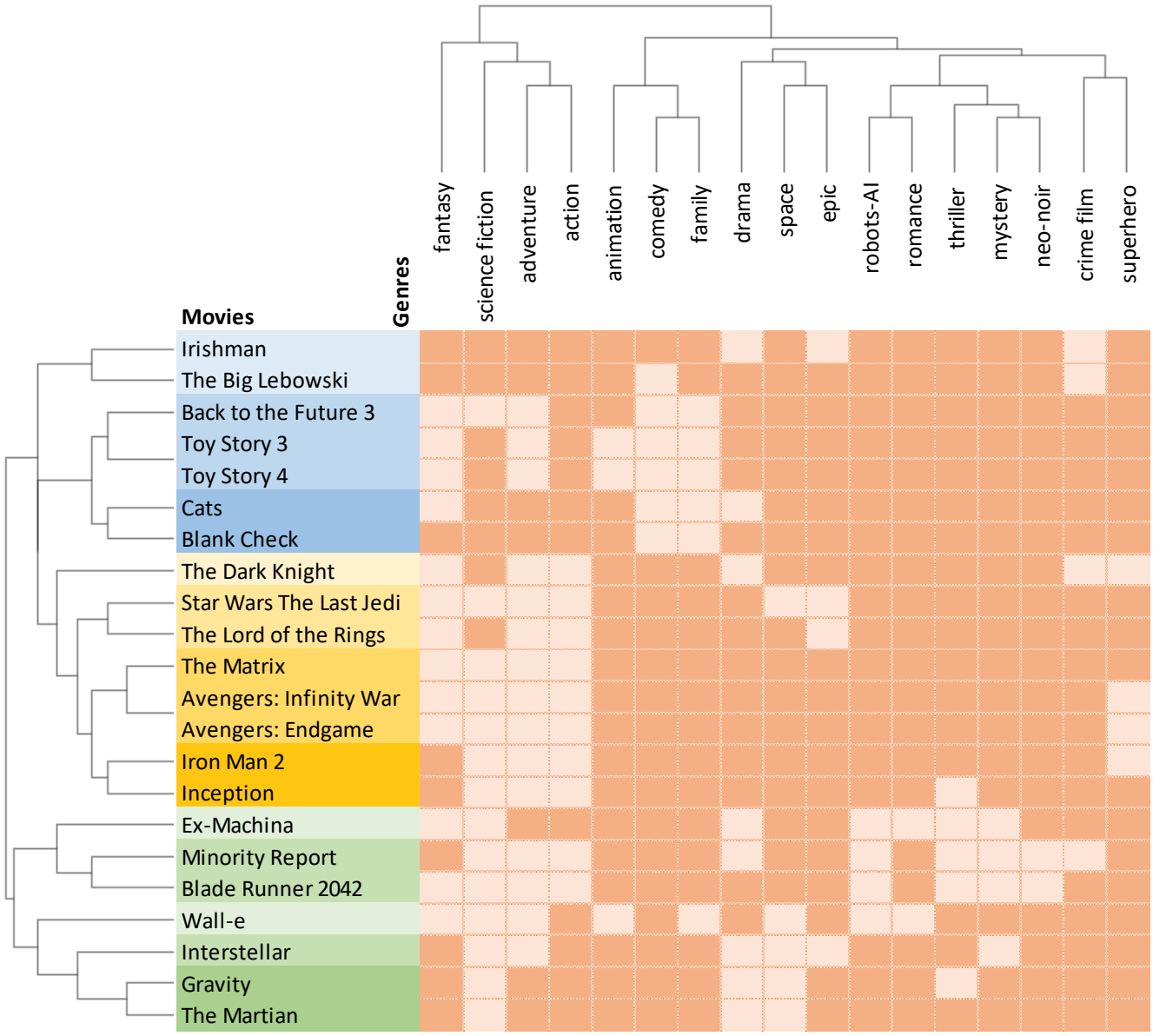
Alisher Siddikov
MSDS 453 - NLP
Assignment 3
6/6/2020

This is part 3, the continuation of the previous assignments. Since I have not seen the ¾ movies, I do not have domain knowledge about their genres. However, I utilized Google knowledge graph and discussion 5 to compile genres of the films and then built a dendrogram to see how they cluster. Further details about the movie genres table can be seen in appendix 1.



There are 22 movies, and the lighter color of the dendrogram represents the movie genres. In high level, they are divided into three major genres:

1. fantasy, science-fiction, adventure, action, and some superhero
    - **The Dark Knight** *(fantasy, adventure, action, superhero)*
    - **Star Wars the Last Jedi** *(fantasy, sci-fi, action, adventure, space)*
    - **The Lord of the Rings** *(fantasy, adventure, action)*
    - **The Matrix** *(fantasy, sci-fi, adventure, action)*
    - **Avengers: Infinity War** *(fantasy, sci-fi, adventure, action, superhero)*
    - **Avengers: Endgame** *(fantasy, sci-fi, adventure, action, superhero)*
    - **Iron Man 2** *(sci-fi, adventure, action, superhero)*
    - **Inception** *(sci-fi, adventure, action, thriller)*
2. family, animation, comedy, and some drama and crime
    - **Irishman** *(crime, drama)*
    - **The Big Lebowski** *(crime, comedy)*
    - **Back to the Future 3** *(family, comedy, fantasy, sci-fi, adventures)*
    - **Toy Story 3** *(family, animation, comedy, fantasy)*
    - **Toy Story 4** *(family, animation, comedy, fantasy)*
    - **Cats** *(family, comedy, drama, comedy, fantasy)*
    - **Blank Check** *(family, comedy)*
3. science fiction, drama, and some space, robotics-AI, romance, thriller, mystery
    - **Ex-Machina** *(sci-fi, fantasy, drama, robotics-AI, romance, thriller)*
    - **Wall-e** *(sci-fi, fantasy, animation, adventure, robotics-AI, romance, space)*
    - **Minority Report** *(sci-fi, robots-AI, thriller, mystery)*
    - **Blade Runner 2042** *(sci-fi, robots-AI, thriller, mystery)*
    - **Interstellar** *(sci-fi, drama, space)*
    - **Gravity** *(sci-fi, drama, space)*
    - **The Martian** *(sci-fi, drama, space)*

So, we can see that movies can be grouped or clustered in many different ways because they share several genres. For example, the following Wall-E can be bundled with several different genres:

- Wall-E and Toy Story - animation, family, fantasy and adventure
- Wall-E and Ex-Machina – robotics-AI, romance, fantasy and sci-fi
- Wall-E and Gravity/The Martian/Interstellar – sci-fi and space
- Wall-E and Back to the Future – futuristic, family, fantasy, sci-fi, and adventures

Let's see how our k-mean clustering groups our movies. To arrive at this conclusion, I expanded my previous stop words list. Now I include all the proper nouns, prepositions, determiners, and less meaningful words (verbs

adjectives and adverbs) to the stop words list by using NLTK library pos tag and lemmatize functions. I spent most of my time to refine the stop word. Please see the stop words in the appendices section. Stop words in Appendix 2 contain about 1728 words.

- **'earth'**: ['home planet', 'world'],
- **'space'**: ['planets', 'planet', 'universe', 'galaxy', 'spacewalking', 'cosmic', 'moon', 'mars', 'star'],
- **'spacecraft'**: ['axiom', 'interstellar', 'space station', 'satellite'],
- **'robot'**: ['robots', 'walles', 'walle', 'wally', 'eve', 'machina', 'conscious machine', 'humanlike creature', 'artificial human', 'intelligent creature', 'machine', 'machines', 'robotic'],
- **'replicant - AI'**: ['artificial intelligence', 'artificially intelligent', 'humanlike intelligence', 'artificial intelligent', 'intelligence'],
- **'technology'**: ['technologies', 'algorithm', 'matrix', 'computer', 'tech', 'optical recognition', 'gadget'],
- **'human'**: ['humanity', 'humans', 'people', 'person', 'mankind'],
- **'family'**: ['family', 'father', 'babies'],
- **'woman'**: ['women', 'female', 'girls'],
- **'kids**: ['teenager', 'teenage', 'teen', 'young', 'children', 'child'],
- **'money'**: ['fund', 'bank', 'banks'],
- **'crime'**: ['criminal', 'organized crime', 'mafia', 'mob', 'bully'],
- **'violence'**: ['destroying', 'destroyed', 'destroyers', 'destroy', 'fighting', 'fights', 'fight', 'revenge'],
- **'killing'**: ['killing', 'killer', 'killed', 'terrorist', 'terrorism', 'militant', 'death'],
- **'kidnap'**: ['kidnapped', 'kidnapping', 'abducted'],
- **'music'**: ['musical', 'dance', 'choreography'],
- **'dream'**: ['dreams', 'inception', 'vision', 'awake'],
- **'nerdism'**: ['whiz', 'geek', 'genius', 'hacker', 'talented', 'talents'],
- **'future'**: ['futuristic', 'alternate vision', 'precogs', 'precrime', 'dystopian'],
- **'fantasy'**: ['science fiction', 'scifi', 'fantastical', 'fantastic']

During our last three class discussions, we came to the following conclusions. Since our corpus was based on the movie reviews rather than movie descriptions, there would be clusters where unrelated movies put together depending on how the reviewer "feel" about the specific movie. Also, we can see that all three reviews of the same movie could not be clustered together because of the reviewers' sentiment. We can see that clusters based on the sentiment of the language in the reviews - for example, certain clusters having similarly harsh critique

while the clustered films appear unrelated in terms of plot/genre. Please note that there are many different ways to write movie reviews, depending upon an author's perceptions. Sometimes reviews take a genre-based approach, in which the focus of the review is in drawing comparisons to other movies. Or, a review might focus more on the plot, or dive deeper into the themes.

Here are the visual friendly k-mean cluster results – I increased the centroids from 5 to 30 to see what terms trigger the clustering. The python version output can be accessed from appendix 3. In general, the movie reviews in cluster 0 falls to fantasy, sci-fi, adventure genres. I can see that Wall-E, Ex-Machine, Minority Report, and Blade Runner were clustered with robotics, sci-fi, and fantasy genres by using robots, artificial intelligent, replicants, humanlike intelligence technology, AI vs. humans, space, and earth. Ex-Machina review mentions the future technology of artificial intelligence, which is an invention of lethal human face robots, and it compares robots to humans, whether artificial intelligent creations think and act like humans. The wall-E review mentions that industrious robots sent to earth from the spacecraft and robotic creatures express a rainbow of emotions. Minority Report review mentions about the spider robots for tracking humans and search buildings. Blade Runner review mentions about six replicants genetically designed artificial human beings intended as slaves on the earth and four replicants take over a space shuttle and return to earth. We can see that Matrix movie review also mentions how AI was a threat to human judgment and how an AI of human creation designed the dream world. Cats and Wall-E were clustered with family and fantasy genres, and I believe that listening to music, musical themes, humanlike intelligence, and human hybrids of cat terms triggered those movies to cluster together. I believe Cats' movie should have been clustered with Blank Check movie because they share comedy and family genres. Inception, Matrix, and Blade Runner movie reviews were clustered with fantasy, adventure, sci-fi, and some action genres by using the following terms: dreams and simulate reality. In the Matrix movie review, we can read the simulated reality – computer-generated dreams and Inception movie review mentions that the minds can be accessed in a dream state. However, clustered terms do not have context; for example, Blade Runner review mentions if androids dream of electric sheep. In here, the dream does not cluster with Inception and Matrix dreams – simulated realities.

Cluster 0:		Cluster 1:		Cluster 2:	
human	nerdism	space	real	woman	road
robot	invent	dream	inspire	money	spot
technology	waste	spacecraft	hero	critic	parody
future	mind	hope	comic	life	earth
earth	evil	violence	move	kids	amuse park
dream	turn	human	money	action	amuse
ai-replicant	plant	mean	light	male	lead
simul	predict	daughter	resist	crime	robot
music	show	earth	crime	family	failure
history	control	kill	mind	creature	park
space	execute	enjoy	fantasy	cash	gender
poll	computer	shoot	lead	human	book
life	develop	kids	young	parent	happy
cruise	real	family	book	transform	power
crime	consumer	life	strike	concept	journey

Cluster 0 titles:	Cluster 1 titles:	Cluster 2 titles:
Back to the Future 3 - 1	Avengers: End Game	Blank Check - 1
Back to the Future 3 - 1	Avengers: Inf. War	Blank Check - 2
Back to the Future 3 - 2	Blade Runner 2042 - 3	Blank Check - 3
Blade Runner 2042 - 1	Gravity - Space	Cats - 1
Blade Runner 2042 - 2	Inception - 2	Ex-Machina - 1
Cats - 2	Inception - 3	Irishman - 2
Cats - 3	Interstellar - 1	Irishman - 3
Ex-Machina - 2	Interstellar - 2	Iron Man 2
Ex-Machina - 3	Interstellar - 3	The Big Lebowski - 3
Inception - 1	Interstellar - Space	The Lord of the Rings - 3
Minority Report - 1	Irishman - 1	Toy Story 4 - 1
Minority Report - 2	Star Wars The Last Jedi - 1	Toy Story 4 - 2
Minority Report - 3	Star Wars The Last Jedi - 2	Toy Story 4 - 3
The Lord of the Rings - 1	Star Wars The Last Jedi - 2	
The Lord of the Rings - 2	Star Wars The Last Jedi - 3	
The Matrix - 1	The Big Lebowski - 1	
The Matrix - 2	The Big Lebowski - 2	
The Matrix - 3	The Dark Knight - 1	
Wall-E - 1	The Dark Knight - 2	
Wall-E - 2	The Dark Knight - 3	
Wall-E - 3	The Martian - Space	

Minority Report, Back to the Future, and Wall-E movie reviews were clustered with family and, thriller, mystery in addition to the sci-fi and adventure genres. The terms make them closer are futuristic, forward-looking, predicting, and comparison to history or current times. Minority Report review talks about preventing murders

before they happen through the future visions of three mutated humans. Back to the Future review mentions changing the history or future history by meeting with his own future mother. The wall-E movie review mentions how 800 years of Future looks when humans leave the trashed earth. The Lord of the Rings does not belong to this cluster because the clustering terms were not used in the same context as other movies in this cluster. For example, the reviewer writes about the future middle earth threatened Sauron obsessed evil ring in The Lord of the Rings movie.

The movie reviews in cluster 1 fall to various genres ranges from sci-fi, drama, space, and crime. I can see that Irishman and The Big Lebowski movie reviews can be clustered together with crime genres. In Irishman review-1 mention organized crimes, gangsters, violence took place and killed without remorse. The Big Lebowski movie review mentions the kidnapping and delivering ransom money, high spending young wife who owes them some money. I can also see that Interstellar, The Martian, Gravity, and Star Wars reviews clustered together as sci-fi, drama, and space genres because they share a mission to leave the earth and travel to space with a spacecraft. These movies also involve hope and life in space. Both Avengers and the Dark Knight movie reviews share the same genres, such as action and superhero films. The Dark Knights also have a crime genre that can group with Irishman and The Big Lebowski movie. Movie Inception movie reviews 2 and 3 should go to the cluster one. It started to show up in this cluster because maze architecture space should be built-in in the dream. So, the space term in Inception does not refer to the galaxy – space. The context here is different. However, the Inception review number one was clustered correctly in our previous cluster group.

The last cluster group almost represents the top dendrogram genre cluster (colored in blue), and this cluster is based on family, animation, comedy, and some drama and crime, and I agree how it was clustered in here. I can see why Blank Checks was clustered with The Big Lebowski in this cluster group. It is because they are a comedy genre and involves money and cash. Toy story is an animated family movie that portrays life between humans/kids and toys; they also include male and female toy genders, which can be grouped with Cats movie reviews. Since our two previous clusters cover AI, robotics, human replicant, dreams, crime, and space, I believe we need to make this cluster to cover family, comedy, and drama. The remaining reviews of Irishman and The Big Lebowski movies can go back to cluster two, where some of their reviews reside there under the crime scene.

We learned that movies could be grouped or clustered in many different ways because they share several genres. Since our corpus was based on movie reviews rather than movie descriptions, there were clusters where unrelated movies put together depending on the reviewer's feelings, sentiment, and intention. We cleaned the movie review corpus, stripped out stop words, created equivalence classes, and then ran the K-mean cluster to see high-level cluster groupings. We explored three different clusters. We can also change the size of the k-mean clusters. When I chose 10 clusters, the different reviews of the same movie usually got clustered together, and it was not that useful. So I decided to stick with a high level of clustering.

Appendices:

Appendix 1. Google Knowledge and Discussion 5

	Movie	Genre
1	Avengers: Endgame	action, adventure, fantasy, science fiction, superhero
2	Avengers: Infinity War	action, adventure, fantasy, science fiction, superhero
3	Back to the Future 3: Reboot: From Archives LA Times	adventure, comedy, family, fantasy, science fiction
4	Blade Runner 2042	action, adventure, fantasy, mystery, neo-noir, science fiction, thriller, robots-AI
5	Blank Check: Check Cashes In	comedy, family
6	Cats	comedy, drama, family, fantasy
7	Ex-Machina	drama, fantasy, mystery, romance, science fiction, thriller, robots-AI
8	Gravity	drama, science fiction, thriller, space
9	Inception: Architecture of the Minds: Dreams on Top: This Time the Dream	action, adventure, science fiction, thriller
10	Interstellar: Off to the Stars	adventure, drama, epic, mystery, science fiction, space
11	Irishman	crime film, drama, epic
12	Iron Man 2	action, adventure, science fiction, superhero
13	Minority Report	action, adventure, crime film, drama, mystery, neo-noir, science fiction, thriller, robots-AI
14	Star Wars the Last Jedi: Review the Latest	action, adventure, epic, fantasy, science fiction, space
15	The Big Lebowski: El-Duderino: Dude Bowls: Film Critics Blind	comedy, crime film
16	The Dark Knight (Batman Begins)	action, adventure, crime film, drama, fantasy, superhero
17	The Lord of the Rings (The Fellowship of the Ring)	action, adventure, epic, fantasy
18	The Martian	drama, science fiction, space
19	The Matrix	action, adventure, fantasy, science fiction
20	Toy Story 3	adventure, animation, comedy, family, fantasy
21	Toy Story 4	adventure, animation, comedy, family, fantasy
22	Wall-e	adventure, animation, family, fantasy, science fiction, romance, robots-AI, space

Appendix 2. Stop words

aaron	audiences	clear	draws	finally	hudson	little	named	provides	scotts	spielbergs	though	voiced
ability	authors	clearly	easily	finding	hudsons	lives	narrative	providing	screen	stand	thought	wanted
abrams	awakens	close	eckhart	finds	ideas	living	nathan	pulls	screenwrite	stands	thoughts	wanting
absolute	baggins	cobbs	effects	first	images	lloyd	needs	purpose	script	stark	three	wants
abstract	barely	collection	efforts	focus	impressive	longer	never	pursued	sebastian	start	throughout	watanabe
according	based	combinatior	eight	follows	including	looked	nolan	question	second	started	times	watching
accustomed	basic	comes	either	footage	indeed	looking	nolans	questions	seeing	statement	title	watney
achieves	become	comfortable	endgame	forget	infinity	looks	normal	quiet	seeks	steve	today	waves
acronym	becomes	coming	ending	forward	initially	lucas	nostalgia	quigley	seemingly	steven	together	wears
across	becoming	common	engage	found	inside	macintosh	nothing	quite	seems	still	tolkien	webber
activity	began	completely	engages	francesca	inspired	mackenzie	novel	rachel	sense	stoner	totems	whatever
actor	begin	consequenc	engaging	franchise	instead	major	numbers	raising	sequel	stories	touches	whats
actors	begins	cooper	enjoy	frank	intended	maker	obvious	rather	sequels	story	toward	whether
actually	behind	could	enjoyed	frankenstein	interest	makes	offers	realize	sequence	stranded	towering	white
addition	beings	course	enjoyment	frankly	interesting	makeup	often	realized	sequences	stripped	tried	whole
allow	belongs	cowrote	enough	frodo	involves	making	opening	reallife	series	strong	tries	whose
almost	benefits	create	entertainme	fully	ironically	manner	original	really	seven	strongly	trilogy	wilson
along	better	created	entire	gamechangi	isaac	martian	oscar	reason	shadows	stuff	truly	winnertakeal
already	beyond	creates	entirely	gamely	issues	marty	others	recent	share	stunning	truth	wires
although	black	critics	escape	games	jackson	marvel	pacino	recently	sheeran	style	trying	wishes
always	blank	currently	especially	garland	james	matthew	palpatine	reflects	shifting	subject	turing	within
amazing	bonnie	daisy	essential	generally	january	maybe	particular	release	showing	suggest	turkel	without
ambition	bonsall	daniels	eventually	genre	jennifer	mcconaughe	particularly	replicants	shows	suggests	turned	witwer
america	boxoffice	david	every	george	johnson	mcfly	performance	report	simple	summer	turns	wonder
american	brand	dawes	everyone	getting	jonathan	mckellen	perhaps	represent	simply	superheroes	tyrell	wonderfully
among	breaking	decade	everything	given	jones	meaning	pesci	represents	simulated	supply	ultimate	wonders
amount	brian	decide	everywhere	giving	judi	meaningful	peter	requires	simultaneou	supposed	ultimately	woody
anderton	bridges	decided	exactly	going	justin	means	philip	response	since	surely	understand	words
andy	brothers	decidedly	except	gordon	keeps	meant	pieces	result	situations	surprised	undoubtedl	worked
animated	brown	decides	exciting	gotham	knight	meanwhile	pixar	return	skywalker	surprising	unexpected	working
animation	bruce	decision	exist	gravity	knowing	meets	place	reveals	slightly	survey	union	works
another	bufalino	decisions	expect	great	known	memorable	plato	review	small	taken	unique	would
answer	built	deckard	experience	grown	knows	merkin	played	ridley	somehow	takes	unlike	wrapping
anthony	bullock	delivers	experiences	hannah	later	merry	plays	ridleys	someone	taking	using	writers
anyone	caleb	denby	explain	happen	latter	michael	pleasant	right	something	telling	usual	writing
anything	called	dench	faces	harrison	leading	might	pleasure	robert	sometimes	tells	usually	written
apparent	casting	depends	falls	harry	leads	million	plenty	rolls	somewhere	tendency	value	wrote
appearance	central	despite	familiar	harvey	learn	minds	point	running	sophisticate	terms	values	youre
appeared	certain	details	famous	hathaway	least	minutes	possible	russell	sought	thanks	varied	zemeckis
appearing	chain	developed	feeling	hauer	leave	mixed	potential	russo	sounds	thats	variety	
appears	character	dicks	feelings	hayward	leaves	moment	potter	saito	sourcedate	theater	version	
approved	characters	didnt	feels	heard	leaving	moments	powerful	sandra	speak	thematic	versus	
aragorn	check	different	fellowship	helps	lebowski	mostly	preston	saruman	speaking	theme	victoria	
archcriminal	choose	difficult	ferrer	highly	ledger	moves	prestons	sauron	special	thing	video	
around	chris	directed	figure	hilarious	level	movie	previous	saying	specializes	things	viewer	
asked	christian	director	filled	hoffa	lightsaber	movies	probably	scale	specifically	think	viewers	
asking	christopher	directors	filmmaker	hollywood	likely	moving	problem	scene	spectacular	thinking	viewing	
assist	cinematic	disney	filmmakers	hooper	lines	multiple	process	scenes	spectacularl	thinks	visions	
athleisurep	claims	document	films	hours	literal	murph	production	scorses	spent	third	visits	
audience	class	doesnt	final	however	literally	murphy	provide	scott	spielberg	thomas	visually	

Appendix 3. K-mean Cluster Python Output

```
Cluster 0:
 human
 robot
 technolog
 futur
 earth
Cluster 0 titles:
 MRR_Doc3_Cats-Review-They.docx,
 RRM_Doc2_Why-Minority-Report.docx,
 AB_Doc2_The_Matrix_Predicted.docx,
 PD_Doc2_AI_Gods_Egos_Ex-Machina.docx,
 AB_Doc1_The_Matrix_20.docx,
 SMN_Doc1_Back-Future-THR.docx,
 SA_DOC1_WhenBladeRunner.docx,
 PD_Doc3_Ex-Machina_AI_moviewithbrains.docx,
 AS_Doc1_Walle2.docx,
 SJB_Doc1_Architecture-of-the-Minds.docx,
 AS_Doc1_Walle3.docx,
 SA_DOC2_WhyBladeRunner.docx,
 SMN_Doc2_Back-Future-Reboot.docx,
 AB_Doc3_AI__What_Is_The_Matrix_.docx,
 MAS_Doc1_TheLordoftheRings_TheFellowshipoftheRing.docx,
 AS_Doc1_Walle1.docx,
 RRM_Doc1_Minority-Report-Predictions.docx,
 MRR_Doc2_Cats-Review-Will.docx,
 SMN_Doc3_From-Archives-LATimes.docx,
 MAS_Doc2_TheLordoftheRings_TheFellowshipoftheRing.docx,
 RRM_Doc3_Minority-Report.docx,

Cluster 1:
 space
 dream
 spacecraft
 hope
 violenc
Cluster 1 titles:
 SA_DOC3_BladeRunnerReview.docx,
 Visually_Stunning_Gravity_A_Movie_That_Changed_Movied_KFB.docx,
 ECC_Doc2_Dude-Bowls.docx,
 CMB_Doc1_Avengers_InfinityWar.docx,
 SL_Doc1_Off-to-the-Stars.docx,
 CMB_Doc2_Avengers_EndGame.docx,
 ECC_Doc1_El-Duderino.docx,
 MCL_Doc2_ReviewTheLatest.docx,
 The_Martian_Review_Matt_Damon_Shines_As_Stranded_Astronaut_KFB.docx,
```

```
 MCL_Doc1_StarWarsThe.docx,
 SJB_Doc3_This-Time-the-Dream_s.docx,
 CR_The_Dark_Knight_1.docx,
 CR_The_Dark_Knight_2.docx,
 SL_Doc3_Interstellar-Review.docx,
 CR_The_Dark_Knight_3.docx,
 SJB_Doc2_Dreams-On-Top.docx,
 Interstellar_Shows_The_Wonder_Of_Worlds_Beyond_KFB.docx,
 SL_Doc2_Interstellar-Review.docx,
 MCL_Doc2_ReviewTheLatest(1).docx,
 MCL_Doc3_StarWarsThe.docx,
 In_The_Irishman_BG.docx,

Cluster 2:
 woman
 money
 critic
 life
 kidskid
Cluster 2 titles:
 MRR_Doc1_Cats-Could-Have.docx,
 PD_Doc1_Ex-Machina_Fembot_Prob.docx,
 ECC_Doc3_Film-Critics-Blind.docx,
 NA_Doc3_Toy-Story-4-Existential-Terror.docx,
 CMB_Doc3_IronMan2.docx,
 CT_DOC2_CheckCashesIn.docx,
 CT_DOC1_BlankCheck.docx,
 The_Irishman_Mob_s_Greatest_Hits_BG.docx,
 CT_DOC3_BlankCheck.docx,
 NA_Doc1_Toy-Story-Trilogy-Epilogue.docx,
 NA_Doc2_Toy-Story-4-Escapes-Curse-Feminized-Sequel.docx,
 The_Irishman_Throwback_To_Scorsese_s_Golden_Age_BG.docx,
 MAS_Doc3_TheLordoftheRings_TheFellowshipoftheRing.docx,
```