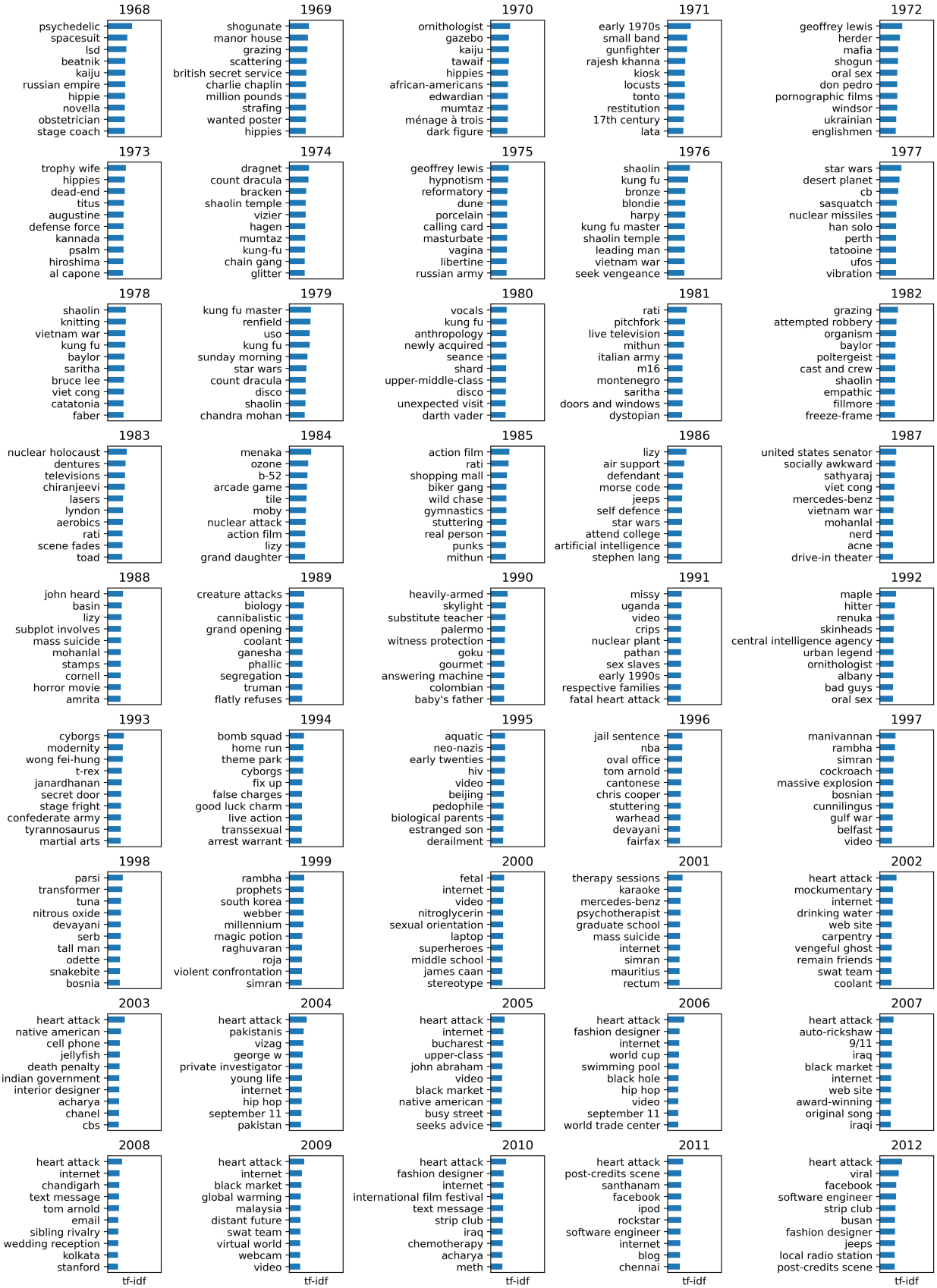
Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[6] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, and Jiawei Han. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30:1825–1837, 2018.

6 Appendix

Figure 6: Top phrases by year




1968
psychedelic
spacesuit
lsd
beatnik
kaiju
russian empire
hippie
novella
obstetrician
stage coach
1969
shogunate
manor house
grazing
scattering
british secret service
charlie chaplin
million pounds
strafing
wanted poster
hippies
1970
ornithologist
gazebo
kaiju
tawaif
hippies
african-americans
edwardian
mumtaz
ménage à trois
dark figure
1971
early 1970s
small band
gunfighter
rajesh khanna
kiosk
locusts
tonto
restitution
17th century
lata
1972
geoffrey lewis
herder
mafia
shogun
oral sex
don pedro
pornographic films
windsor
ukrainian
englishmen
1973
trophy wife
hippies
dead-end
titus
augustine
defense force
kannada
psalm
hiroshima
al capone
1974
dragnet
count dracula
bracken
shaolin temple
vizier
hagen
mumtaz
kung-fu
chain gang
glitter
1975
geoffrey lewis
hypnotism
reformatory
dune
porcelain
calling card
masturbate
vagina
libertine
russian army
1976
shaolin
kung fu
bronze
blondie
harpy
kung fu master
shaolin temple
leading man
vietnam war
seek vengeance
1977
star wars
desert planet
cb
sasquatch
nuclear missiles
han solo
perth
tatooine
ufos
vibration
1978
shaolin
knitting
vietnam war
kung fu
baylor
saritha
bruce lee
viet cong
catatonia
faber
1979
kung fu master
renfield
uso
kung fu
sunday morning
star wars
count dracula
disco
shaolin
chandra mohan
1980
vocals
kung fu
anthropology
newly acquired
seance
shard
upper-middle-class
disco
unexpected visit
darth vader
1981
rati
pitchfork
live television
mithun
italian army
m16
montenegro
saritha
doors and windows
dystopian
1982
grazing
attempted robbery
organism
baylor
poltergeist
cast and crew
shaolin
empathic
fillmore
freeze-frame
1983
nuclear holocaust
dentures
televisions
chiranjeevi
lasers
lyndon
aerobics
rati
scene fades
toad
1984
menaka
ozone
b-52
arcade game
tile
moby
nuclear attack
action film
lizy
grand daughter
1985
action film
rati
shopping mall
biker gang
wild chase
gymnastics
stuttering
real person
punks
mithun
1986
lizy
air support
defendant
morse code
jeeps
self defence
star wars
attend college
artificial intelligence
stephen lang
1987
united states senator
socially awkward
sathyaraj
viet cong
mercedes-benz
vietnam war
mohanlal
nerd
acne
drive-in theater
1988
john heard
basin
lizy
subplot involves
mass suicide
mohanlal
stamps
cornell
horror movie
amrita
1989
creature attacks
biology
cannibalistic
grand opening
coolant
ganesha
phallic
segregation
truman
flatly refuses
1990
heavily-armed
skylight
substitute teacher
palermo
witness protection
goku
gourmet
answering machine
colombian
baby's father
1991
missy
uganda
video
crips
nuclear plant
pathan
sex slaves
early 1990s
respective families
fatal heart attack
1992
maple
hitter
renuka
skinheads
central intelligence agency
urban legend
ornithologist
albany
bad guys
oral sex
1993
cyborgs
modernity
wong fei-hung
t-rex
janardhanan
secret door
stage fright
confederate army
tyrannosaurus
martial arts
1994
bomb squad
home run
theme park
cyborgs
fix up
false charges
good luck charm
live action
transsexual
arrest warrant
1995
aquatic
neo-nazis
early twenties
hiv
video
beijing
pedophile
biological parents
estranged son
derailment
1996
jail sentence
nba
oval office
tom arnold
cantonese
chris cooper
stuttering
warhead
devayani
fairfax
1997
manivannan
rambha
simran
cockroach
massive explosion
bosnian
cunnilingus
gulf war
belfast
video
1998
parsi
transformer
tuna
nitrous oxide
devayani
serb
tall man
odette
snakebite
bosnia
1999
rambha
prophets
south korea
webber
millennium
magic potion
raghuvaran
roja
violent confrontation
simran
2000
fetal
internet
video
nitroglycerin
sexual orientation
laptop
superheroes
middle school
james caan
stereotype
2001
therapy sessions
karaoke
mercedes-benz
psychotherapist
graduate school
mass suicide
internet
simran
mauritius
rectum
2002
heart attack
mockumentary
internet
drinking water
web site
carpentry
vengeful ghost
remain friends
swat team
coolant
2003
heart attack
native american
cell phone
jellyfish
death penalty
indian government
interior designer
acharya
chanel
cbs
2004
heart attack
pakistanis
vizag
george w
private investigator
young life
internet
hip hop
september 11
pakistan
2005
heart attack
internet
bucharest
upper-class
john abraham
video
black market
native american
busy street
seeks advice
2006
heart attack
fashion designer
internet
world cup
swimming pool
black hole
hip hop
video
september 11
world trade center
2007
heart attack
auto-rickshaw
9/11
iraq
black market
internet
web site
award-winning
original song
iraqi
2008
heart attack
internet
chandigarh
text message
tom arnold
email
sibling rivalry
wedding reception
kolkata
stanford
2009
heart attack
internet
black market
global warming
malaysia
distant future
swat team
virtual world
webcam
video
2010
heart attack
fashion designer
internet
international film festival
text message
strip club
iraq
chemotherapy
acharya
meth
2011
heart attack
post-credits scene
santhanam
facebook
ipod
rockstar
software engineer
internet
blog
chennai
2012
heart attack
viral
facebook
software engineer
strip club
busan
fashion designer
jeeps
local radio station
post-credits scene
tf-idf
tf-idf
tf-idf
tf-idf
tf-idf