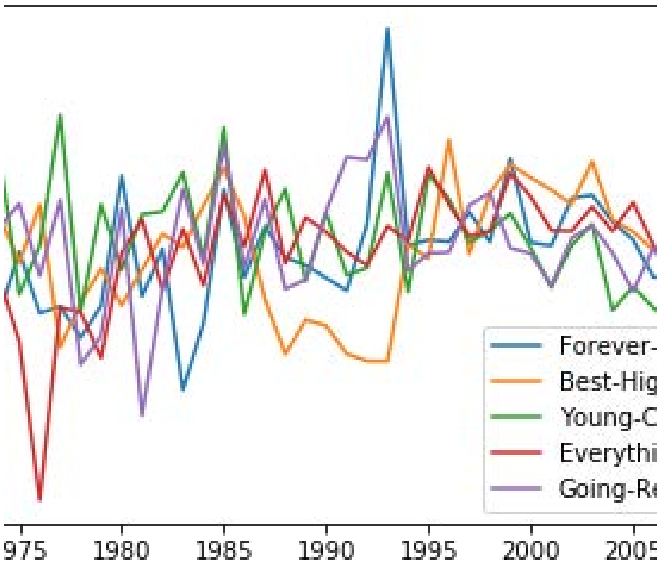
Text Intro

School of Information Studies
Syracuse University

Finding Meaning in Text Data

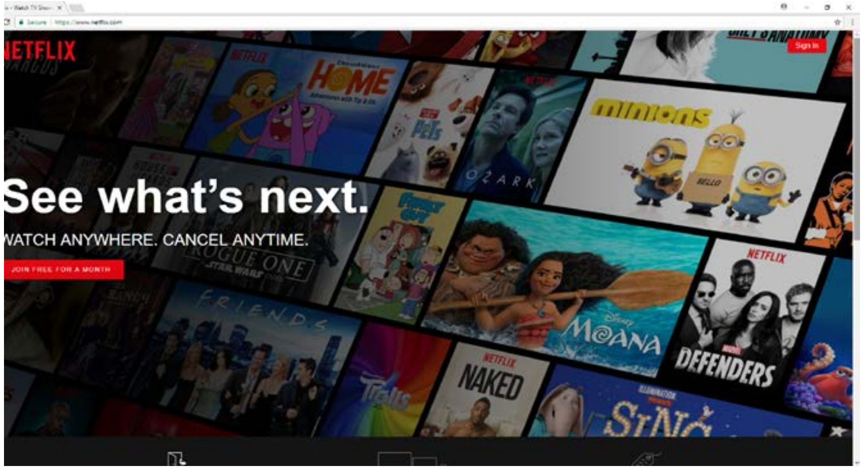
- **Obtain**
- **Scrub**
- **Explore**
- Model
- iNterpret

• One of the Funniest Police Shows Ever	(1974)
• Mosaiken und Fresken als Zeugen	(1975)
• Let all the poisons that lurk in the mud	(1976)
...	
• 12 days that shook Chile	(2011)
• The Movie Unrated and Uncensored	(2012)
• Pretty on the outside, ugly crying on the inside	(2013)



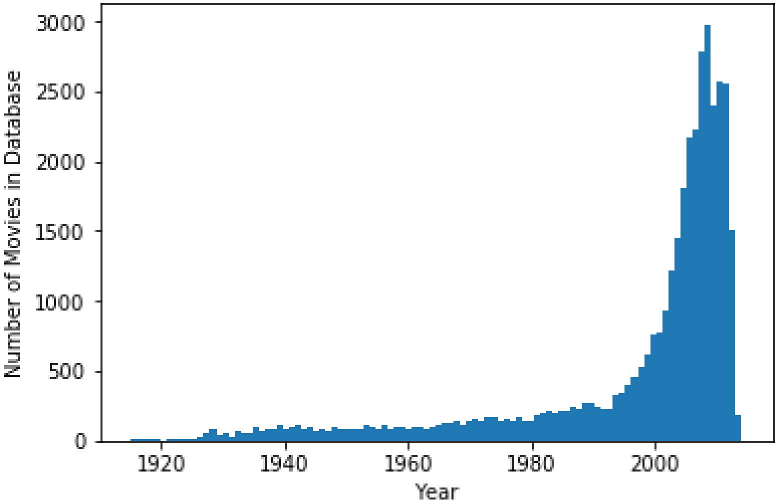
Our Challenge This Week?

What Are the Movie Trends?

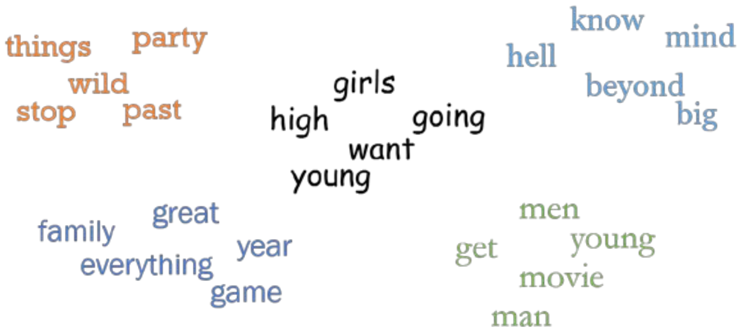


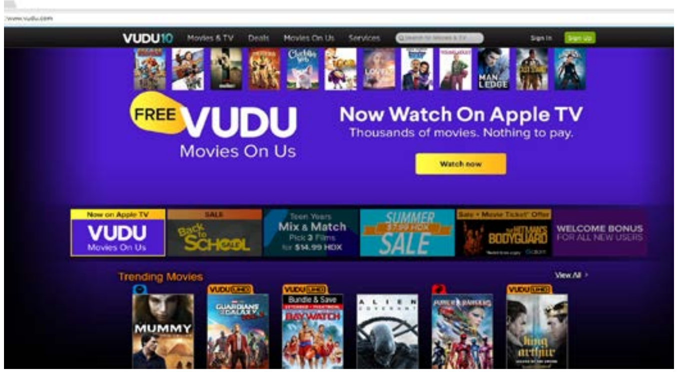
See what's next.

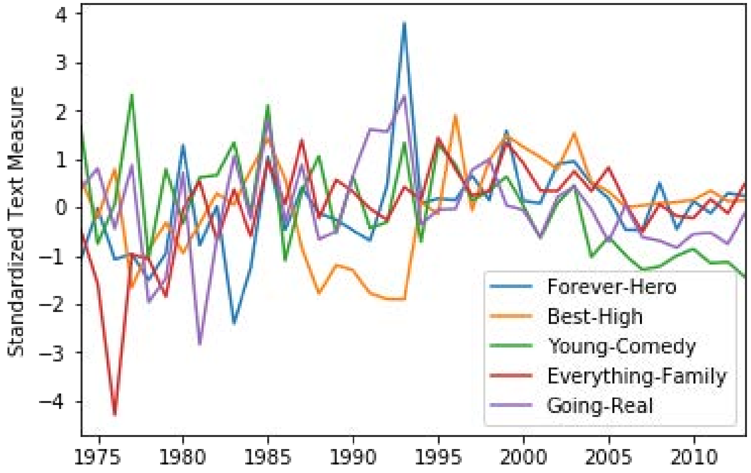
WATCH ANYWHERE. CANCEL ANYTIME.



Can We Use Text Analytics to Detect Trends?



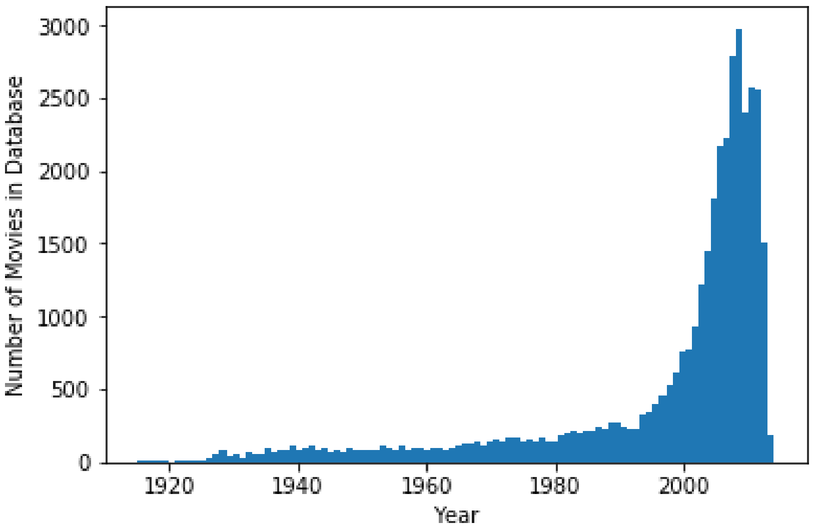




Data Review

School of Information Studies
Syracuse University

Text Data



actors
black
early
enjoy
films
full
get
good
great
guys
humor
just
last
later
lead
like
little
loved
men
minutes

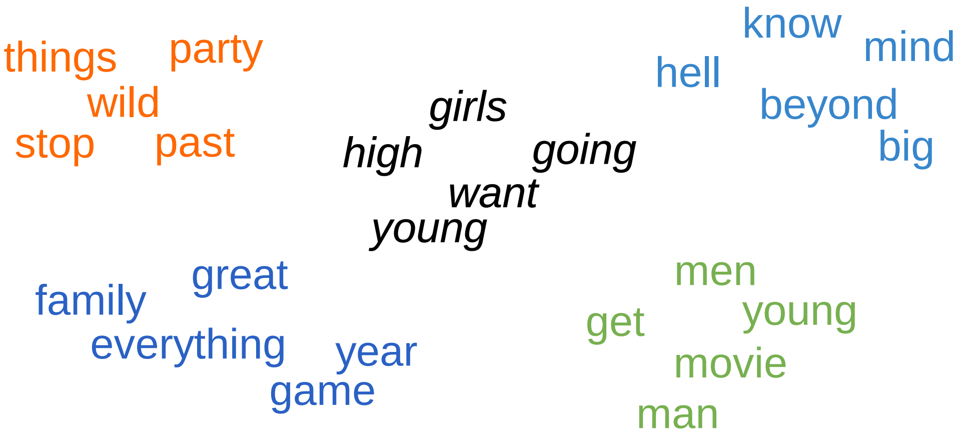
often
one
part
perhaps
say
supposed
thought
three
two
along
american
another
best
comedy
definitely
despite
direction
director
end
entertaining

far
film
funny
goes
however
long
love
make
picture
plot
quite
rather
script
still
story
stupid
war
way
well
worth

years
acting
action
actually
back
can
characters
done
enjoyed
especially
every
excellent
face
feel
finally
first
gives
got
heard
history

idea
kids
line
look
many
movie
never
new
night
now
original
people
plays
read
real
right
role
scene
scenes
screen

Bag of Words



things party wild stop past

girls high going want young

hell know mind beyond big

family great everything year game

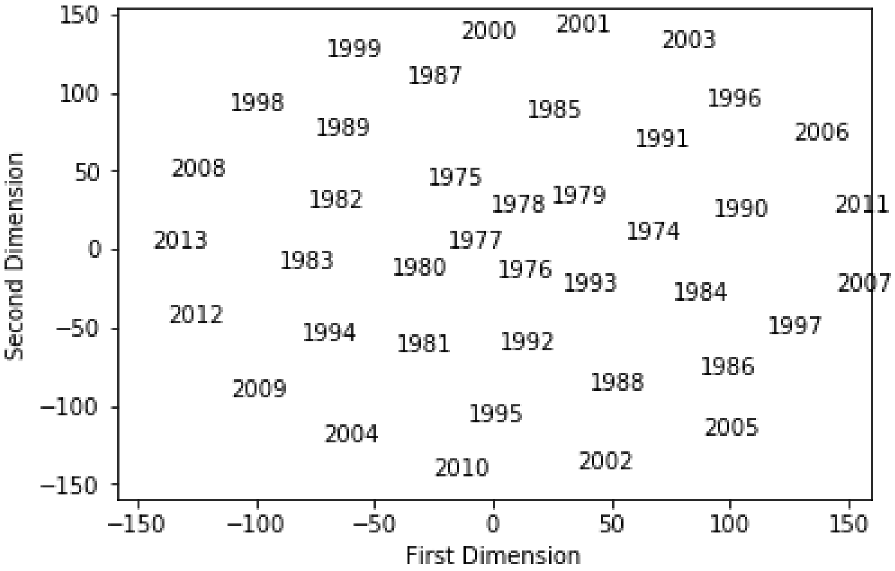
get men young movie man

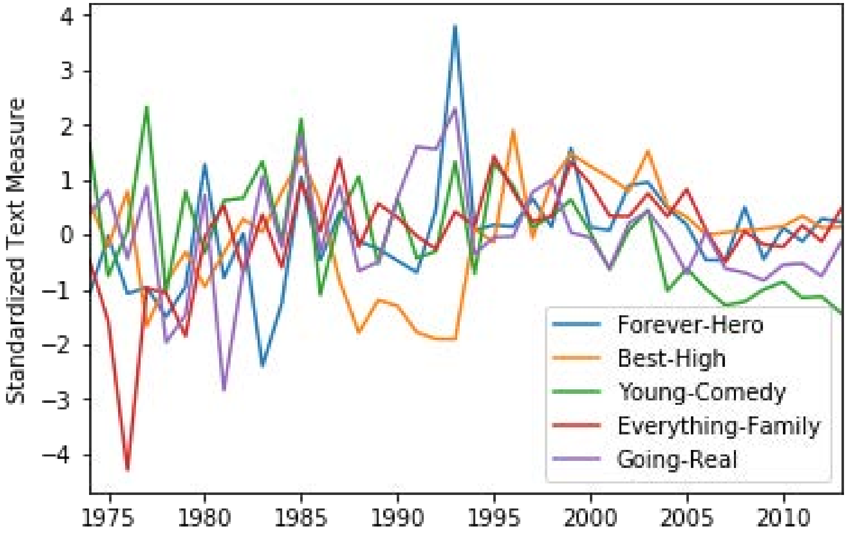
- One of the Funniest Police Shows Ever (1974)
- Mosaiken und Fresken als Zeugen (1975)
- Let all the poisons that lurk in the mud (1976)

...

- 12 Days That Shook Chile (2011)
- The Movie Unrated and Uncensored (2012)
- Pretty on the Outside, Ugly Crying on the Inside (2013)

Bag of Words (cont.)





Sentiment Analysis

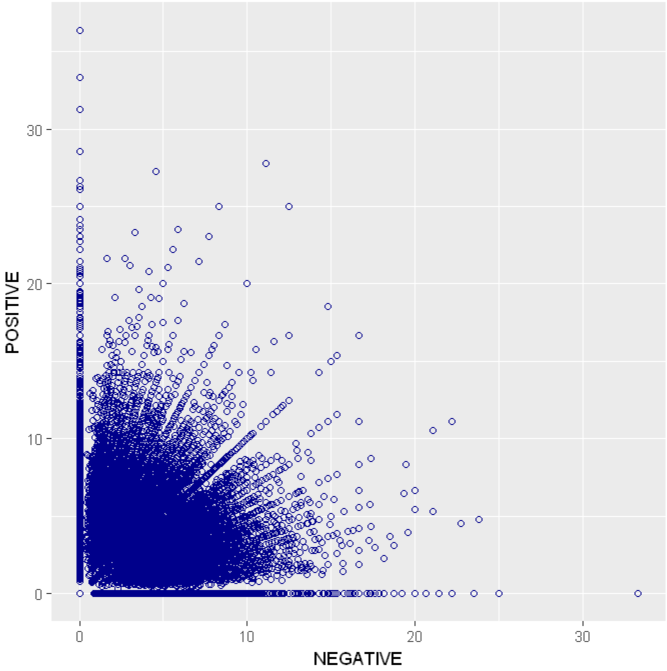
actors
black
early
enjoy
films
full
get
good
great
guys
humor
just
last
later
lead
like
little
loved
men
minutes

often
one
part
perhaps
say
supposed
thought
three
two
along
american
another
best
comedy
definitely
despite
direction
director
end
entertaining

far
film
funny
goes
however
long
love
make
picture
plot
quite
rather
script
still
story
stupid
war
way
well
worth

years
acting
action
actually
back
can
characters
done
enjoyed
especially
every
excellent
face
feel
finally
first
gives
got
heard
history

idea
kids
line
look
many
movie
never
new
night
now
original
people
plays
read
real
right
role
scene
scenes
screen

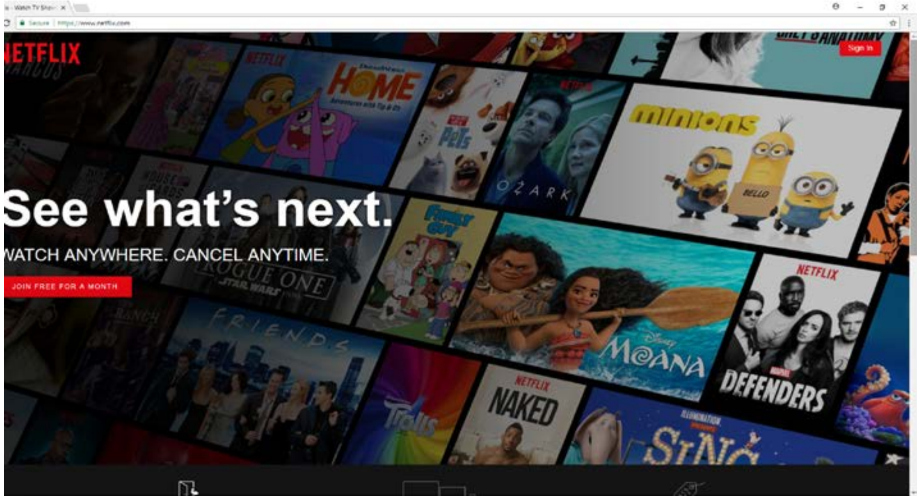


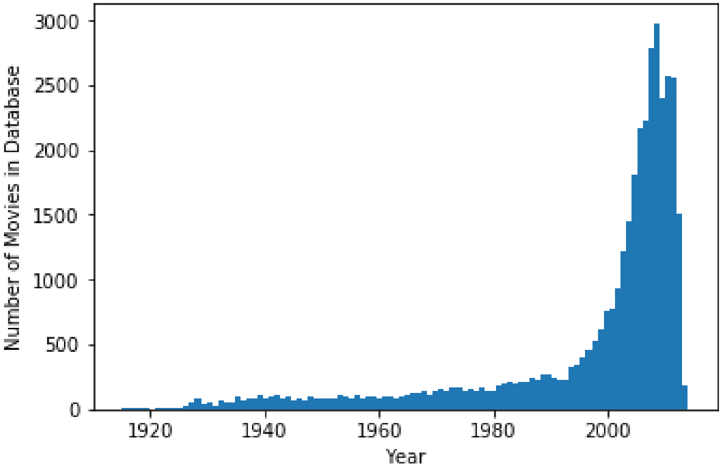
Recommendation

School of Information Studies
Syracuse University

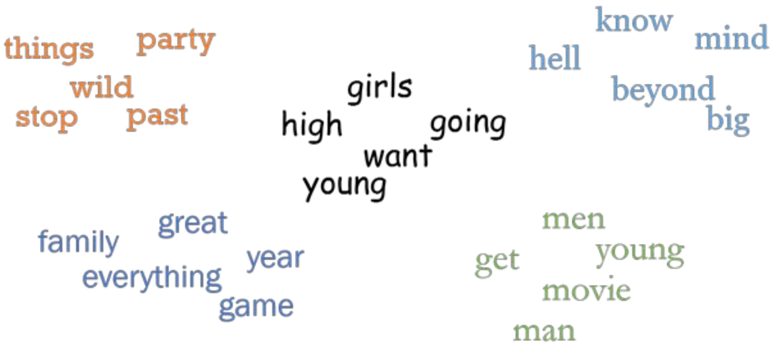
Our Challenge This Week?

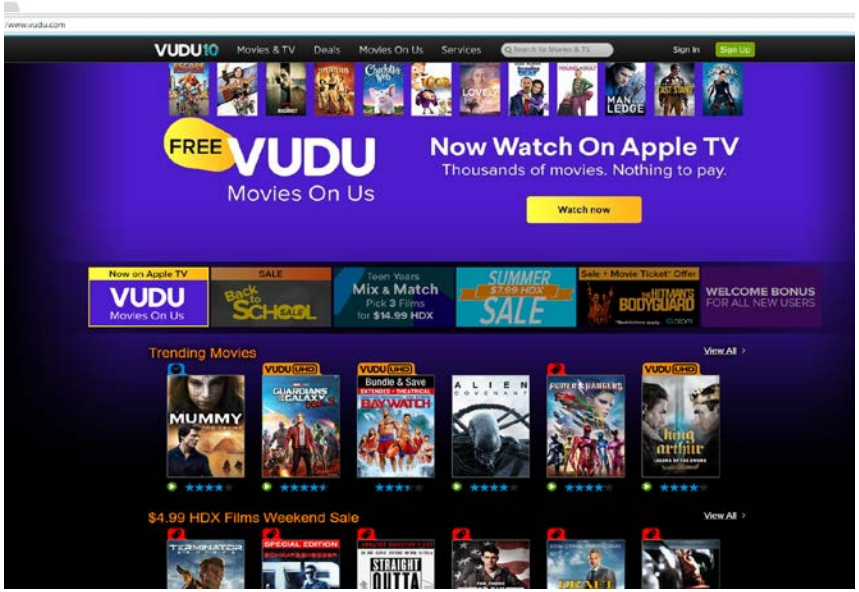
How Might We Discover Them?



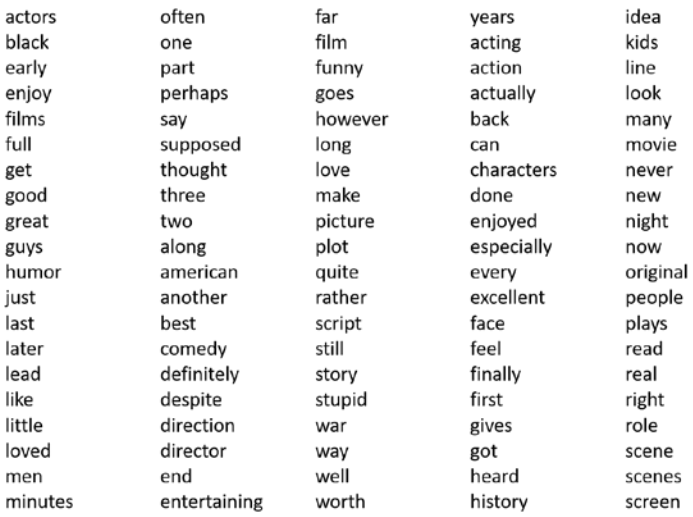


How Might We Discover Them? (cont.)

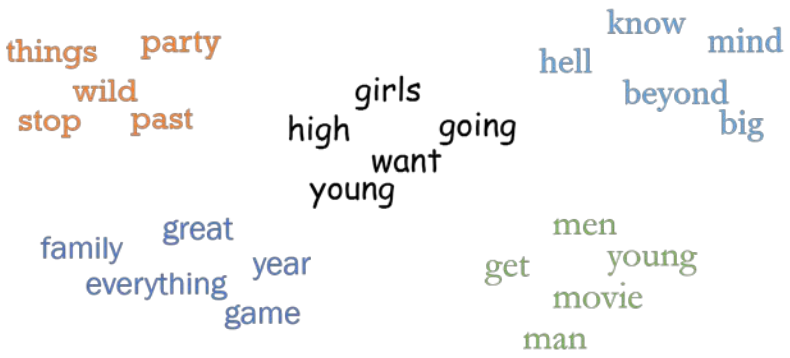




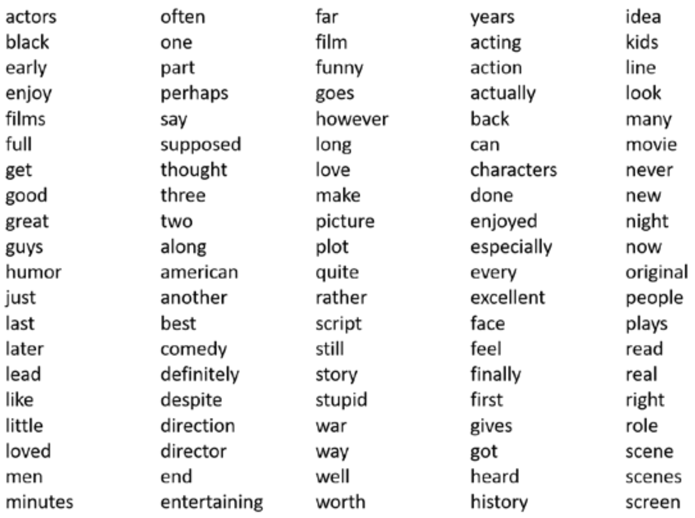
Bag of Words



actors
black
early
enjoy
films
full
get
good
great
guys
humor
just
last
later
lead
like
little
loved
men
minutes
often
one
part
perhaps
say
supposed
thought
three
two
along
american
another
best
comedy
definitely
despite
direction
director
end
entertaining
far
film
funny
goes
however
long
love
make
picture
plot
quite
rather
script
still
story
stupid
war
way
well
worth
years
acting
action
actually
back
can
characters
done
enjoyed
especially
every
excellent
face
feel
finally
first
gives
got
heard
history
idea
kids
line
look
many
movie
never
new
night
now
original
people
plays
read
real
right
role
scene
scenes
screen



Sentiment Analysis



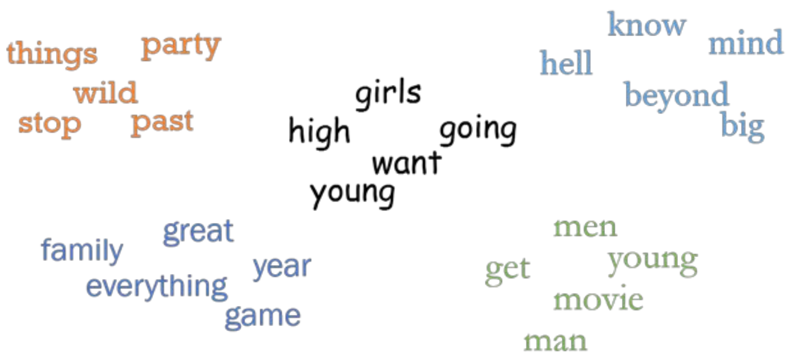
actors
black
early
enjoy
films
full
get
good
great
guys
humor
just
last
later
lead
like
little
loved
men
minutes

often
one
part
perhaps
say
supposed
thought
three
two
along
american
another
best
comedy
definitely
despite
direction
director
end
entertaining

far
film
funny
goes
however
long
love
make
picture
plot
quite
rather
script
still
story
stupid
war
way
well
worth

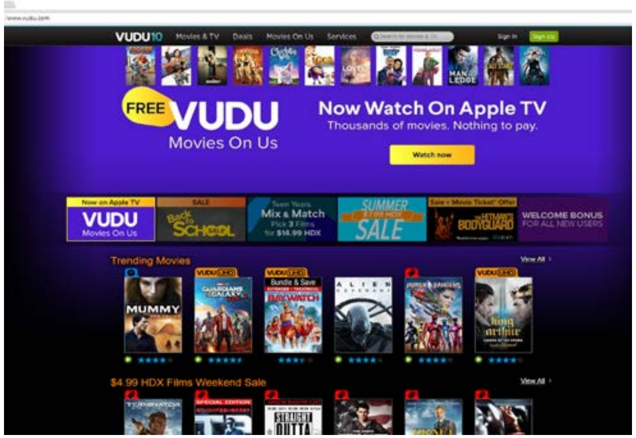
years
acting
action
actually
back
can
characters
done
enjoyed
especially
every
excellent
face
feel
finally
first
gives
got
heard
history

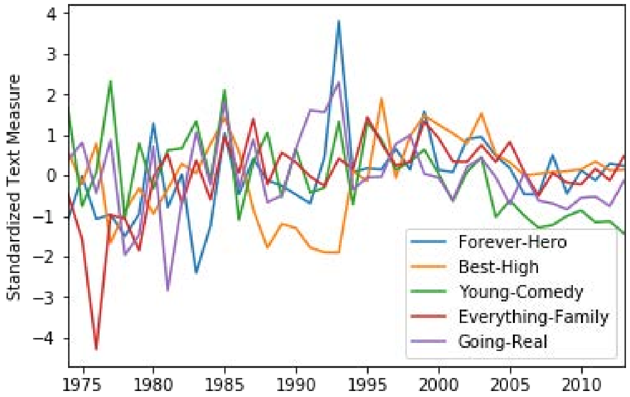
idea
kids
line
look
many
movie
never
new
night
now
original
people
plays
read
real
right
role
scene
scenes
screen



Recommendation

- Text analytics requires a move from unstructured to structured
- Some primary methods
  - Bag of words
  - Natural language processing
  - Sentiment analysis
- Text analytics gives new insight into consumer preferences





Text Processing

School of Information Studies
Syracuse University

Patterns in Text Data

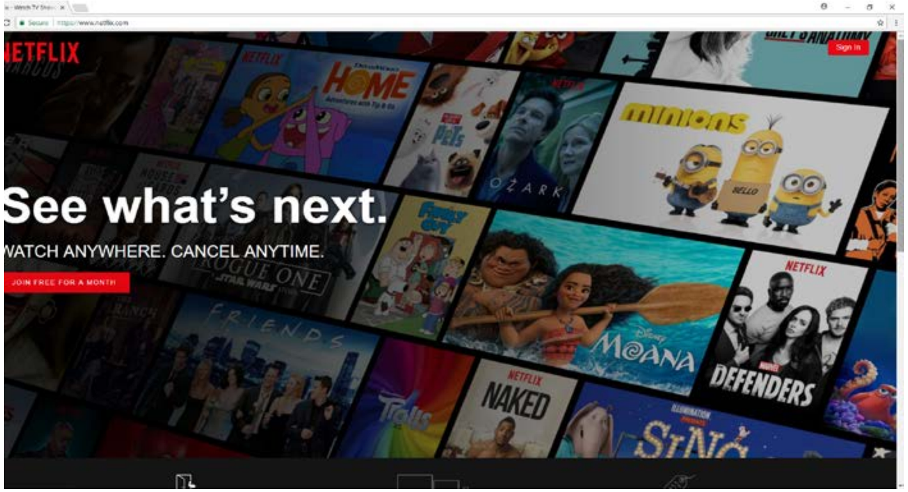
actors
black
early
enjoy
films
full
get
good
great
guys
humor
just
last
later
lead
like
little
loved
men
minutes

often
one
part
perhaps
say
supposed
thought
three
two
along
american
another
best
comedy
definitely
despite
direction
director
end
entertaining

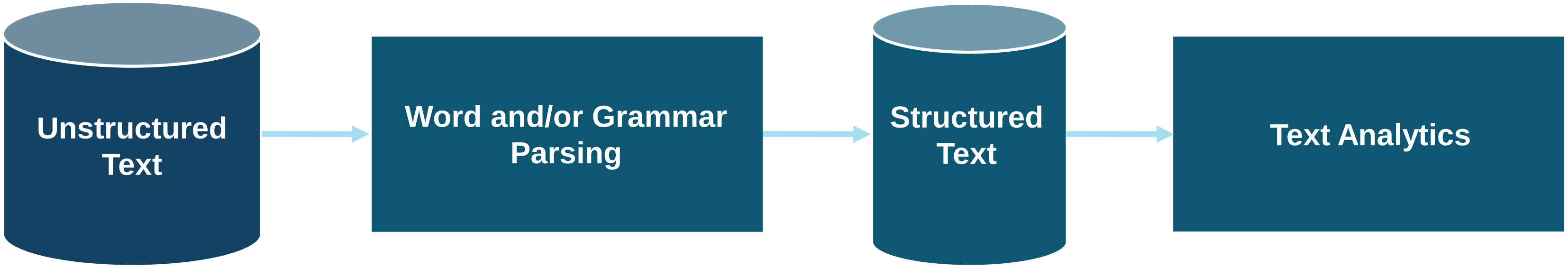
far
film
funny
goes
however
long
love
make
picture
plot
quite
rather
script
still
story
stupid
war
way
well
worth

years
acting
action
actually
back
can
characters
done
enjoyed
especially
every
excellent
face
feel
finally
first
gives
got
heard
history

idea
kids
line
look
many
movie
never
new
night
now
original
people
plays
read
real
right
role
scene
scenes
screen



Text Processing



Source: Adapted from Miller (2005).

Term Document Matrix

Covert (2017)

You don’t get more back to the future than this wholehearted embrace of sci-fi and golden oldies nostalgia. No franchise is so gifted at exploiting the popularity of superhero movies while satirizing the genre … “Guardians of the Galaxy Vol. 2” is just right. It’s hyperbolic nonsense wrapped in the colors of a neon rainbow, bouncing from one artfully wacky scenario to the next. Here it’s sleekly futuristic, there it’s older than mud. It’s the galaxy’s silliest thrill show.

	Document		
Term	Covert	Lapin	…
future	1	0	
oldies	1	1	
movie	1	1	
galaxy	1	1	
neon	1	0	
rainbow	1	0	
battle	0	1	
families	0	1	
…			

Lapin (2017)

Swapping out yet another giant, ponderous battle for a cute creature and Oldies radio is exactly the sort of action movie send-up the genre needs right now, so it's a shame the rest of the film doesn't follow that same spirit of rule-breaking. Instead, we get straight-laced morals about families that feel like warmed-over *Fast and Furious* wisdom.

Source: Adapted from Miller (2005).

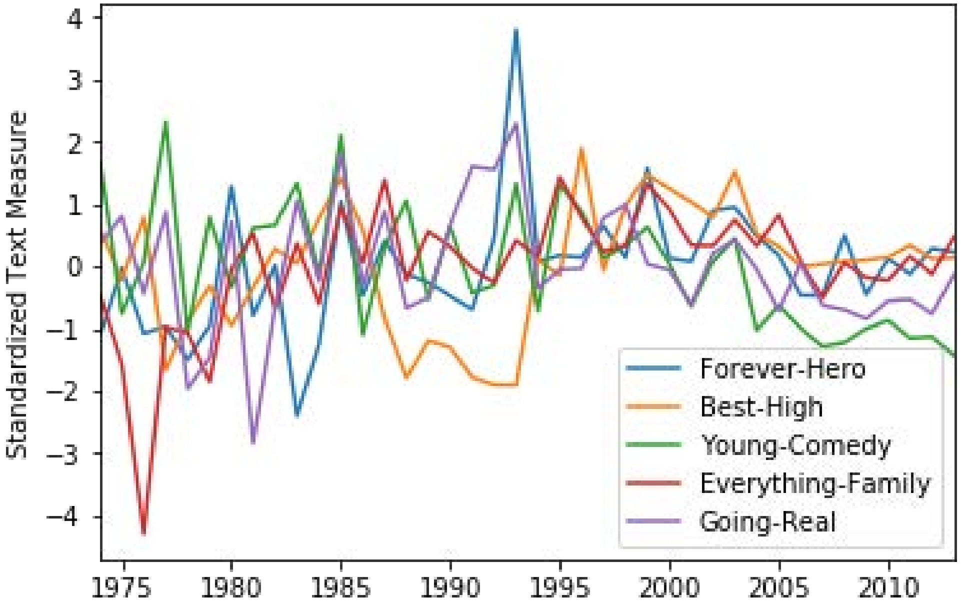
Text Analytics

Unsupervised

- No response or class
- Patterns or trends over time/population

Supervised

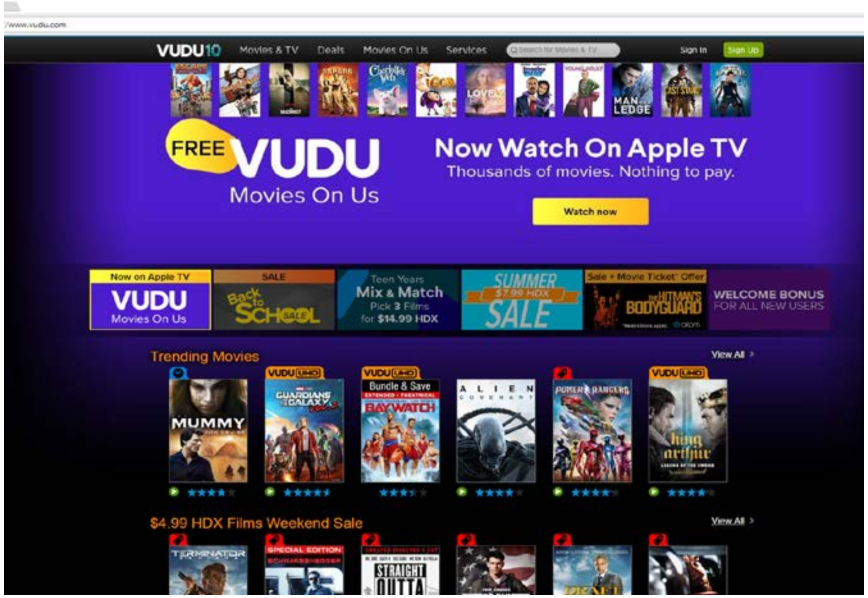
- Response or class
- Spam or relevant search

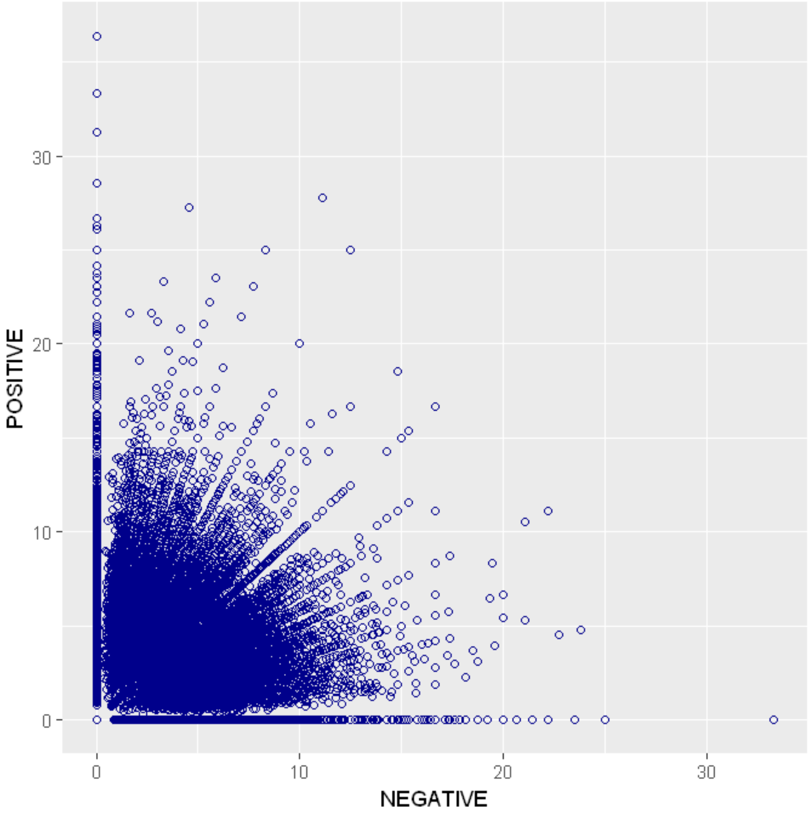


Foundation

School of Information Studies
Syracuse University

Patterns in Text Data





Sentiment Analysis

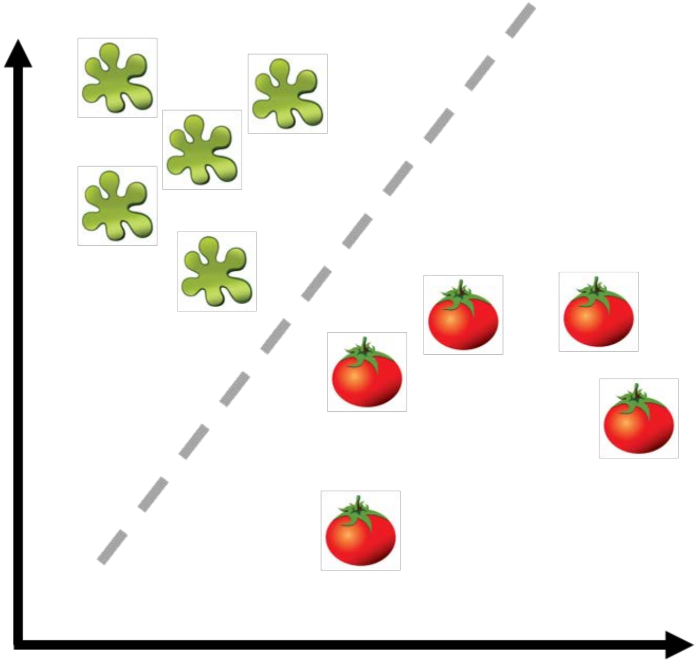
- Measurement focused
  - Positive words
  - Negative words
- Dictionary
  - Not alphabetized definitions
  - Repository of lists
  - Positive and negative words
  - Synonyms and antonyms

Simple Difference

- Compute difference scores
- Cutoff predicted on training set
- Positive minus negative scores

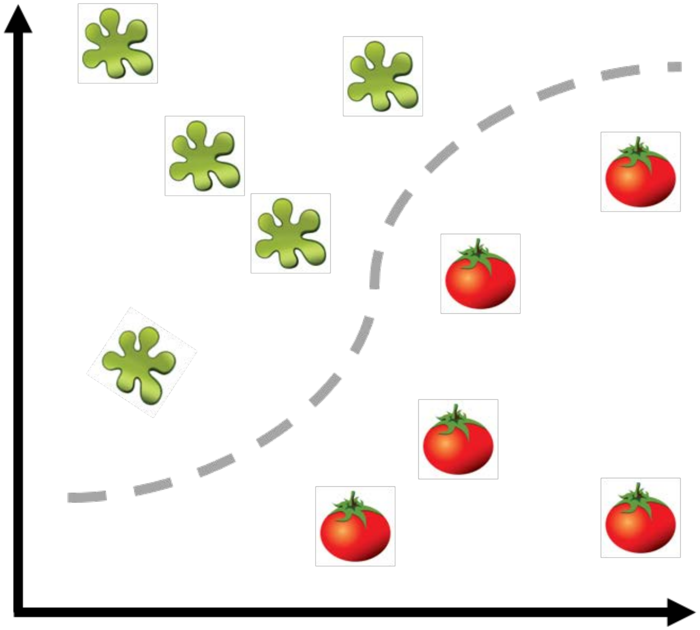
Regression Difference

- Linear regression to determine weights
- Combine positive and negative scores into a linear predictor
- Predicted rating becomes a cutoff



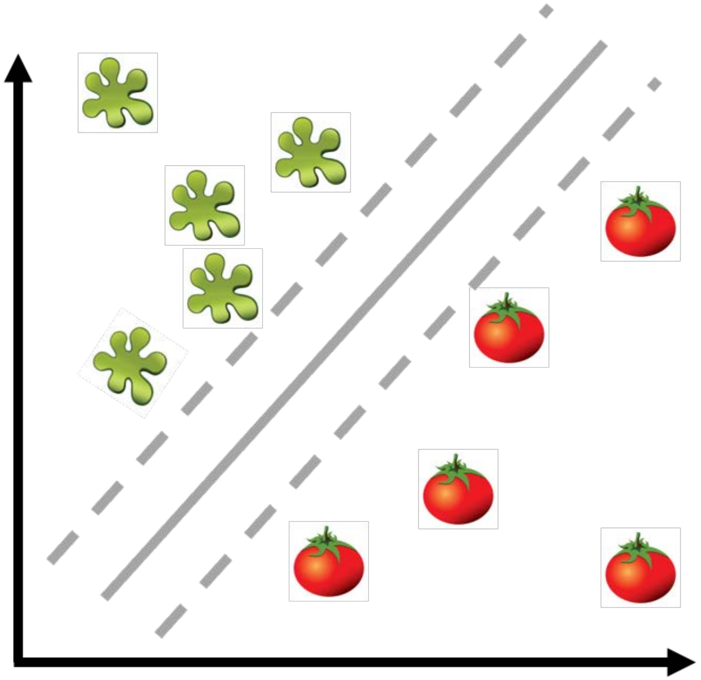
Logistic Regression

- Predicting a binary response
- Select useful predictors from sentiment words
- Weights calculated using maximum likelihood



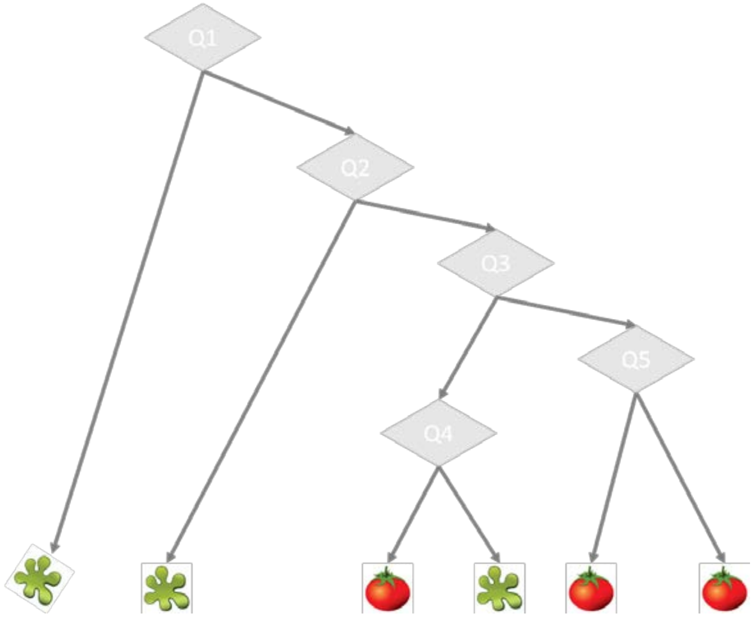
Support Vector Machines

- Effective in problems with large numbers of explanatory variables
- Classifies using a separating hyperplane
- Memory efficient



Random Forests

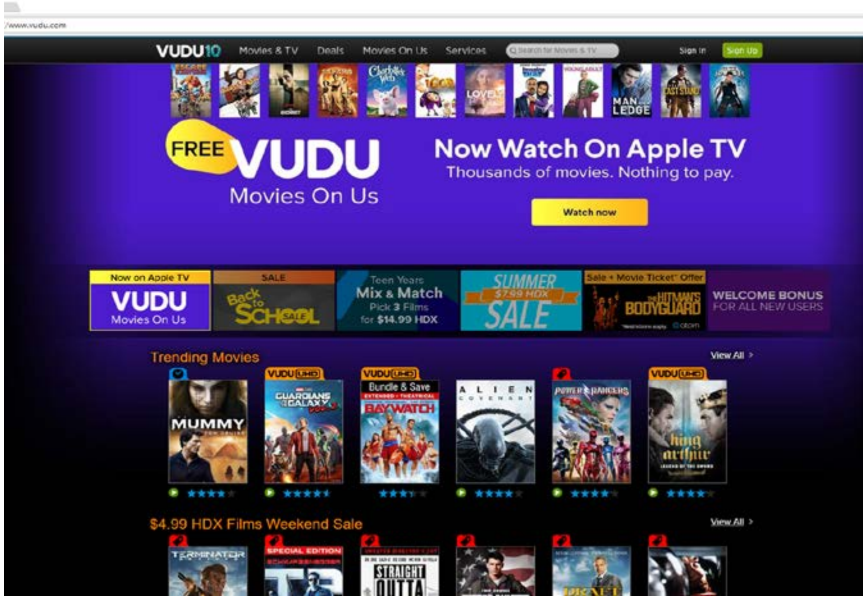
- Ensemble method using multiple decision trees
- Recursive partitioning on the training set
- Effective with large number of explanatory variables

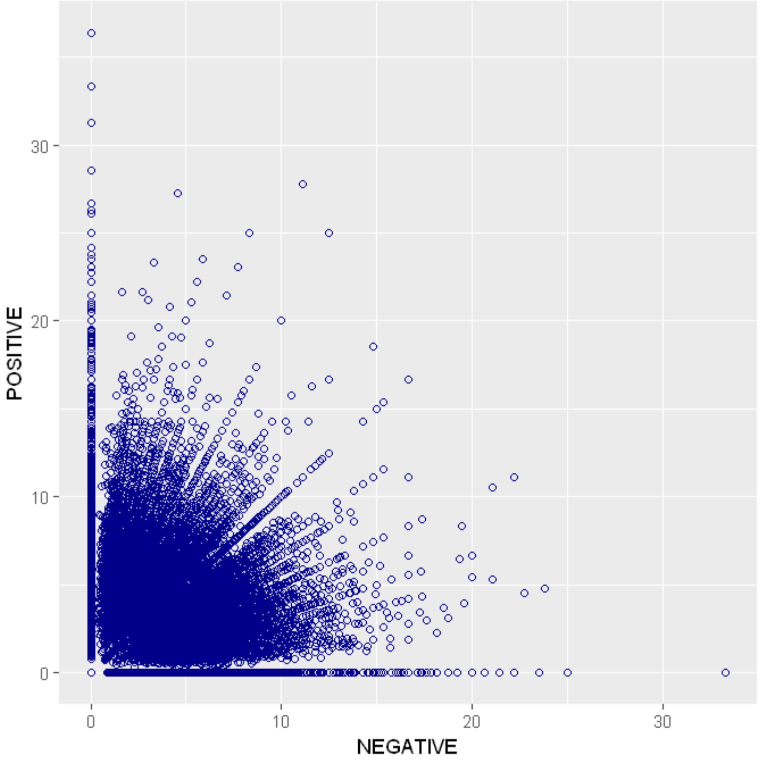


Alternatives

School of Information Studies
Syracuse University

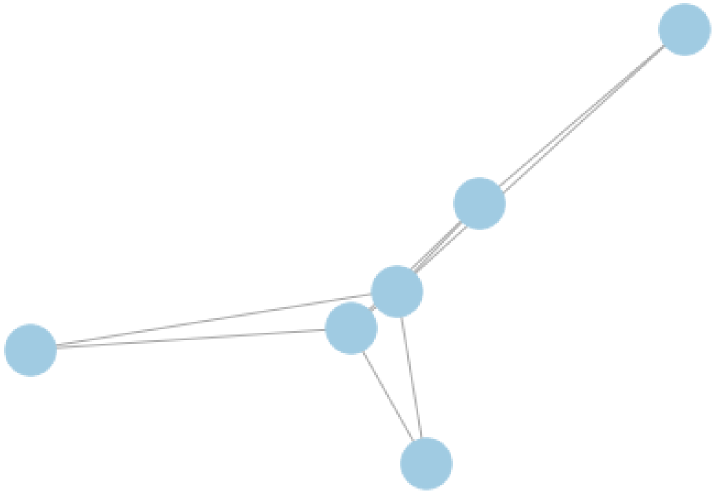
Alternatives for Text Data





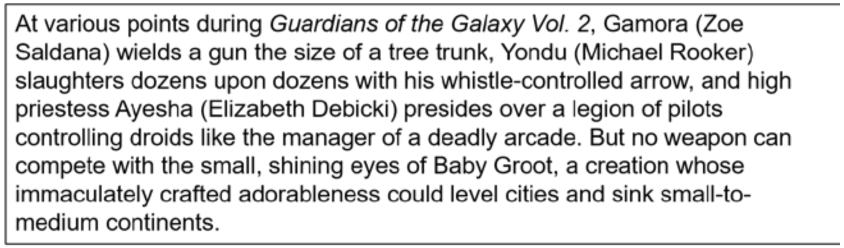
Latent Semantic Analysis

- Identify relationships between words
- Concepts within the documents
- Eliminate the noise and provide simple representation of the data



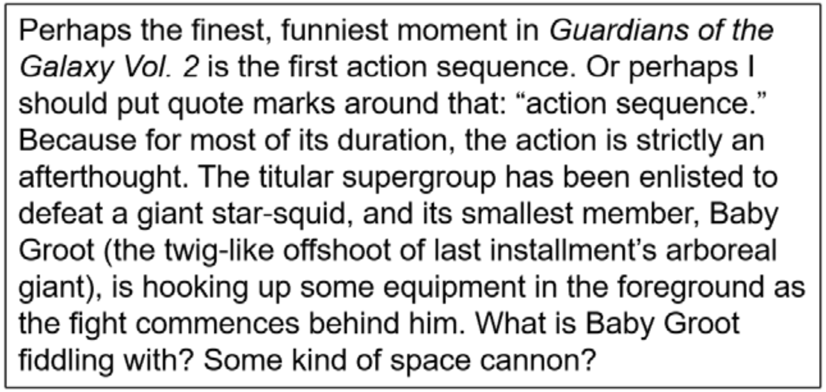
Latent Semantic Analysis (cont.)

Willmore (2017)



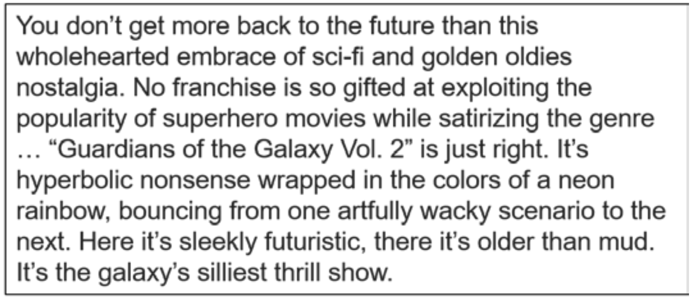
At various points during *Guardians of the Galaxy Vol. 2*, Gamora (Zoe Saldana) wields a gun the size of a tree trunk, Yondu (Michael Rooker) slaughters dozens upon dozens with his whistle-controlled arrow, and high priestess Ayesha (Elizabeth Debicki) presides over a legion of pilots controlling droids like the manager of a deadly arcade. But no weapon can compete with the small, shining eyes of Baby Groot, a creation whose immaculately crafted adorableness could level cities and sink small-to-medium continents.

Orr (2017)



Perhaps the finest, funniest moment in *Guardians of the Galaxy Vol. 2* is the first action sequence. Or perhaps I should put quote marks around that: “action sequence.” Because for most of its duration, the action is strictly an afterthought. The titular supergroup has been enlisted to defeat a giant star-squid, and its smallest member, Baby Groot (the twig-like offshoot of last installment’s arboreal giant), is hooking up some equipment in the foreground as the fight commences behind him. What is Baby Groot fiddling with? Some kind of space cannon?

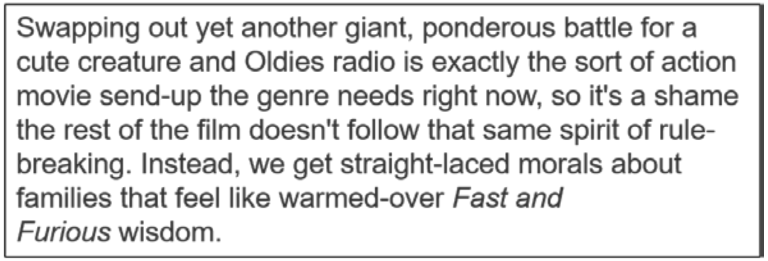
Covert (2017)



You don’t get more back to the future than this wholehearted embrace of sci-fi and golden oldies nostalgia. No franchise is so gifted at exploiting the popularity of superhero movies while satirizing the genre … “Guardians of the Galaxy Vol. 2” is just right. It’s hyperbolic nonsense wrapped in the colors of a neon rainbow, bouncing from one artfully wacky scenario to the next. Here it’s sleekly futuristic, there it’s older than mud. It’s the galaxy’s silliest thrill show.

GG2 delivers easy summer movie fare, has a message of family, stars Baby Groot, and is generally liked by reviewers.

Lapin (2017)



Swapping out yet another giant, ponderous battle for a cute creature and Oldies radio is exactly the sort of action movie send-up the genre needs right now, so it's a shame the rest of the film doesn't follow that same spirit of rule-breaking. Instead, we get straight-laced morals about families that feel like warmed-over *Fast and Furious* wisdom.

Latent Semantic Analysis (cont.)

From our term document matrix:

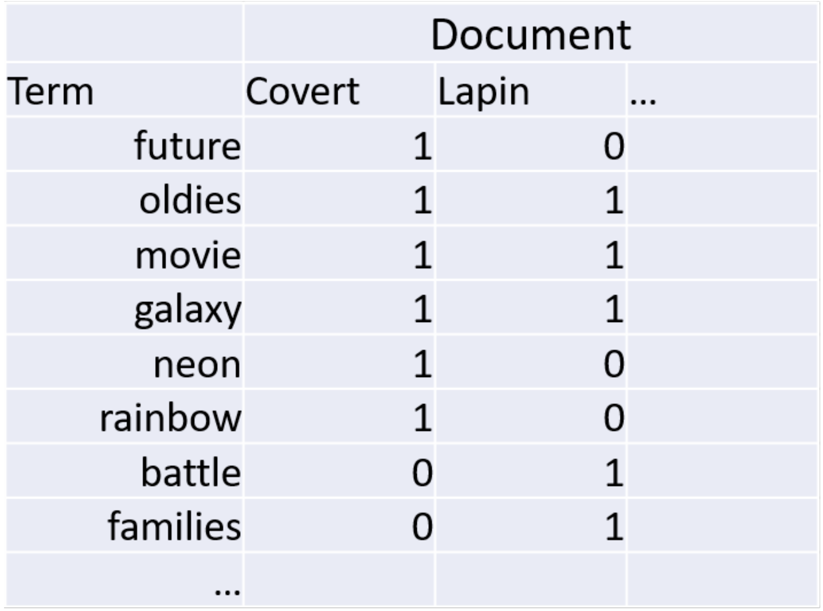
$$M = USV^T$$

M is $m \times n$

U is $m \times k$

S is $k \times k$

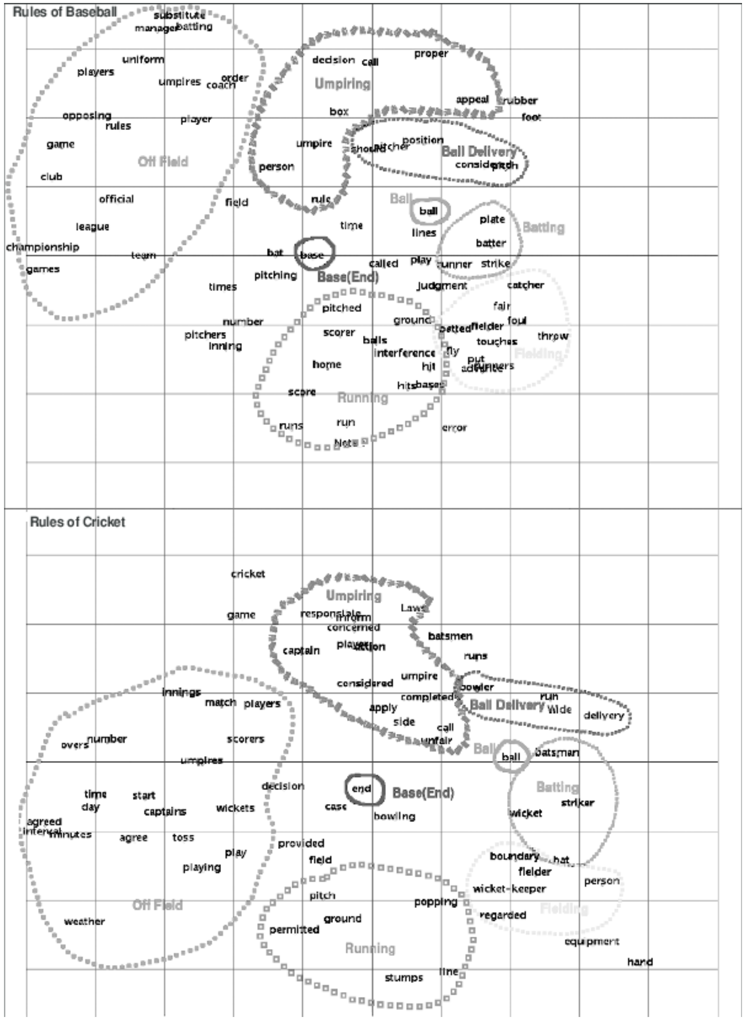
V is $k \times n$



	Document		
Term	Covert	Lapin	…
future	1	0	
oldies	1	1	
movie	1	1	
galaxy	1	1	
neon	1	0	
rainbow	1	0	
battle	0	1	
families	0	1	
…			

Some Warnings

- The “good-bad” dimension
- Specific terms vs. latent semantic dimensions
- Relation to business outcomes and insights



Rules of Baseball
substitute
manager
batting
uniform
players
umpires
coach
order
opposing
rules
player
game
Off Field
club
official
league
championship
team
games
decision
call
proper
Umpiring
appeal
rubber
foot
box
umpire
should
pitcher
position
Ball Delivery
considered
pitch
person
field
rule
Ball
ball
time
lines
plate
Batting
batter
bat
base
called
play
runner
strike
pitching
Base(End)
times
judgment
catcher
pitched
fair
number
ground
batted
fielder
foul
pitchers
inning
scorer
balls
touches
throw
interference
fly
home
hit
put
advance
runners
Fielding
score
hits
bases
Running
runs
run
Note
error
Rules of Cricket
cricket
Umpiring
game
responsible
inform
concerned
Laws
batsmen
captain
player
action
runs
considered
umpire
bowler
innings
completed
run
match
players
apply
Ball Delivery
Wide
delivery
side
call
overs
number
scorers
unfair
Ball
ball
batsman
umpires
decision
end
Base(End)
Batting
time
start
striker
day
captains
wickets
case
bowling
wicket
agreed
interval
minutes
agree
toss
provided
play
field
boundary
bat
playing
fielder
person
wicket-keeper
pitch
popping
Off Field
regarded
Fielding
weather
ground
permitted
equipment
Running
hand
stumps
line