The Effect of Proper Nouns on K-Means Clustering: A Peek Into TF-IDF Sensitivity

Christopher J. Rico

Northwestern University

Introduction and Problem Statement

Term frequency–inverse document frequency, or TF-IDF, is a numerical statistic often used in natural language processing. TF-IDF scores are intended to reflect how important a word is to a particular document in a corpus. These scores can then be used to create “clusters” of documents that are most alike using a machine learning modeling method called K-Means clustering.

The corpus in question consists of film reviews covering a broad swath of films. Within a corpus this broad, the common important terms among documents that review the same film (or same film series) are likely to be the characters, actors, director, and the film title itself. However, this presents a problem: we are interested in employing NLP on these documents to understand how they relate and differ from one another on a level deeper than this. If we wish to do more than simply understand which documents mention the same people, it may be of value to examine the role that proper nouns play on TF-IDF term extraction and scoring.

My analysis attempts to evaluate the effect of proper nouns on TF-IDF k-means document clustering. Rather simply, I intend to perform k-means document clustering with various *k*-values on the corpus, both with and without proper nouns included during term extraction and document vectorization.

Dataset

The dataset consists of a single corpus of 58 documents, each contributed by a student in the course. Within the corpus, each document is a film review written in English, truncated to roughly 500 words by the student who submitted it.

Manipulation of the corpus was performed in Google Colaboratory (a cloud-based Jupyter notebook) running a Python 3 interpreter. A variety of Python functions and NLP libraries were used to prepare the data for term extraction and document vectorization. Each document was first tokenized and stripped of punctuation and non-alphabetical characters. Next, tokens shorter than 4 characters were removed, as well as any stopwords included in the Natural Language Toolkit (NLTK) English stopwords list. Finally, all characters were converted to lowercase.

Research Design and Methods

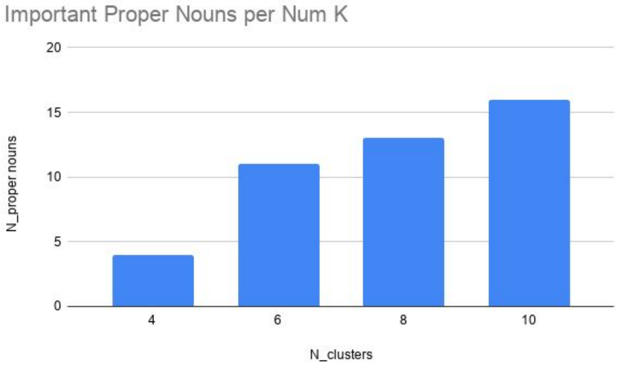
This experiment was run using the provided code in the environment described above. Various industry-standard NLP libraries were used to perform the term extraction and document vectorization: SciKit-Learn's TF-IDF model. Computing k-means clusters was performed using SciKit-Learn's Kmeans library.

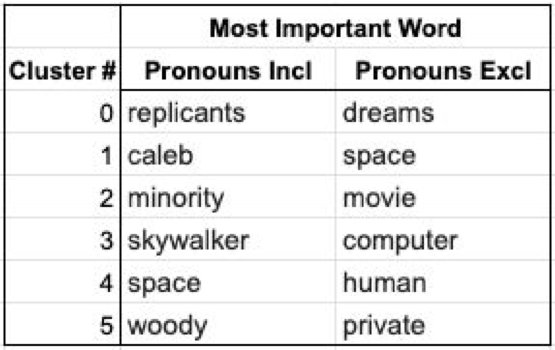
To compare the effects of proper nouns on term extraction and vectorization, this experiment employs a 4x4 fully-crossed design. On the documents in the corpus, clustering will be performed using *k*-values of 4, 6, 8, and 10 on TF-IDF results with and without proper nouns, as well as Doc2Vec results with and without proper nouns. These values were chosen after examining preliminary results that showed clusters with $3 > k < 10$ had groupings that were either so broad or so specific as to be without meaning. Clustering results were then compared to one another to gain qualitative insight into how proper noun exclusion affects the clustering of the corpus.

Proper nouns were discovered iteratively by performing k-means TF-IDF extraction/clustering on the corpus, and simply looking at the 'most important' terms within each cluster. Selected proper nouns were then added to an auxiliary stopwords list (see appendix), and removed during corpus pre-processing. In general, TF-IDF extraction on this corpus tended to favor proper nouns as the most important terms, so this was an effective way to remove those which had the most influence on term extraction and (presumably) document vectorization.

Results

Change in values of *k* had major effects on the "most important words" in each cluster of TF_IDF results. In the TF-IDF groups which included proper nouns, as *k* increased, the number of proper nouns included in important words did as well. In both TF-IDF groups, increased *k* caused an increase in the uniqueness of important words.





	Most Important Word	
Cluster #	Pronouns Incl	Pronouns Excl
0	replicants	dreams
1	caleb	space
2	minority	movie
3	skywalker	computer
4	space	human
5	woody	private

Change in the values of *k* had major effects on the composition of each cluster. When proper nouns were included, higher values of *k* led to documents that reviewed the same film (or films in the same series) to be clustered together. With proper nouns excluded, document clustering was less cohesive with regards to film and film series, regardless of *k* values.

Analysis

With proper nouns included in the corpus documents, TF-IDF had a tendency to extract and rank the most important terms as primarily proper nouns. Some examples of important terms extracted include “walle”, “skywalker”, “batman”, “joker”, “woody”, “matrix”, and “lebowski”.

By excluding proper nouns, TF-IDF was forced to rank more frequent terms as important. Indirectly, this had the effect of removing the k-clustering algorithm’s ability to group films primarily based on the people involved in them, or the film titles mentioned in them. Without these highly specific terms, clustering decisions were made increasingly by using terms important to film topic, such as “action”, “dreams”, “space”, “terrorists”, “reality”, “crime”, and “intelligence”. However, some terms with less discriminatory ability also became ranked as more important: “film”, “along”, “still”, and “point”. That said, these less-discriminative terms were still in lower importance positions than most of the terms related to film topics.

Removing proper nouns from the corpus could be an effective way to ensure that document vectorization, as well as any subsequent attempts to employ classification, are unaffected by the names of the film and people involved. This could help a classifier better generalize to film reviews it has not been exposed to before: because the classifier will be trained using terms and document vectors more related to film *topic*, it won’t be caught up trying to classify documents based on *people*.

Design and Implementation Considerations

Initially, my plan was to compare clustering ability across different *n* terms extracted. However, after testing a variety of *n* terms in the range 50-150, it was clear that this number had no effect whatsoever on the clustering results. Realistically, this makes sense: both TFIDF and

Doc2Vec algorithms are designed to extract the most explanatory terms and vectorize documents using the most explanatory terms, so the term values and vector representations of documents in the corpus are unlikely to change much as the number of allowed terms is reduced. As k-means clustering relies on these term and vector values to perform its partitioning, it would hardly be affected by a reduction in less-explanatory terms. With this in mind, I settled on using the default value of 100 terms for both TF-IDF term extraction.

Conclusions

In conclusion, I was able to qualitatively observe the effect of excluding proper nouns on TF-IDF term extraction results within our corpus. The effect of excluding proper nouns was examined qualitatively by comparing k-means clustering decisions. Exploration of these effects allowed for a better understanding of the TF-IDF extraction process, and has given greater insight into which terms may be most valuable for inclusion in a vector for clustering and classification.

In the future, it would be prudent to develop more programmatic quantitative methods for comparing variable effects. Put simply, I was not able to program in a way to access the TF-IDF values or the document vector values so that the effect of proper nouns on TF-IDF scores could be quantified concretely.

Appendix

Proper nouns removed from documents:

abrams	cobbs	hannah	knight	nolan	scotts
avengers	cooper	harrison	lebowski	nolans	skywalker
batman	deckard	harvey	ledger	pacino	spielberg
blade	denby	hauer	machina	palpatine	stark
blade runner	eckhart	hayward	martian	pesci	tolkien
blank	ex machina	hobbits	marty	preston	tyrell
blank check	fellowship	hoffa	marvel	quigley	victoria
bonnie	forky	hooper	matrix	replicants	walle
bonsall	frank	inception	mcconaughey	ridley	wargames
brand	frodo	infinity	mcfly	rings	woody
bridges	gandalf	interstellar	merkin	runner	zemeckis
caleb	garland	irishman	michael	russel	
check	george	jackson	minority	russell	
christian	gordon	jeff	minority report	scorsese	
christopher	gotham	joker	nathan	scott	

“Important Words” Per Cluster With Proper Nouns Included

N_Clusters	Cluster	IW 1	IW 2	IW 3	IW 4
4	0	check	preston	blank	blank check
4	1	walle	space	human	would
4	2	woody	movie	story	fellowship
4	3	lebowski	movie	critics	bowling
6	0	replicants	blade	blade runner	runner
6	1	caleb	lebowski	would	reality
6	2	minority	matrix	report	minority report
6	3	skywalker	abrams	force	trilogy
6	4	space	interstellar	batman	nolan
6	5	woody	forky	frank	hobbits

8	0	skywalker	abrams	movie	trilogy
8	1	fellowship	frodo	frank	hobbits
8	2	dreams	replicants	inception	blade
8	3	walle	space	interstellar	future
8	4	matrix	reality	intelligence	artificial intelligence
8	5	batman	joker	lebowski	movie
8	6	check	preston	blank	blank check
8	7	woody	forky	story	bonnie
10	0	skywalker	abrams	trilogy	force
10	1	fellowship	frodo	hobbits	hooper
10	2	dreams	replicants	inception	blade
10	3	space	interstellar	future	mcfly
10	4	matrix	reality	intelligence	artificial intelligence
10	5	batman	joker	lebowski	movie
10	6	check	preston	blank	blank check
10	7	woody	forky	story	bonnie
10	8	frank	irishman	pesci	pacino
10	9	walle	robots	earth	though

“Important Words” Per Cluster With Proper Nouns Excluded

N_Clusters	**Cluster**	**IW 1**	**IW 2**	**IW 3**	**IW 4**
4	0	computer	movie	terrorists	games
4	1	movie	story	dreams	characters
4	2	space	earth	planet	robots
4	3	intelligence	human	reality	artificial
6	0	dreams	reality	dream	without
6	1	space	earth	movie	black
6	2	movie	force	trilogy	franchise
6	3	computer	movie	story	report
6	4	human	intelligence	female	future
6	5	private	excon	instead	clown

8	0	funny	movie	cosmic	excon
8	1	action	movie	wizard	characters
8	2	movie	crime	plenty	point
8	3	story	movie	franchise	trilogy
8	4	dreams	dream	minds	ideas
8	5	space	earth	mission	human
8	6	intelligence	future	female	would
8	7	bowling	inspired	shaggy	private
10	0	cosmic	funny	movie	enough
10	1	action	movie	wizard	characters
10	2	movie	power	crime	violence
10	3	story	along	lightsaber	franchise
10	4	dreams	dream	minds	ideas
10	5	space	earth	movie	mission
10	6	intelligence	female	human	reality
10	7	bowling	inspired	shaggy	private
10	8	movie	computer	report	still
10	9	replicant	directors	versus	version