Discovering Patterns: Clusters & Topics

On Netflix Data

Harry Xiong

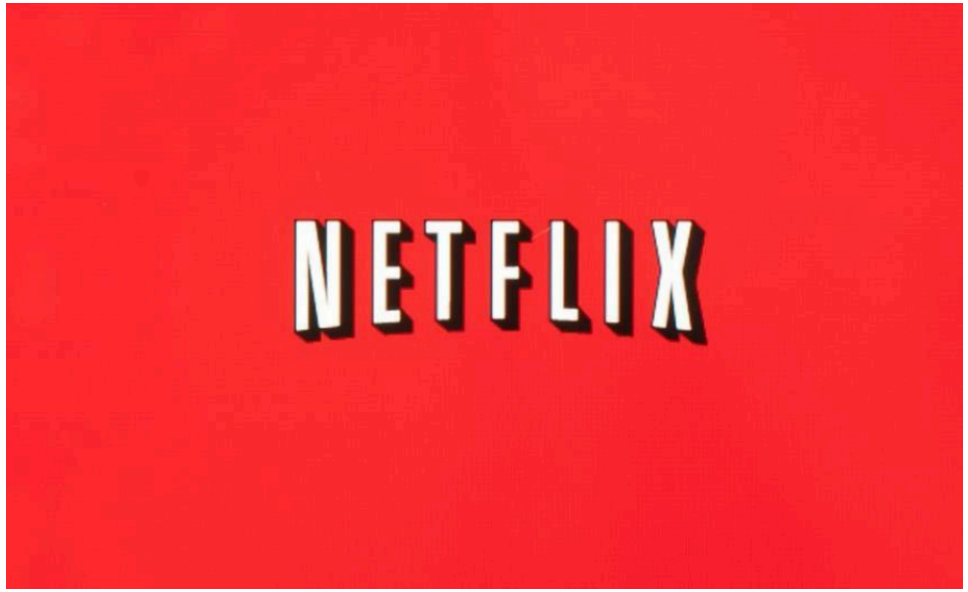
Goal: Discover Meaningful Patterns in Text Data

- Clusters of Words, Topics and Documents

Methods: Unsupervised Learning Models

- K-Means Clustering (Flat)
- Silhouette Method
- Ward's Method (Hierarchical)
- Latent Dirichlet Allocation (LDA)

Intro | **Data** | Flat Clustering | Hierarchical Clustering | Topic Modeling | Network

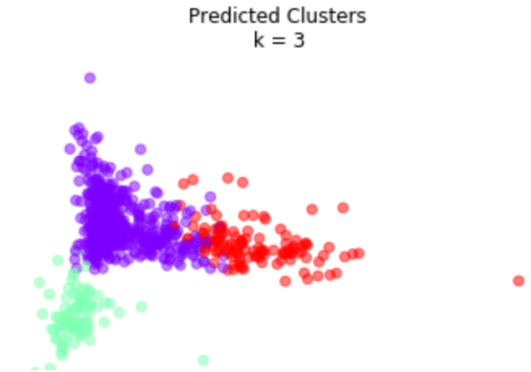


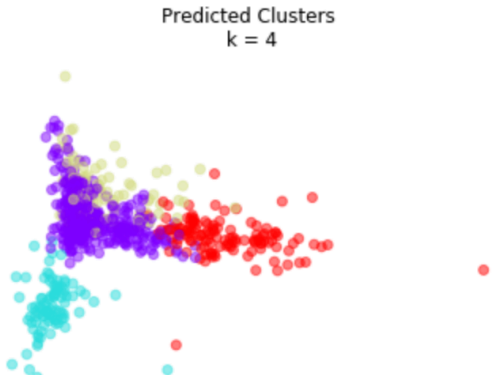
	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	81145628	Movie	Norm of the North: King Sized Adventure	Richard Finn, Tim Maltby	Alan Marriott, Andrew Toth, Brian Dobson, Cole...	United States, India, South Korea, China	September 9, 2019	2019	TV-PG	90 min	Children & Family Movies, Comedies	Before planning an awesome wedding for his gra...
1	80117401	Movie	Jandino: Whatever it Takes	NaN	Jandino Asporaat	United Kingdom	September 9, 2016	2016	TV-MA	94 min	Stand-Up Comedy	Jandino Asporaat riffs on the challenges of ra...
2	70234439	TV Show	Transformers Prime	NaN	Peter Cullen, Sumalee Montano, Frank Welker, J...	United States	September 8, 2018	2013	TV-Y7-FV	1 Season	Kids' TV	With the help of three human allies, the Autob...
3	80058654	TV Show	Transformers: Robots in Disguise	NaN	Will Friedle, Darren Criss, Constance Zimmer, ...	United States	September 8, 2018	2016	TV-Y7	1 Season	Kids' TV	When a prison ship crash unleashes hundreds of...
4	80125979	Movie	#realityhigh	Fernando Lebrija	Nesta Cooper, Kate Walsh, John Michael Higgins...	United States	September 8, 2017	2017	TV-14	99 min	Comedies	When nerdy high schooler Dani finally attracts...

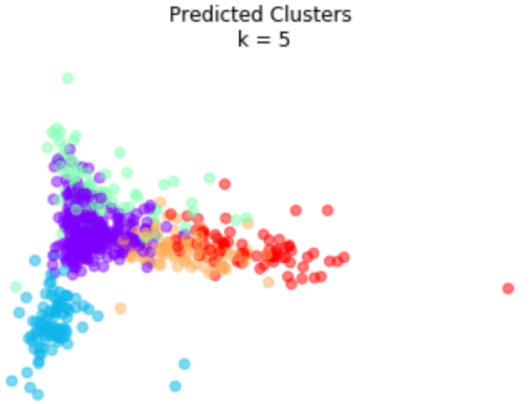
- 6234 TV shows or Movies
- Description as text data
- Listed_in as labels
- Selected 3 most common labels: Stand-Up Comedy, Documentary and Dramas (820 rows)

https://www.kaggle.com/shivamb/netflix-shows

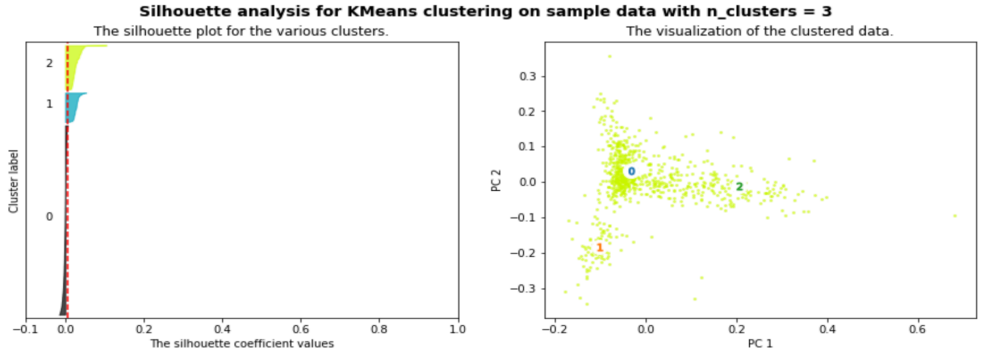
- "Simple" Movie from Netflix
- 12 Characters, 6 Main Ones
- Analyze Script
- Topic Models & Network Analysis



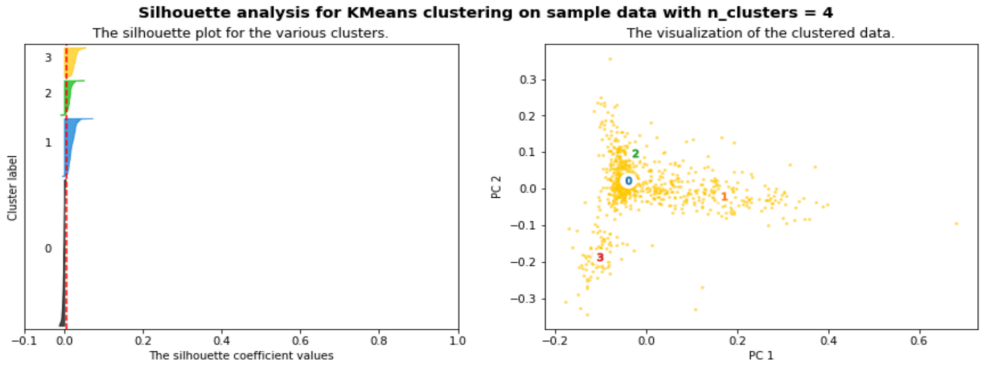




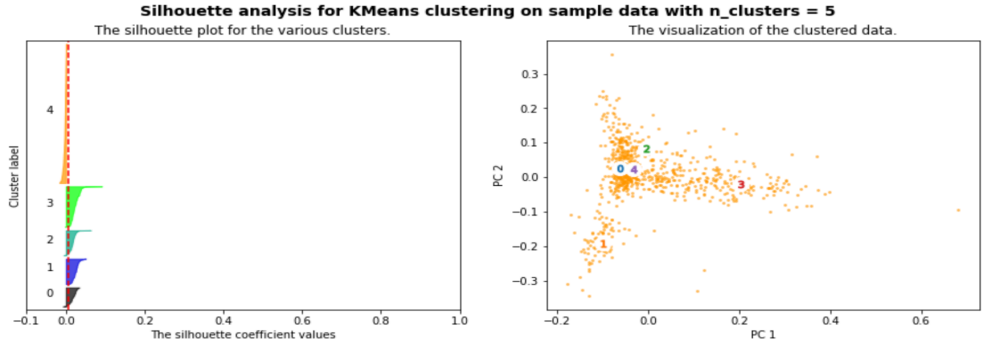
- Pretend we don't know the real number of clusters
- We tried k = 3, 4, 5
- Visually we can see k = 4, 5 have less distinguishable borders
- Compare with the "true" plot
- Good job on picking out the bottom left part (documentary)



For n_clusters = 3, The average silhouette_score is : 0.007



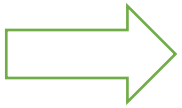
For n_clusters = 4, The average silhouette_score is : 0.008



For n_clusters = 5, The average silhouette_score is : 0.008

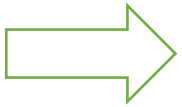
- The silhouette score is a measure of cohesion and separation
- It ranges from -1 to 1 where higher value suggests object being better matched with own clusters and poorly matched with other clusters
- The scores are pretty similar for k = 3, 4, 5
- K = 4, 5 even perform better than k = 3

```
Cluster 0:
 life
 young
 man
 family
 comedian
 new
 woman
 world
 comic
 film
```



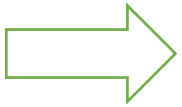
Drama

```
Cluster 1:
 documentary
 follows
 explores
 examines
 american
 world
 chronicles
 family
 high
 people
```



Documentary

```
Cluster 2:
 stand
 special
 comedy
 comedian
 comic
 life
 jokes
 takes
 delivers
 shares
```

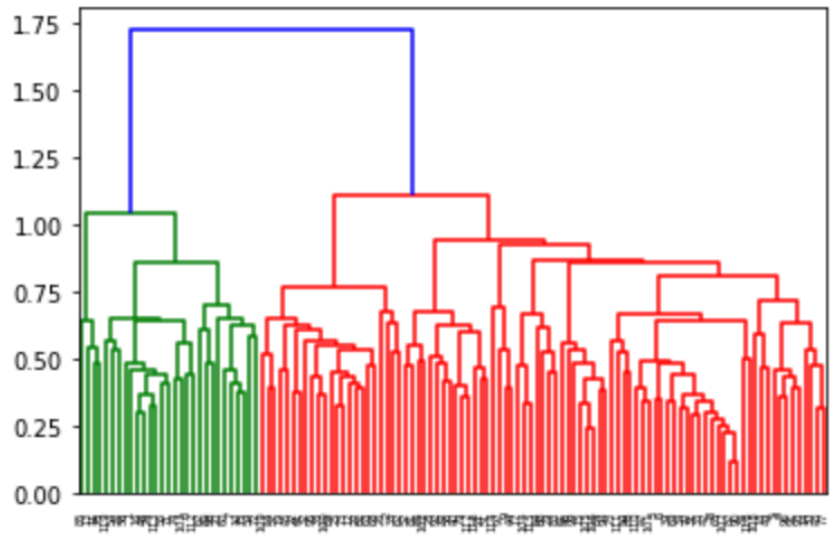


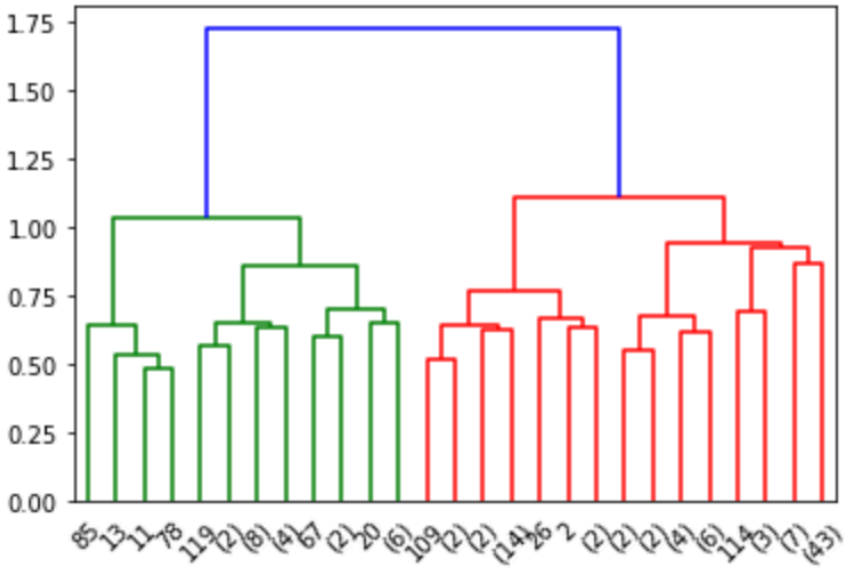
Stand-up Comedy

```
For our clusters:
Homogeneity: 0.272
Completeness: 0.389
V-measure: 0.320
Adjusted Rand Score: 0.120
```

- From cluster top terms, we get clear bags of words for documentaries and stand-up comedies and we can make a decent guess for dramas

- Homogeneity score: the degree that all of its clusters contain only data points which are members of a single class
- Completeness score: the degree that all data points of a given class are also elements of the same cluster.
- V-measure: the harmonic mean of *Homogeneity* and *Completeness*
- Adjusted Rand Index: another measure of prediction accuracy





- Hierarchical clustering gives nested clusters at any resolution
- At the finest resolution, every document is its own cluster
- Top-left graph visualizes the dendrogram of 40 Netflix movies/shows per category
- Bottom-left graph restricts the number of visible branches to 4
- Only able to identify 2

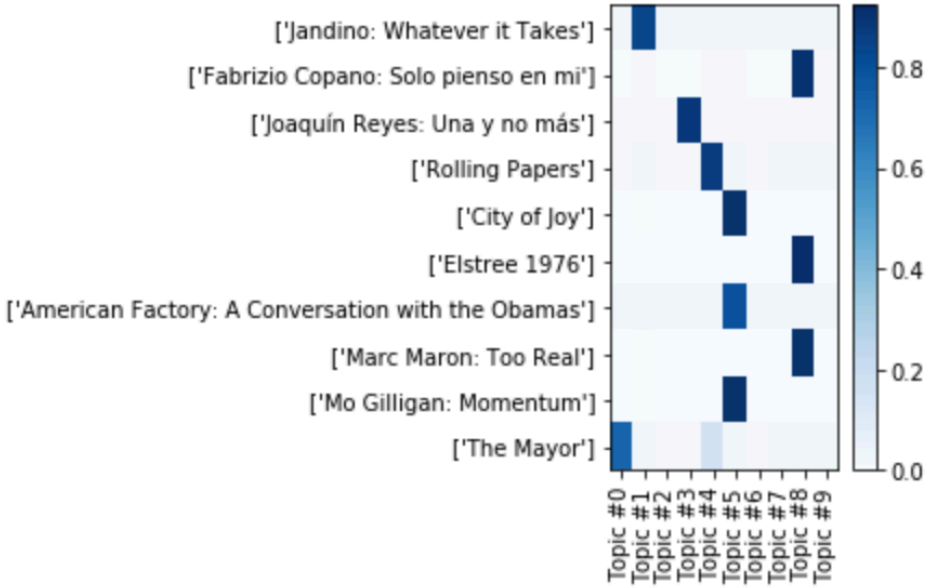
```
For our complete clusters:
Homogeneity: 0.397
Completeness: 0.423
V-measure: 0.409
Adjusted Rand Score: 0.261
```

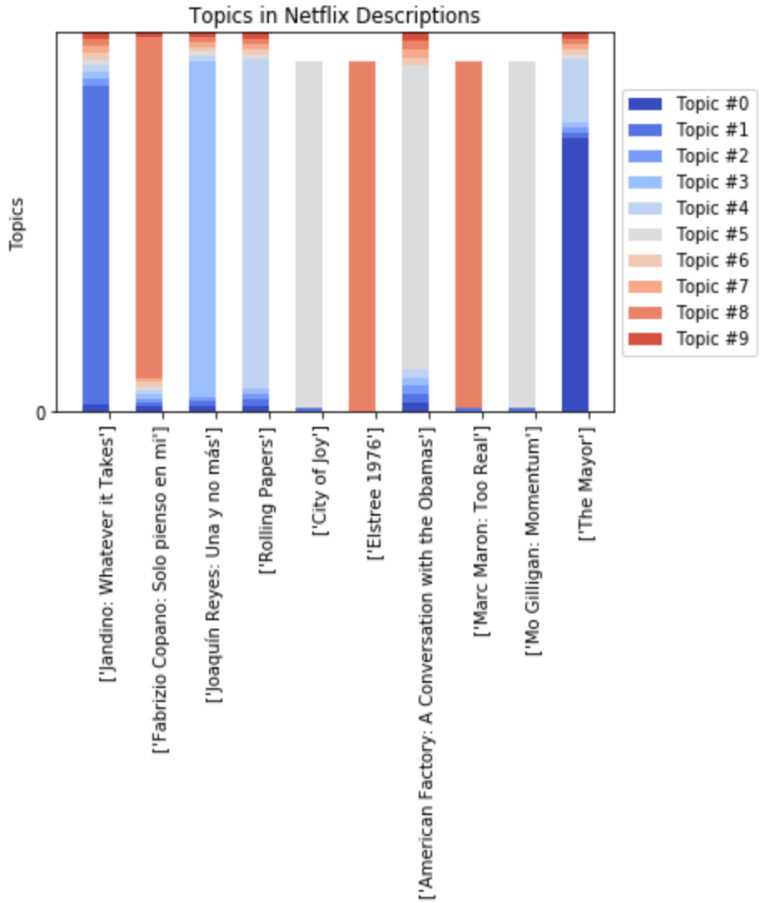
- The metrics look better than flat clustering all around!
- Homogeneity: 0.397 > 0.272
- Completeness: 0.423 > 0.389
- V-measure: 0.409 > 0.320

- Used Gensim
- Based on LDA
- # of topics = 10
- Many repeated words, but each topic is still different

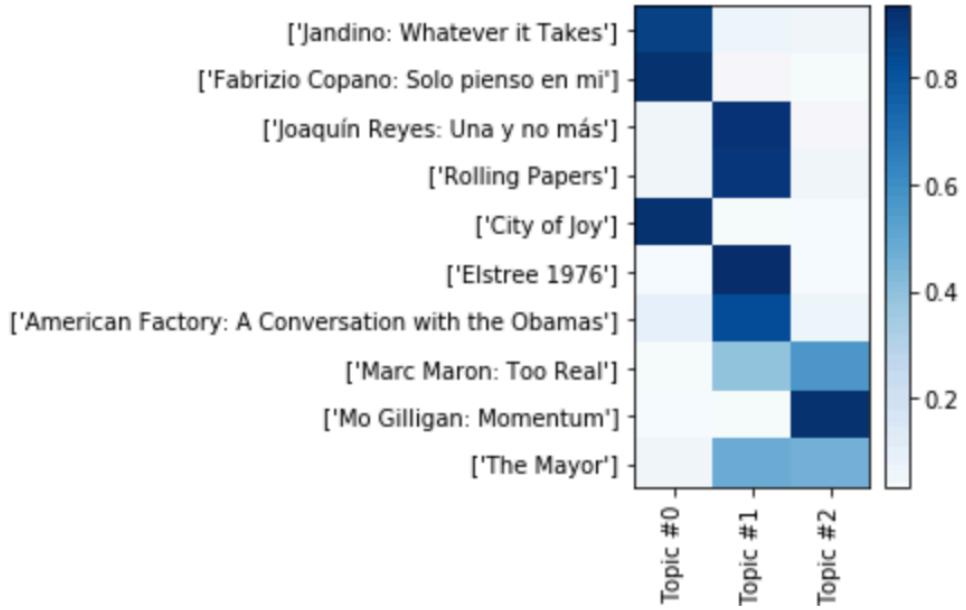
	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Topic_6	Topic_7	Topic_8	Topic_9
0	family	documentary	documentary	comedian	special	stand	woman	life	family	stand
1	special	old	live	life	stand	man	young	live	comedian	comic
2	stand	stand	career	young	comic	special	live	world	life	comedian
3	comic	change	stage	love	comedy	comedian	life	father	set	special
4	turn	life	stand	woman	life	new	stand	home	stand	life
5	man	sex	man	friend	new	life	comedian	story	film	film
6	share	comedian	death	marriage	relationship	documentary	man	return	comic	comedy
7	documentary	live	come	struggle	comedian	family	mother	comedy	documentary	star
8	struggle	world	offer	world	year	mother	husband	man	man	filmmaker
9	stage	man	group	documentary	woman	drama	form	young	history	family

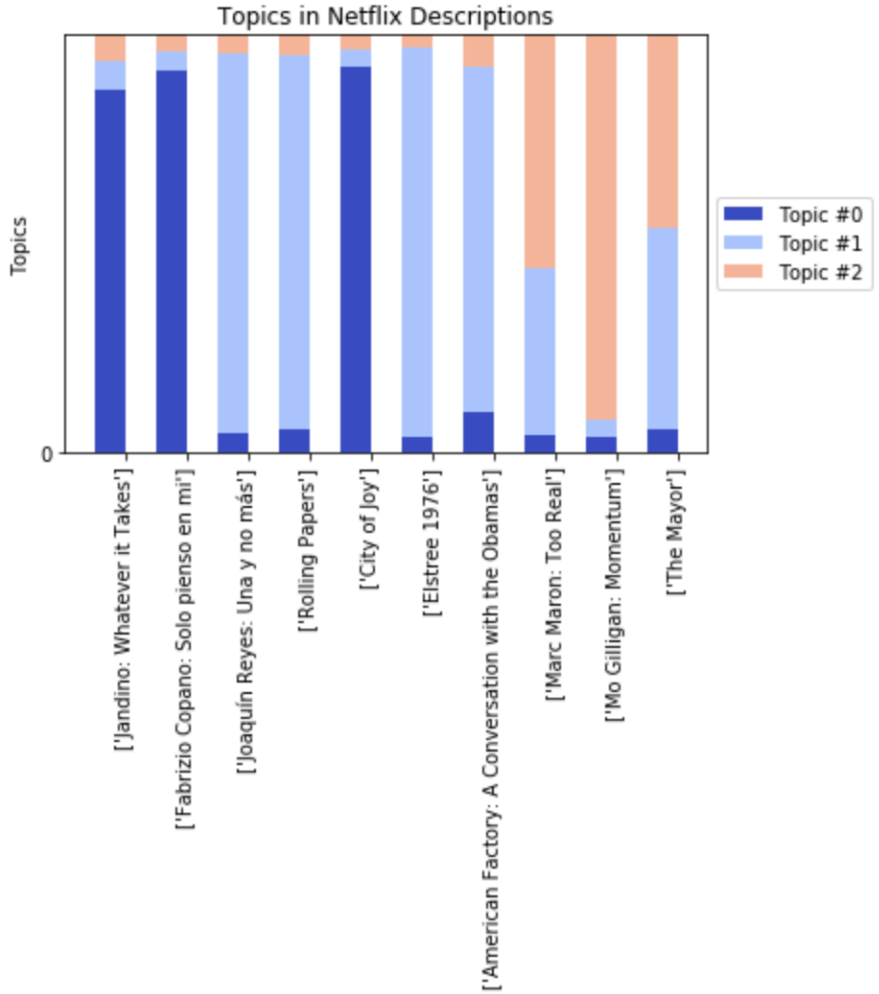
- Used Gensim
- Based on LDA
- # of topics = 10
- Many repeated words, but each topic is still different
- The stacked bar and heat map illustrate that most shows only have one major topic





- Used Gensim
- Based on LDA
- # of topics = 3
- Very similar to the flat clustering results
- Relatively good results too, so 10 topics can be overfitting

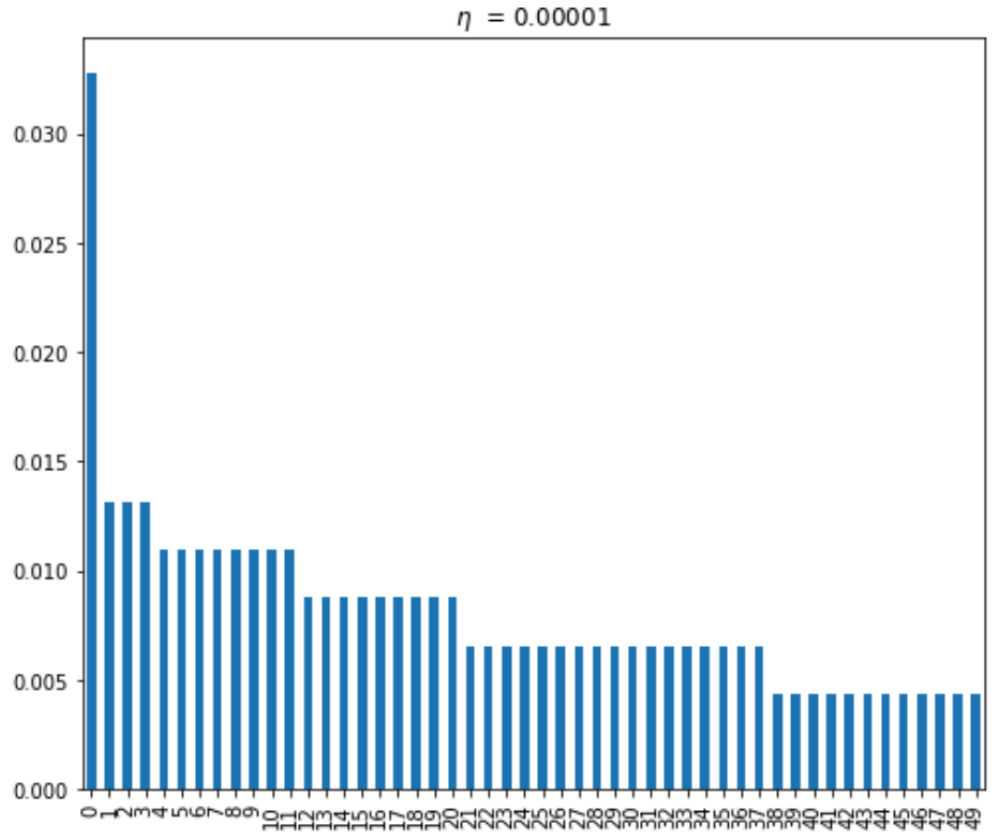


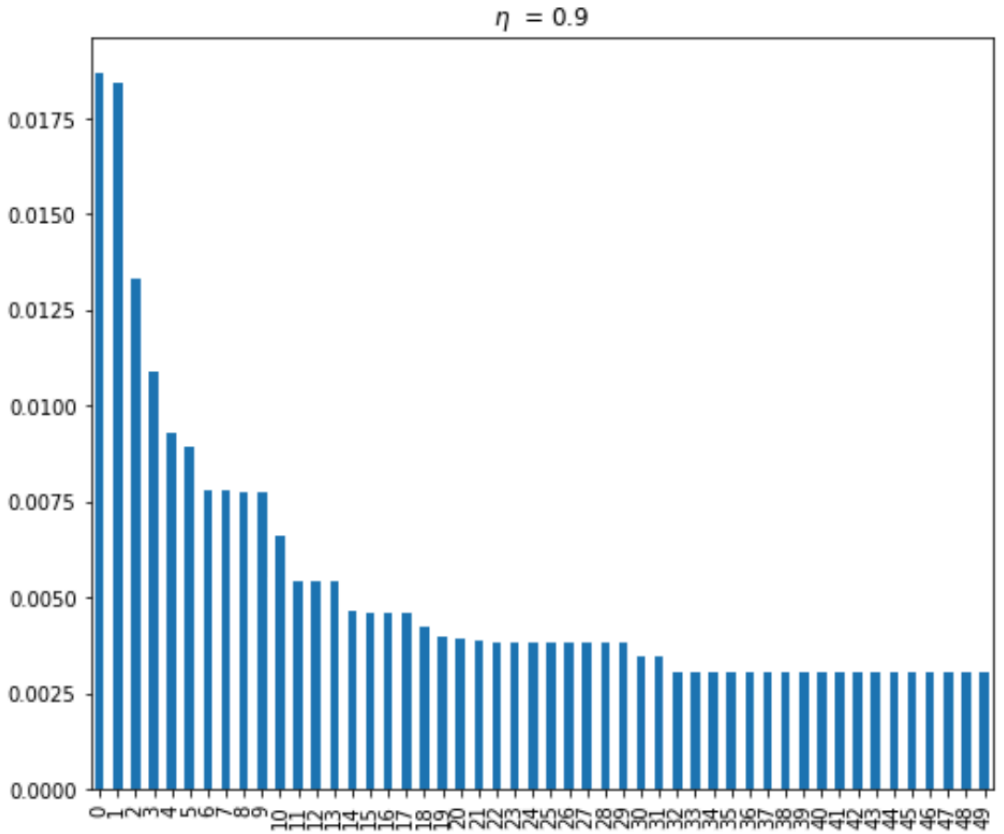


- Used Gensim
- Based on LDA
- # of topics = 10
- Some hyperparameters can be tuned

- α: the sparsity of document-topic loadings
- η: the sparsity of topic-word loadings
- # of topics

Probability Distribution of Words, Topic 1



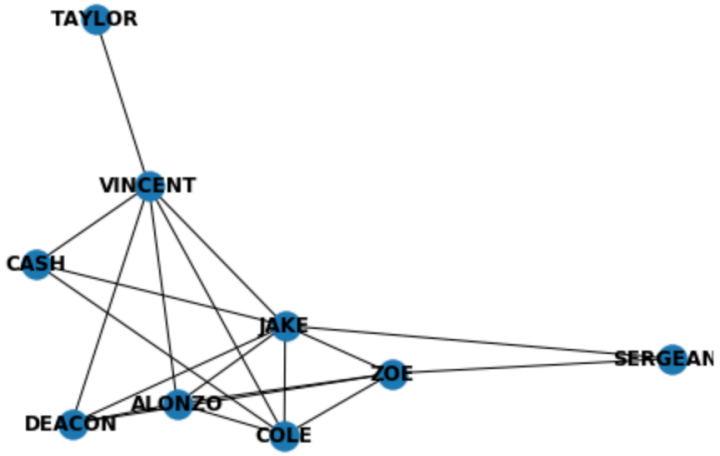


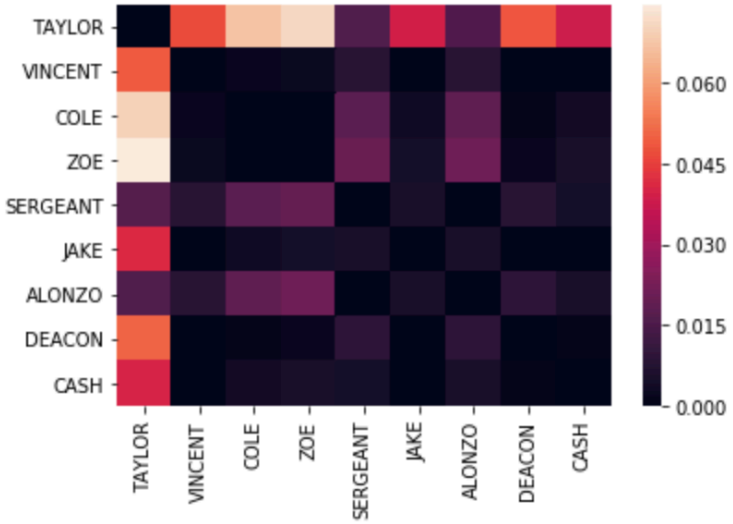
- Topic modeling can lead to more complex analysis such as networks
- Although movie descriptions do not interact with each other, we can dive into a particular movie and investigate its character interactions
- The Marine 5: Battleground

```
ALONZO Screw Vincent I 'm running this I 'll call him when it 's done

SERGEANT Captain ai n't gon na like this Shit

DEACON That 's from the truck that did the hit
```

- Analyze text by parsing out the lines by characters and to whom are the lines directed to
- Graph the network
- Topic modeling
- KL divergence of characters' topics

- Topic modeling can lead to more complex analysis such as networks
- Although movie descriptions do not interact with each other, we can dive into a particular movie and investigate its character interactions
- The Marine 5: Battleground





- Analyze text by parsing out the lines by characters and to whom are the lines directed to
- Graph the network
- Topic modeling
- KL divergence of characters' topics