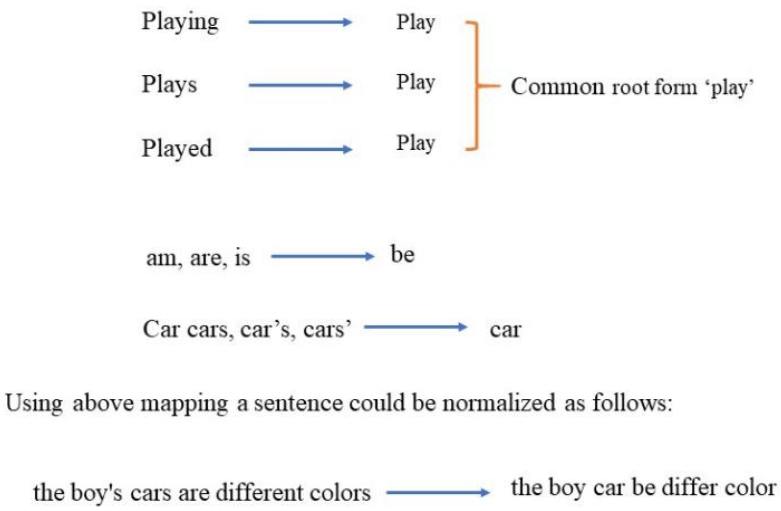
Alisher Siddikov
MSDS 453 - NLP
Assignment 1, 2, 3
5/16/2020

Part 1

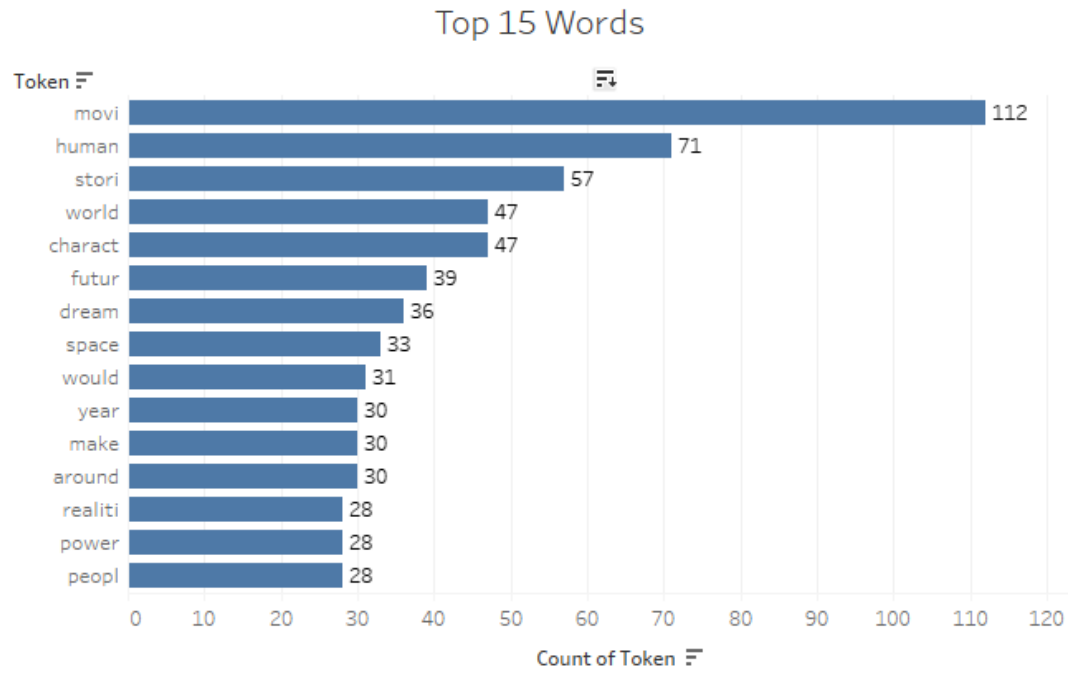
I started this text analytics assignment by splitting the text of a collection of movie reviews (or corpus) into the tokens (single words) and then cleaning them: removed punctuations, non-alphabetical (special characters and numbers), and short words (less than 4). I also removed stop words, the words that are not useful for our analysis, typically extremely common words such as “th”, “o”, “t”, and so forth in English. I also converted each token to a lower case and then normalized them by stemming. Stemming is the process of reducing inflection in words to their common root forms, such as mapping a group of words to the same stem even if the stem itself is not a valid word in English. Here is a DataCamp example of how the stemming works.



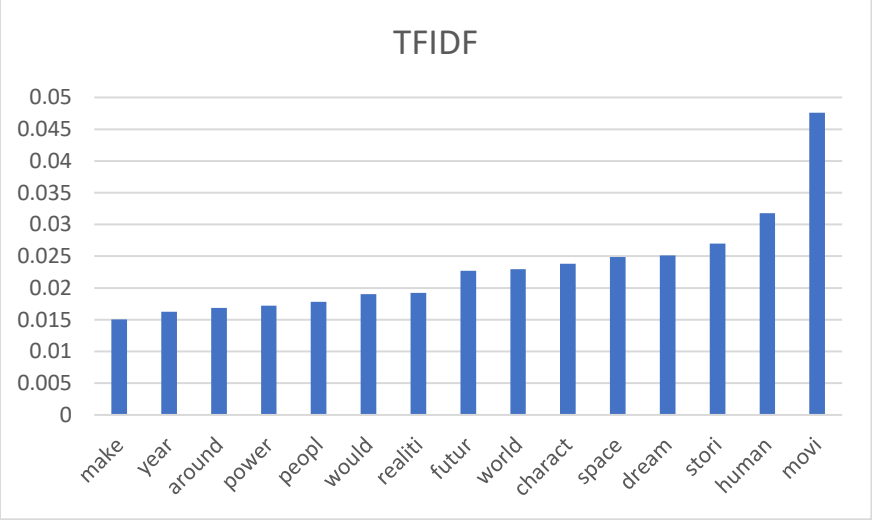
Using term frequency and inverse document frequency (TF-IDF) allowed me to find words that were characteristic for one movie review within a collection of movie reviews. TF means that the number of times a word appears in a document divided by the total number of words in the document. Every document has its own term frequency. IDF means that the log of the number of documents divided by

the number of documents that contain the word. Inverse data frequency determines the weight of rare words across all documents in the corpus. (Maklin, n.d.)

Here are the top fifteen words after we parse and clean the corpus. We can eliminate some of those



We can explore TFIDF score of the top fifteen words. We can see that words are less meaningful when the TFIDF average score is lower. As it gets higher, the word gets more meaningful.



Below are the k-means clustering results for TFIDF about the Wall-E movie. After comparing my three Wall-E movie reviews to the rest of the corpus, I was able to identify a single class that I believe characterizes my three documents, and that would be“robots” I can see that all three Wall-E reviews were clustered in the same cluster group along with Ex-Machina and Matrix movies, which contain similar themes and elements. Those movies were clustered more on robots, artificial intelligence, machines, and humanoid robots, and we can see those terms below. I believe that the El-Duderino movie does not fit this cluster.

```
 walle
 robot
 intellig
 caleb
 human
 nathan
 machina
 artifici
 machin
 suppli
Cluster 7 titles: PD_Doc1_Ex-Machina_Fembot_Prob.docx,
PD_Doc2_AI_Gods_Egos_Ex-Machina.docx, PD_Doc3_Ex-
Machina_AI_moviewithbrains.docx, AS_Doc1_Walle2.docx, AS_Doc1_Walle3.docx,
AS_Doc1_Walle1.docx, AB_Doc3_AI__What_Is_The_Matrix_.docx, ECC_Doc1_El-
Duderino.docx
```

Some of the terms that could fall within the“robot” class are“artificial intelligence," “machine,” and “humanoid.” Caleb Nathan is a person who meets the female robot, Ex-Machina. Therefore, Caleb, Nathan, and Machina are proper nouns. We can exclude them in the second round of our assignment. I believe the “robot” class meets the criteria for identifying classes because it is at an appropriate level of abstraction, and there is an adequate number of terms that can support the class.

However, Dec2Vec classification tells a different story than the TF-IDF classification. When I tried the Doc2Vec, I can see that cluster is not that meaningful. My three review documents are placed in two different clusters. The first cluster is more about robots, dreams, and sci-fi movies like Matrix, Ex-Machina, Avengers, Back to the Future, etc. The sixth cluster is about sci-fi movies, but it also includes

wars and crimes such as Star Wars and Irishman. I noticed that when I restarted the Python kernel, I got different clustering results for Doc2Vec even though k-means had a seed number set to 89.

```
1: ['SA_DOC3_BladeRunnerReview.docx',
  'CMB_Doc1_Avengers_InfinityWar.docx',
  'AB_Doc1_The_Matrix_20.docx',
  'SL_Doc1_Off-to-the-Stars.docx',
  'SMN_Doc1_Back-Future-THR.docx',
  'SA_DOC1_WhenBladeRunner.docx',
  'ECC_Doc3_Film-Critics-Blind.docx',
  'ECC_Doc1_El-Duderino.docx',
  'PD_Doc3_Ex-Machina_AI_moviewithbrains.docx',
  'CT_DOC2_CheckCashesIn.docx',
  'AS_Doc1_Walle2.docx',
  'AS_Doc1_Walle3.docx',
  'SA_DOC2_WhyBladeRunner.docx',
  'SJB_Doc3_This-Time-the-Dream_s.docx',
  'AB_Doc3_AI__What_Is_The_Matrix_.docx',
  'RRM_Doc1_Minority-Report-Predictions.docx',
  'NA_Doc1_Toy-Story-Trilogy-Epilogue.docx',
  'MAS_Doc3_TheLordoftheRings_TheFellowshipoftheRing.docx',
  'SMN_Doc3_From-Archives-LATimes.docx'],
6: ['AB_Doc2_The_Matrix_Predicted.docx',
  'NA_Doc3_Toy-Story-4-Existential-Terror.docx',
  'MAS_Doc1_TheLordoftheRings_TheFellowshipoftheRing.docx',
  'AS_Doc1_Walle1.docx',
  'CT_DOC3_BlankCheck.docx',
  'MRR_Doc2_Cats-Review-Will.docx',
  'MCL_Doc3_StarWarsThe.docx',
  'In_The_Irishman_BG.docx'],
```

I believe that Doc2Vec did not cluster well because we need a larger and higher quality dataset than the corpus we currently have. However, the clusters for TFIDF were adequate for my Wall-E review. When we compared the TFIDF and Doc2Vec cluster results, we can see that TFIDF clustered movies better. During our discussion exercise, the following suggestion emerged: TFI-IDF might be useful for weighing how important terms are to a document. At the same time, Doc2Vec is useful in diving deeper into the document and measuring similarities between sentences.

Part 2

In part 1, I started clustering the movies with similar topics. However, at that time, I did not include the customized stop words and equivalence class (EC). However, in part 2, after expanding my stop word list, building my EC, and changing the TFIDF k-means from 8 to 10, I was able to cluster similar movies

together better. Also, as you can see that most of the reviews of the same movies clustered together. All three reviews of the Wall-E movie clustered with all three reviews of Ex-Machina movies. I am thrilled about how my algorithm clustered the movie topics together.

```
Cluster 0:
 dream
 dream dream
 realiti dream
 world
 realiti
 human
 corpor
 layer
 rival
 skill
Cluster 0 titles:
 SJB_Doc1_Architecture-of-Minds,
 SJB_Doc3_This-Time-the-Dream_s,
 SJB_Doc2_Dreams-On-Top,
 Interstellar_Shows_KFB,

Cluster 1:
 technolog
 detect
 futur
 human
 realiti
 ai replicant
 world
 simul
 cruis
 money
Cluster 1 titles:
 AB_Doc2_The_Matrix_Predicted,
 AB_Doc1_The_Matrix_20,
 AB_Doc3_AI__What_Is_The_Matrix_,
 SA_DOC1_WhenBladeRunner,
 SA_DOC2_WhyBladeRunner,
 SA_DOC3_BladeRunnerReview,
 CT_DOC3_BlankCheck,
 RRM_Doc1_Minority-Report,
 RRM_Doc3_Minority-Report,
 RRM_Doc2_Why-Minority-Report,

Cluster 2:
 superhero
 space
 duel
 resist
 violenc
 swashbuckl adventur
 swashbuckl
 johnson
 violenc space
 forc
Cluster 2 titles:
 CMB_Doc1_Avengers_InfinityWar,
 MCL_Doc2_ReviewTheLatest,
 MCL_Doc2_ReviewTheLatest(1),
 MCL_Doc3_StarWarsThe,

Cluster 3:
 space
 spacecraft
 music
 human
 mission
 digit
 odyssey
 jellicl
 fantasi
 daughter
Cluster 3 titles:
 MRR_Doc1_Cats-Could-Have,
 MRR_Doc3_Cats-Review-They,
 MRR_Doc2_Cats-Review-Will,
 SL_Doc1_Off-to-the-Stars,
 SL_Doc3_Interstellar-Review,
 SL_Doc2_Interstellar-Review,
 Visually_Stunning_Gravity_KFB,
 The_Martian_Review_Matt_KFB,

Cluster 4:
 robot
 woman
 ai replicant
 human
 world
 fantasi
 gender
 spacecraft
 wast
 vulner
Cluster 4 titles:
 PD_Doc1_Ex-Machina_Fembot_Prob,
 PD_Doc2_AI_Gods_Egos_Ex-Machina,
 PD_Doc3_Ex-Machina_AI,
 AS_Doc1_Walle2,
 AS_Doc1_Walle3,
 AS_Doc1_Walle1,

Cluster 5:
 scorses
```

```
 superhero
 crime
 violenc
 frank
 power
 gangster
 mobster
 jimmi
 goodfella
Cluster 5 titles:
 The_Irishman_Mob_s_Greatest_BG,
 The_Irishman_Throwback_To_BG,
 In_The_Irishman_BG,
 CR_The_Dark_Knight_3,

Cluster 6:
 futur
 reboot
 human
 kids
 money
 mcfli
 excon
 futur reboot
 delorean
 glover
Cluster 6 titles:
 SMN_Doc1_Back-Future-THR,
 SMN_Doc2_Back-Future-Reboot,
 SMN_Doc3_From-Archives-LATimes,
 CT_DOC1_BlankCheck,
 CT_DOC2_CheckCashesIn,

Cluster 7:
 superhero
 wizard
 villain
 quest
 action
 imagin
 comedi
 human
 tolkien
 wizard wizard
Cluster 7 titles:
 CMB_Doc2_Avengers_EndGame,
 ECC_Doc3_Film-Critics-Blind,
 MCL_Doc1_StarWarsThe,
 CR_The_Dark_Knight_1,
 CR_The_Dark_Knight_2,
 MAS_Doc1_TheLordoftheRings,
 MAS_Doc3_TheLordoftheRings,
 MAS_Doc2_TheLordoftheRings,

Cluster 8:
 bowl
 privat
 shaggi
 polit
 russian
 brilliant
 kidnap
 malibu
 seattl
 millionair
Cluster 8 titles:
 ECC_Doc2_Dude-Bowls,
 ECC_Doc1_El-Duderino,
 CMB_Doc3_IronMan2,

Cluster 9:
 toys
 toys toys
 woman
 bonni
 woodi
 rescu
 andi
 existenti
 concept
 antiqu
Cluster 9 titles:
 NA_Doc3_Toy-Story-4-Existential,
 NA_Doc1_Toy-Story-Trilogy,
 NA_Doc2_Toy-Story-4-Escapes-Curse,
```

Let’s focus on the cluster number four, where the Wall-E movie reviews were clustered with Ex-Machina. This cluster was based on fantasy, vulnerability, robots, artificial intelligence, replicants, female gender, humans, waste compactor, wasteland, earth, and spacecraft. And most of the terms were derived by EC. Here is the full list of ECs terms, and I highlighted in red the ones which used in the clustering Wall-E movie.

- **'world'**: ['home planet', 'earth'],
- **'space'**: ['planets', 'planet', 'universe', 'galaxy', 'spacewalking', 'alien', 'moon', 'mars'],
- **'spacecraft'**: ['axiom', 'interstellar', 'space station', 'satellite'],
- **'robot'**: ['robots', 'walles', 'walle', 'wally', 'eve', 'machina', 'conscious machine', 'humanlike creature', 'artificial human', 'intelligent creature', 'machine', 'machines'],
- **'ai replicant'**: ['artificial intelligence', 'artificially intelligent', **'humanlike intelligence'**, 'artificial intelligent', 'intelligence', 'replicant', 'humanoid', 'android', 'genetically', 'superiority', 'sentient'],
- **'technology'**: ['technologies', 'algorithm', 'ripley', 'matrix', 'computer', 'tech', 'optical recognition', 'gadget'],
- **'human'**: ['humanity', 'humans', 'people', 'family', 'father', 'person', 'babies'],
- **'woman'**: ['women', 'female', 'girls'],
- **'kids'**: ['teenager', 'teenage', 'teen', 'young', 'children', 'child'],
- **'detective'**: ['blade runner', 'minority report'],
- **'superhero'**: ['batman', 'star wars', 'hobbits', 'lord of the rings', 'lord rings', 'irishman', 'avengers'],
- **'villain'**: ['joker'],
- **'toys'**: ['woody', 'forky', 'spork', 'bonnie'],
- **'money'**: ['fund', 'bank', 'banks'],
- **'crime'**: ['criminal', 'organized crime', 'mafia', 'mob'],
- **'violence'**: ['destroying', 'destroyed', 'destroyers', 'destroy', 'fighting', 'fights', 'fight', 'revenge', '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'],
- **'wizard'**: ['gandalf'],
- **'fantasy'**: ['science fiction', 'scifi', 'fantastical', 'fantastic']

Also, adding the proper nouns and less meaningful words (verbs adjectives, and adverbs) to the stop words list helped to improve the topic clustering dramatically. I spent most of my time to refine the stop word and EC lists. Please see the stop words in the appendices section. There are about 640 stop words. Per further discussions and comments, I will refine this project. For now, I am satisfied with the results.

Bibliography

Maklin, C. (n.d.). *TF IDF | TFIDF Python Example*. Retrieved from Towards Data Science: https://towardsdatascience.com/natural-language-processing-feature-engineering-using-tf-idf-e8b9d00e7e76

Appendices:

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	unexpectedl	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	