Clustering Wikipedia Articles

Lane Aasen

Department of Computer Science University of Washington Seattle, WA 98105 aaasen@cs.washington.edu

Abstract

Clustering Wikipedia articles using K-Means. Exploration into K-Means including choice of K, K-Means++, the effect of random seeds, and tf-idf thresholds.

1 Dataset

The provided dataset contains 15,903 Wikipedia articles in term frequency inverse document frequency (tf-idf) format. Most of the articles are in English, but some are not. Sorting the documents by language would be an interesting project in itself! There are 10,574 unique words in this dataset. Each document is represented as a sparse vector with one dimension for each word. Stop words have been removed from the dataset, but uncommon words remain.

2 K-Means Clustering

My first goal was to implement a basic K-Means clustering algorithm from scratch. The code is available at https://github.com/aaasen/wiki-cluster.

2.1 Choosing K

2.1.1 Minimizing Distortion

Given K clusters $C_1, C_2, ..., C_K$ where each cluster is a set of document vectors and μ_i is the centroid of C_i , the total distortion is defined as follows:

$$\sum_{i=1}^{K} \sum_{d \in C_i} ||d - \mu_i||^2$$

To minimize the distortion, we could set K equal to the number of documents, but then the clusters would be meaningless. We want to choose a K with low distortion that also results in interpretable clusters. Figure 1 shows a plot of K versus total distortion. When $1 \le K \le 16$, adding additional clusters has a large impact on the distortion, but once K > 16, adding additional clusters has little impact on the distortion. From this alone, it makes sense to set K = 16 since it provides a good balance of distortion and interpretability.

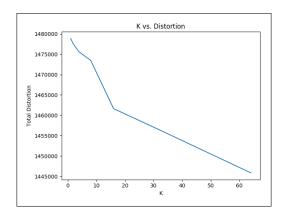


Figure 1: K versus total distortion for $K \in \{1, 2, 4, ..., 256\}$

2.2 K and Cluster Size

As K increases, the clusters become more sparse. Once K=256, over half of the clusters have only one document, and are essentially useless. When K=16, the median cluster size is 8.5, and the cluster sizes are as follows:

$$[10061, 3013, 1128, 909, 707, 30, 23, 13, 4, 4, 3, 2, 2, 2, 1, 1]$$

Over half of the clusters are very small, and one of the clusters is too large to be interpretable. This indicates that the data has significant outliers and may lack a structure conducive to clustering.

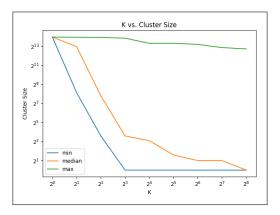


Figure 2: K versus minimum, median, and maximum cluster size for $K \in \{1, 2, 4, ..., 256\}$ with a log_2 scale on both axes.

2.3 Exploring Clusters

Table 1 shows the clusters with at least 10 documents for K-Means clustering with K=16. The words in each cluster are the dimensions of the centroid with the largest magnitude. The documents shown are those that are closest to the centroid of the cluster.

Overall, the generated clusters make sense, but there are some points of confusion:

- The words that make up cluster 0 have little relation to each other. This cluster contains the majority of the documents.
- Cluster 1 contains churches as well as colleges.

• Cluster 3 contains documents related to TV shows and sports because both contain the word "season." Another explanation for this is that sports players appear on Dancing with the Stars.

Table 1: K-Means clusters with K=16 and at least 10 documents.

Cluster	Size	Words	Documents		
0	10061	females	mcgillpainquestionnaire		
		station	historyofthefamily		
		family	thetussaudsgroup		
		located	nadiraactress		
		north	mansfieldsummithighschool		
1	3013	church	edmondscommunitycollege		
		college	helderbergcollege		
		students	oberlincongregationalchurch		
		published	lundbyoldchurch		
		institute	dioceseoflimerickandkillaloe		
2	1128	party	partyidentification		
		served	labourfarmerparty		
		general	democraticalliancesouthafrica		
		member	liberaldemocratsitaly		
		senate	christiancreditparty		
3	909	season	dancingwiththestars		
		club	davidmccracken		
		playing	gilbertcurgenven		
		seasons	bjsamsamericanfootball		
		player	livingstonewalker		
4	707	album	thegreatestdaytakethatalbum		
		released	conflictingemotions		
		songs	primalscream		
		records	leftbacklp		
5	30	rock nba	elisamartin		
)	30	basketball	kcjones		
			hakeemolajuwon		
		points	albertkingbasketball		
		season	ballstatecardinalsmensbasketball		
6	23	seasons riots	201011southfloridabullsmensbasketballteam sowetouprising		
	23	police	1992losangelesriots		
		murder	nikolaybogolepov		
		captured	josephlamothe		
		robbery	jenmi		
7	13	congo	republicofcabinda		
'		subtropical	brownrumpedbunting		
		republic	copperbeltprovince		
		zambia	leptopelisviridis		
		zimbabwe	yellowthroatedpetronia		

2.4 K-Means++

K-Means++ is a method for selecting the initial cluster centers for K-Means. In the normal version of K-Means, initial cluster centroids are selected by sampling random documents from the dataset. This can produce suboptimal clusters, increase time to convergence, and increase variability in clustering performance.

K-Means++ solves this by selecting initial centroids one at a time in order to minimize distortion. Each document is given a weight proportional to the distance to the nearest centroid and the next centroid is selected at random using these weights until K centroids have been chosen. This results in better centroids than random sampling and lessens the time until convergence. However, it can

be quite expensive when K is large since the distance from each document to each cluster centroid must be computed at every iteration.

I implemented K-Means++ and expected it to increase clustering performance considerably. However, I found that it actually had a slightly negative impact on distortion and noticeably increased cluster sparsity.

I think that this is because outliers were chosen as initial centroids and the clusters never expanded to include other points.

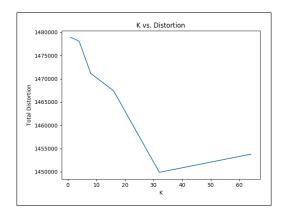


Figure 3: K versus total distortion for $K \in \{1, 2, 4, ..., 64\}$ using K-Means++.

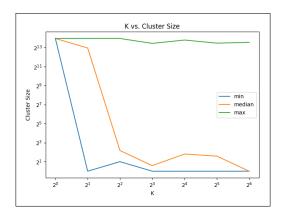


Figure 4: K versus minimum, median, and maximum cluster size for $K \in \{1, 2, 4, ..., 64\}$ with a log_2 scale on both axes. Using K-Means++.

2.5 Fighting Cluster Sparsity

After implementing K-Means++ with no success, it became apparent that there was a problem with the underlying data, and not with the clustering algorithm.

I began to investigate the provided dataset, which was in tf-idf format with no documentation. I found that stop words had been removed from the dataset, but rare words had been left in. Half of the words in the dictionary were used in 16 or fewer datasets (less than 0.1%). Tables 2, 3, and 4 show the least common, somewhat common, and most common words in the dataset, respectively.

2.5.1 Selecting a Threshold

Since the provided dataset does not contain stop words, I decided not to filter out common words. The most common words in the dataset appear to contain useful information for clustering.

Table 2: 10 least common words with the number and percentage of documents that they appear in.

Word	Documents	Percentage of Documents
frazioni	0	0.0%
threeletter	1	0.0062881%
budjovice	1	0.0062881%
gmina	1	0.0062881%
ortsgemeinden	1	0.0062881%
headwater	2	0.012576%
baronetage	3	0.018864%
breaststroke	3	0.018864%
voronezh	3	0.018864%
rosettes	3	0.018864%

Table 3: 10 words near the median with the number and percentage of documents that they appear in.

Word	Documents	Percentage of Documents
multimillion	16	0.10061%
pisa	16	0.10061%
pranks	16	0.10061%
pesticides	16	0.10061%
hesitation	16	0.10061%
convection	16	0.10061%
ortiz	16	0.10061%
stagnation	16	0.10061%
gonzlez	16	0.10061%
cummins	16	0.10061%

Table 4: 10 most common words with the number and percentage of documents that they appear in.

Word	Documents	Percentage of Documents
well	4009	25.209%
second	3538	22.247%
high	2836	17.833%
family	2579	16.217%
group	2412	15.167%
north	2364	14.865%
major	2298	14.45%
large	2227	14.004%
general	2187	13.752%
long	2164	13.607%

Selecting a lower threshold involves finding the right balance between information loss and cluster sparsity. Words that are only used in one document should be ignored since they provide no useful clustering information. The problem gets murkier with words that are used in just a few datasets. Table 3 lists ten words that are somewhat common, in that their counts are exactly at the median word count. Some of these words seem useful for clustering, such as "multimillion", "pesticides", and "convection". Others are names which do not necessarily indicate document similarity.

One of the large problems here is that we are working with only 15,000 of the 5 million English Wikipedia articles. If we were working with more data, we could produce more fine-grained clusters with rarer words. However, since we only have a tiny subset of Wikipedia, we just don't have enough examples to produce these narrow clusters.

There is no correct way to choose a tf-idf threshold. I ended up choosing to get rid of the least common half of words, but many other choices would be just as valid.

2.6 Effect of Seed on Distortion

After observing some unexpected variability in my results, I decided to investigate the effect of random seeds on distortion.

I found that the choice of seed can have a large impact on cluster quality and distortion. Because of this, it is very important to run K-Means with at least a few different seeds and select the best one. While this can be too computationally expensive to do when running experiments, it is important to do when generating the final clusters.

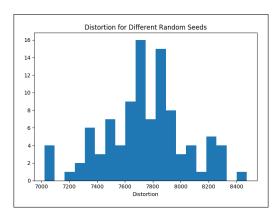


Figure 5: Distortion using K-Means++ on 100 training documents with different random seeds.

3 Conclusion

Table 5 shows the final clusters. While 64 clusters were generated, only the largest 10 are shown. Only words that occurred in at least 16 documents, which is about half of all words, were considered. K-Means++ was used to generate the initial cluster centroids. Names have been added to each cluster.

3.1 Issues with K-Means

3.1.1 Variability in Cluster Size

The most glaring issue with these results is that the sizes of the clusters vary greatly. One cluster contains about half of the documents while the other 63 contain the other half. There are many useless one document clusters.

It is difficult to know if this is an intrinsic part of the data, or an issue with the K-Means algorithm. I suspect that it is a combination of both, and that K-Means is not the best model for the data.

While exploring the data, I did try to visualize it using Multidimensional Scaling (MDS), but I found that too much information was lost in the dimensionality reduction for it to be of any use. Textual data is high dimensional and sparse by its nature, which makes it somewhat unintuitive to work with in this context. It would be interesting to further explore dimensionality reduction techniques as well as alternate clustering algorithms and see if they could help address the problem of cluster sparsity.

3.1.2 Different Meanings of the Same Word

Some clusters contain documents that are related by a single word, where each document refers to a different meaning of the word. For example, cluster 9 contains documents about mars. Its main words are "mars", "crater", "astronomical", "expedition", and "planets". However, there are two documents in this cluster that have nothing to do with mars. "hostagefilm" (*Hostage* (film)) is an

American action movie whose main character is named Marshall "Mars" Krupcheck. "marsrapper" (*Mars* (rapper)) is a Bay Area rapper who goes by "Mars".

3.2 Further Research

I had intended to explore Latent Dirichlet Allocation but ended up focusing on K-Means instead. As a starting point, I used SciKit's LDA implementation, which seemed to work quite well. The words in each topic are related just as they should be. This would have been be a great topic to research more, if I had the time.

Table 5: 16 topics generated using Latent Dirichlet Allocation.

Topics
magazine, published, journal, format, awards
painting, gallery, paintings, painted, station
greek, empire, ancient, jews, dynasty
served, appointed, general, senate, member
systems, system, design, device, czech
party, security, elections, rights, police
station, air, route, railway, traffic
season, females, club, miles, tournament
church, building, house, located, castle
group, swedish, users, stockholm, program
engine, cars, comics, ford, novels
students, college, institute, campus, program
player, characters, red, color, blue
album, released, songs, musical, records
human, family, published, man, god
common, temperature, chemical, protein, oil

Table 6: 10 largest K-Means clusters with K=64. Using words that appear in at least 16 (0.1%) documents with K-Means++ for initializing cluster centroids.

published daughter royal stmaryschurchgrodno paris dioceseoffimerickandkillaloe 3 1001 party partyidentification served labourfarmerparty serbianliberalparty pronwennaher democraticalliancesouthafrica season bjsamsamericanfootball seasons playing davidmecracken seasons johnflanaganfootballer player gilbertcurgenven 5 640 album primalscream Rock Musterseam released thegreatestdaytakethatalbum songs records japan astorhousehotelshanghai19221959 chinese japan astorhousehotelshanghai19221959 characters listofficlepisodes 7 44 pop popmusic popmusic chart talkinginyoursleepcrystalgaylesong britishpopmusic chart talkinginyoursleepcrystalgaylesong binney june johnlittlechild doctor unitedstatesvlibby	Cluster	Size	Words	Documents	Label
family north colimneads located students system high stargateschool program technology govthazimuhammadmohsincollege high stargateschool program technology govthazimuhammadmohsincollege daughter royal paris dioceseoflimerickandkillaloe paris dioceseoflimerickandkillaloe paris served labourfarmerparty served labourfarmerparty served labourfarmerparty serianliberalparty bronwenmaher senate democraticalliancesouthafrica bisamsamericanfootball seasons playing seasons playing seasons player gilbertcurgenven player gilbertcurgenven griderickingemotions records songs conflictingemotions records bornthisway rock sweetkisses pearl imperialjapanesearmyairforce characters listofficlepisodes 7 44 pop songs album characters listofficlepisodes 8 27 investigation martensvillesatanicsexscandal june johnlittlechild unitedstatesvilibby	0	8486	females	kotavamsa	
north located mladenjiiubiofbribir edmondscommunitycollege College: edmondscommunitycollege College: helderbergcollege stargateschool miltonhighschoolmiltongeorgia govthazimuhammadmohsincollege Churche published daughter molire oberlincongregationalchurch lundbyoldchurch royal stmaryschurchgrodno dioceseoflimerickandkillaloe Political Pa served member general senate democraticalliancesouthafrica senate democraticalliancesouthafrica season johnflanaganfootballer playing davidmccracken seasons johnflanaganfootballer gilbertcurgenven primalscream Rock Muster Rock Mus			station	tornadoesof1993	
located students students edmondscommunitycollege students edmondscommunitycollege helderbergeollege high stargateschool miltongeorgia govthazimuhammadmohsincollege church published daughter royal strayschurchgrodno paris dioceseoflimerickandkillaloe paris eseason ilaburfarmerparty served laburfarmerparty serbianliberalparty bronwenmaher senate democraticalliancesouthafrica bisamsamericanfootball season johnflanaganfootballer gilbertcurgenwen gilbertcurgenwen released thegreatestdaytakethatalbum conflictingemotions records papan astorhousehotelshanghai19221959 japan chinese japanese dard telepop songs album britishpopmusic chart talkinginyoursleepcrystalgaylesong brollithild doctor unitedstates vlibby			family	whitby	
1 2226 students edmondscommunitycollege helderbergcollege high stargateschool miltonhighschoolmiltongeorgia govthazimuhammadmohsincollege molire molire oberlincongregationalchurch lundbyoldchurch stargyschurchgrodno dioceseoflimerickandkillaloe paris dioceseoflimerickandkillaloe paris dioceseoflimerickandkillaloe paris dioceseoflimerickandkillaloe paris dioceseoflimerickandkillaloe paris dioceseoflimerickandkillaloe paris dioceseoflimerickandkillaloe political Pa partyidentification abourfarmerparty served abourfarmerparty bronwenmaher senate democraticalliancesouthafrica bjsamsamericanfootball Sports davidmccracken johnflanaganfootballer player gilbertcurgenven player gilbertcurgenven primalscream thegreatestdaytakethatalbum conflictingemotions records bornthisway songs sweetkisses frederickringer japan astorhousehotelshanghai19221959 japanesebadger imperialjapanesearmyairforce characters listofficlepisodes Pop Mus digitalmonetarytrust andyhayman system martensvillesatanicsexscandal johnlittlechild doctor unitedstatesvlibby			north	colinmeads	
1 2226 students edmondscommunitycollege helderbergcollege high stargateschool miltonhighschoolmiltongeorgia govthazimuhammadmohsincollege molire molire oberlincongregationalchurch lundbyoldchurch stargyschurchgrodno dioceseoflimerickandkillaloe paris dioceseoflimerickandkillaloe paris dioceseoflimerickandkillaloe paris dioceseoflimerickandkillaloe paris dioceseoflimerickandkillaloe paris dioceseoflimerickandkillaloe paris dioceseoflimerickandkillaloe political Pa partyidentification abourfarmerparty served abourfarmerparty bronwenmaher senate democraticalliancesouthafrica bjsamsamericanfootball Sports davidmccracken johnflanaganfootballer player gilbertcurgenven player gilbertcurgenven primalscream thegreatestdaytakethatalbum conflictingemotions records bornthisway songs sweetkisses frederickringer japan astorhousehotelshanghai19221959 japanesebadger imperialjapanesearmyairforce characters listofficlepisodes Pop Mus digitalmonetarytrust andyhayman system martensvillesatanicsexscandal johnlittlechild doctor unitedstatesvlibby			located	mladeniiiubiofbribir	
high program miltonhighschoolmiltongeorgia govthazimuhammadmohsincollege 2 2085 church published daughter royal stmaryschurchgrodno paris dioceseoflimerickandkillaloe 3 1001 party partyidentification served labourfarmerparty member serbianliberalparty general bronwenmaher senate democraticalliancesouthafrica bjsamsamericanfootball season johnflanaganfootballer gilbertcurgenven player gilbertcurgenven player gilbertcurgenven songs conflictingemotions records bornthisway rock sweetkisses pearl chiarcs part imperialjapanesearmyairforce characters listofficlepisodes 7 44 pop popmusic talkinginyoursleepcrystalgaylesong lyne governock songs album britishpopmusic chart talkinginyoursleepcrystalgaylesong songlian money system martensvillesatanicsexscandal june johnlittlechild doctor unitedstatesylibby Sturche Oberthick oberthis agovthazmandmohsincollege molicen govthazimuhammadmohsincollege Churche govthazimuhammadmohsincollege Churche molice Ghurche oberthical govthazimuhammadmohsincollege Churche molice Churche oberthical govthazimuhammadmohsincollege Churche molice Oberthical govthazimuhammadmohsincollege Churche molice Churche oberthical govthazimuhammadmohsincollege Churche oberthical lundbyloidchurch stmaryschurche lundbyloidchurch stmaryschurche democraticalnalchurch lundbyloidchurch stmaryschurche democraticalnalchurch lundbyloidchurch stmaryschurche democraticalnalchurch lundbyloidchurch stmaryschurche democraticalnalchurch lundbyloidchurch stmaryschurche democraticalnachillaloe 4 932 Caso dioceseofficerickinger servial party govthazime g	1	2226	students	edmondscommunitycollege	Colleges
Program technology govthazimuhammadmohsincollege molire oberlincongregationalchurch published daughter royal stmaryschurchgrodno dioceseoflimerickandkillaloe Political Paserved labourfarmerparty served labourfarmerparty serbianliberalparty bronwenmaher senate democraticalliancesouthafrica season bjsamsamericanfootball seasons johnflanaganfootballer gilbertcurgenven Political Paserved labourfarmerparty serbianliberalparty bronwenmaher senate democraticalliancesouthafrica season bjsamsamericanfootball seasons johnflanaganfootballer gilbertcurgenven Political Paserved labourfarmerparty serbianliberalparty bronwenmaher democraticalliancesouthafrica seasons johnflanaganfootballer gilbertcurgenven Political Paserved labourfarmerparty serbianliberalparty bronwenmaher democraticalliancesouthafrica seasons johnflanaganfootballer gilbertcurgenven Popmalscream released thegreatestdaytakethatalbum conflictingemotions records bornthisway sweetkisses Frederickringer japan astorhousehotelshanghai19221959 japanesebadger imperialjapanesearmyairforce characters listofflelepisodes Pop Mus songs teenpop album britishpopmusic talkinginyoursleepcrystalgaylesong blahblahlahlbum digitalmonetarytrust andyhayman system june johnlittlechild unitedstatesvlibby			system	helderbergcollege	
technology govthazimuhammadmohsincollege molire published daughter poberlincongregationalchurch lundbyoldchurch stmaryschurchgrodno dioceseoflimerickandkillaloe party served member serbianliberalparty general bronwenmaher democraticalliancesouthafrica bisamsamericanfootball davidmecracken johnflanaganfootballer gilbertcurgenven primalscream released thegreatestdaytakethatalbum songs records papan chinese japan chinese parl investigation money system june doctor unitedstates vlibby 8 27 investigation molire molire molire moleval public darcingwithtlestars davidmecracken johnflanaganfootballer gilbertcurgenven primalscream Rock Muster papan gilbertcurgenven primalscream released thegreatestdaytakethatalbum songs conflictingemotions records bornthisway sweetkisses 6 280 japanese frederickringer astorhousehotelshanghai 19221959 japanesebadger imperialjapanesearmyairforce characters listofflclepisodes 7 44 pop popmusic teenpop britishpopmusic talkinginyoursleepcrystalgaylesong blablablablablablum money system june johnlittlechild unitedstatesvlibby			high	stargateschool	
2 2085 church published oberlincongregationalchurch published daughter royal stmaryschurchgrodno paris dioceseoflimerickandkillaloe 3 1001 party partyidentification labourfarmerparty served labourfarmerparty serbianliberalparty bronwenmaher senate democraticalliancesouthafrica bjsamsamericanfootball dancingwiththestars playing season long season player gilbertcurgenven 5 640 album primalscream thegreatestdaytakethatalbum songs conflictingemotions records papanese frederickringer astorhousehotelshanghai19221959 japanese pearl imperialjapanesearmyairforce characters listofficlepisodes 7 44 pop popmusic teenpop songs album chart trock talkinginyoursleepcrystalgaylesong blahblahblahalbum money system june johnlittlechild unitedstatesvlibby Churche oberlincongregationalchurch lundbedruch oberlincongregationalchurch lundbedruch aberlincongregationalchurch lundbedruch aberlincongregationalchurch lundbedruch oberlincongregationalchurch lundbedruch oberlincongregationalchurch lundbedruch oberlincongregationalchurch lundbedruch aberlincongregationalchurch lundbedruch aberlincongregationalchurch lundbedruch oberlincongregationalchurch lundbedruch oberlincongregationalchurch lundbedruch aberlincongregationalchurch lundbedruch aberlincongregationalchurch lundbedruch aberlincongregationalchurch lundbedruch potential partyidentification partyidentification oberlincongregationalchurch lundbedruch protect alburghout party served labourfarmerparty serbianlibenderickandkillaloe partyidentification partyidentificatio			program	miltonhighschoolmiltongeorgia	
published daughter royal stmaryschurchgrodno paris dioceseoflimerickandkillaloe 3 1001 party partyidentification served labourfarmerparty serbianliberalparty general bronwenmaher democraticalliancesouthafrica season bjsamsamericanfootball seasons playing davidmecracken johnflanaganfootballer player gilbertcurgenven 5 640 album primalscream released thegreatestdaytakethatalbum conflictingemotions records japan astorhousehotelshanghai19221959 japanese pearl imperialjapanesearmyairforce characters listofflelepisodes 7 44 pop popmusic chart talkinginyoursleepcrystalgaylesong brokes investigation money system investigation digitalmonetarytrust martensvilles atanics exscandal june johnlittlechild doctor unitedstatesvlibby			technology	govthazimuhammadmohsincollege	
daughter royal stmaryschurchgrodno dioceseoflimerickandkillaloe 3 1001 party party partyidentification served labourfarmerparty member general season bjsamsamericanfootball seasons player gilbertcurgenven seasons player gilbertcurgenven songs conflictingemotions portions papan astorhousehotelshanghai 19221959 japanese parl instofficies papan songs album primalspanesearmyairforce characters listofficlepisodes 7 44 pop popmusic chart talkinginyoursleepcrystalgaylesong system investigation dioceseoflimerickandkillaloe 8 27 investigation money system investigation in dioceseoflimerickandkillaloe 3 1001 party stmaryschurchgrodno dioceseoflimerickandkillaloe party demonstration protection protection protection party demonstration protection party demonstration protection party demonstration protection party demonstration party demonstra	2	2085	church	molire	Churches
royal paris dioceseoflimerickandkillaloe 3 1001 party served labourfarmerparty member serbianliberalparty bronwenmaher democraticalliancesouthafrica bisansamericanfootball club dancingwiththestars playing davidmecracken seasons johnflanaganfootballer player gilbertcurgenven 5 640 album primalscream Rock Mustreleased bornthisway rock sweetkisses records bornthisway rock sweetkisses 6 280 japanese frederickringer japan chinese japanesebadger imperialjapanesearmyairforce characters listofflelepisodes 7 44 pop popmusic teenpop britishpopmusic chart talkinginyoursleepcrystalgaylesong rock system martensvillesatanicsexscandal june johnlittlechild doctor unitedstatesvlibby			published	oberlincongregationalchurch	
paris dioceseoflimerickandkillaloe 3 1001 party served labourfarmerparty member general senate democraticalliancesouthafrica 4 932 season bjsamsamericanfootball seasons playing seasons johnflanaganfootballer player gilbertcurgenven 5 640 album primalscream released thegreatestdaytakethatalbum conflictingemotions records bornthisway rock sweetkisses japanese frederickringer japan astorhousehotelshanghai19221959 chinese pearl imperialjapanesearmyairforce listofflelepisodes 7 44 pop popmusic teenpop songs album britishpopmusic talkinginyoursleepcrystalgaylesong rock investigation money system june johnlittlechild doctor unitedstatesvlibby			daughter	lundbyoldchurch	
paris dioceseoflimerickandkillaloe 3 1001 party served labourfarmerparty member general senate democraticalliancesouthafrica 4 932 season bjsamsamericanfootball seasons playing seasons johnflanaganfootballer player gilbertcurgenven 5 640 album primalscream released thegreatestdaytakethatalbum conflictingemotions records bornthisway rock sweetkisses japanese frederickringer japan astorhousehotelshanghai19221959 chinese pearl imperialjapanesearmyairforce listofflelepisodes 7 44 pop popmusic teenpop songs album britishpopmusic talkinginyoursleepcrystalgaylesong rock investigation money system june johnlittlechild doctor unitedstatesvlibby					
3 1001 party served labourfarmerparty served labourfarmerparty serbianliberalparty bronwenmaher democraticalliancesouthafrica bjsamsamericanfootball Sports dancingwiththestars davidmecracken johnflanaganfootballer gilbertcurgenven S 640 album primalscream released thegreatestdaytakethatalbum conflictingemotions bornthisway rock sweetkisses frederickringer japan astorhousehotelshanghai19221959 japanese pearl imperialjapanesearmyairforce listofffclepisodes					
served member general senate democraticalliancesouthafrica bisansamericanfootball club dancingwiththestars playing season johnflanaganfootballer gilbertcurgenven gilbertcurgenven released thegreatestdaytakethatalbum conflictingemotions records bornthisway rock sweetkisses 6 280 japanese frederickringer japan astorhousehotelshanghai 19221959 japanesebadger pearl imperialjapanesearmyairforce listofficlepisodes 7 44 pop popmusic teenpop album britishpopmusic chart talkinginyoursleepcrystalgaylesong rock investigation money system martensvillesatanicsexscandal june johnlittlechild unitedstatesvlibby	3	1001			Political Parties
member general bronwenmaher democraticalliancesouthafrica 4 932 season bjsamsamericanfootball club dancingwiththestars davidmccracken seasons johnflanaganfootballer gilbertcurgenven 5 640 album primalscream released thegreatestdaytakethatalbum conflictingemotions bornthisway rock sweetkisses 6 280 japanese frederickringer japan astorhousehotelshanghai19221959 chinese japanesebadger pearl imperialjapanesearmyairforce listofficlepisodes 7 44 pop popmusic teenpop album britishpopmusic chart talkinginyoursleepcrystalgaylesong rock investigation money system martensvillesatanicsexscandal june johnlittlechild doctor unitedstatesvlibby			-		
general senate democraticalliancesouthafrica 4 932 season bjsamsamericanfootball Sports club dancingwiththestars playing davidmccracken seasons johnflanaganfootballer player gilbertcurgenven 5 640 album primalscream released thegreatestdaytakethatalbum songs conflictingemotions records bornthisway rock sweetkisses 6 280 japanese frederickringer japan astorhousehotelshanghai19221959 chinese japanesebadger pearl imperialjapanesearmyairforce characters listofflclepisodes 7 44 pop popmusic Pop Mus songs teenpop album britishpopmusic chart talkinginyoursleepcrystalgaylesong rock blahblahblahalbum 8 27 investigation digitalmonetarytrust money system martensvillesatanicsexscandal june johnlittlechild doctor unitedstatesvlibby			member		
senate democraticalliancesouthafrica bjsamsamericanfootball Sports club dancingwiththestars			general		
club playing davidmccracken johnflanaganfootballer gilbertcurgenven 5 640 album primalscream released thegreatestdaytakethatalbum conflictingemotions records bornthisway rock sweetkisses 6 280 japanese frederickringer japan astorhousehotelshanghai19221959 chinese japanesebadger pearl imperialjapanesearmyairforce characters listofflclepisodes 7 44 pop popmusic popmusic teenpop album britishpopmusic chart talkinginyoursleepcrystalgaylesong rock blablahblahalbum 8 27 investigation money andyhayman system martensvillesatanicsexscandal june johnlittlechild unitedstatesvlibby			_		
club playing davidmccracken johnflanaganfootballer gilbertcurgenven 5 640 album primalscream released thegreatestdaytakethatalbum conflictingemotions records bornthisway rock sweetkisses 6 280 japanese frederickringer japan astorhousehotelshanghai19221959 chinese japanesebadger pearl imperialjapanesearmyairforce characters listofflclepisodes 7 44 pop popmusic popmusic teenpop album britishpopmusic chart talkinginyoursleepcrystalgaylesong rock blablahblahalbum 8 27 investigation money andyhayman system martensvillesatanicsexscandal june johnlittlechild unitedstatesvlibby	4	932		bjsamsamericanfootball	Sports
playing seasons johnflanaganfootballer gilbertcurgenven 5 640 album primalscream released thegreatestdaytakethatalbum conflictingemotions bornthisway rock sweetkisses 6 280 japanese frederickringer japan astorhousehotelshanghai19221959 chinese japanesebadger pearl imperialjapanesearmyairforce characters listofflclepisodes 7 44 pop popmusic popmusic songs teenpop album britishpopmusic chart talkinginyoursleepcrystalgaylesong rock blahblahblahalbum 8 27 investigation money andyhayman system martensvillesatanicsexscandal june johnlittlechild unitedstatesvlibby			club		1
seasons player gilbertcurgenven 5 640 album primalscream released thegreatestdaytakethatalbum songs conflictingemotions records bornthisway rock sweetkisses 6 280 japanese frederickringer astorhousehotelshanghai19221959 chinese japan esebadger pearl imperialjapanesearmyairforce characters listofflclepisodes 7 44 pop popmusic Pop Mus songs teenpop album britishpopmusic chart talkinginyoursleepcrystalgaylesong rock blahblahblahalbum 8 27 investigation money system martensvillesatanicsexscandal june johnlittlechild doctor unitedstatesvlibby			plaving		
player gilbertcurgenven 5 640 album primalscream released thegreatestdaytakethatalbum songs conflictingemotions records bornthisway rock sweetkisses 6 280 japanese frederickringer japan astorhousehotelshanghai19221959 chinese japanesebadger pearl imperialjapanesearmyairforce characters listofflclepisodes 7 44 pop popmusic songs teenpop album britishpopmusic chart talkinginyoursleepcrystalgaylesong rock blahblahblahalbum 8 27 investigation money system martensvillesatanicsexscandal june johnlittlechild unitedstatesvlibby Rock Must					
5 640 album released thegreatestdaytakethatalbum released songs conflictingemotions records bornthisway rock sweetkisses 6 280 japanese frederickringer japan astorhousehotelshanghai19221959 chinese japanesebadger imperialjapanesearmyairforce characters listofflclepisodes 7 44 pop popmusic portishpopmusic teenpop album chart talkinginyoursleepcrystalgaylesong rock talkinginyoursleepcrystalgaylesong blahblahblahalbum digitalmonetarytrust money system martensvillesatanicsexscandal johnlittlechild unitedstatesvlibby Rock Muster Muster Muster Market Muster M					
released songs conflictingemotions records bornthisway rock sweetkisses 6 280 japanese frederickringer japan astorhousehotelshanghai19221959 chinese japanesebadger imperialjapanesearmyairforce characters listofflclepisodes 7 44 pop popmusic Pop Mus songs teenpop album britishpopmusic chart talkinginyoursleepcrystalgaylesong rock blabblahblahalbum 8 27 investigation money andyhayman system june johnlittlechild doctor unitedstatesvlibby	5	640			Rock Music
songs records bornthisway rock sweetkisses 6 280 japanese frederickringer japan astorhousehotelshanghai19221959 chinese japanesebadger imperialjapanesearmyairforce characters listofflclepisodes 7 44 pop popmusic Pop Mus songs teenpop album britishpopmusic chart talkinginyoursleepcrystalgaylesong rock blahblahblahalbum 8 27 investigation money andyhayman system june johnlittlechild unitedstatesvlibby			released		
records rock sweetkisses 6 280 japanese frederickringer astorhousehotelshanghai19221959 chinese japanesebadger imperialjapanesearmyairforce characters listofflclepisodes 7 44 pop popmusic Pop Mus songs teenpop album britishpopmusic chart talkinginyoursleepcrystalgaylesong rock blahblahblahalbum 8 27 investigation money andyhayman system june johnlittlechild unitedstatesvlibby					
rock sweetkisses frederickringer japan astorhousehotelshanghai19221959 chinese japanesebadger pearl imperialjapanesearmyairforce characters listofflclepisodes 7					
6 280 japanese japan astorhousehotelshanghai19221959 japanesebadger pearl imperialjapanesearmyairforce characters listofflclepisodes 7 44 pop popmusic teenpop album britishpopmusic talkinginyoursleepcrystalgaylesong rock token investigation money system june johnlittlechild doctor limperial James Imperial James Imperia					
japan astorhousehotelshanghai19221959 chinese japanesebadger pearl imperialjapanesearmyairforce characters listofflclepisodes 7	6	280		frederickringer	Imperial Japan
chinese japanesebadger imperialjapanesearmyairforce characters listofflclepisodes 7			_	astorhousehotelshanghai19221959	
pearl imperialjapanesearmyairforce characters listofflclepisodes 7					
characters listofflclepisodes 7			pearl		
7 44 pop popmusic reenpop album britishpopmusic talkinginyoursleepcrystalgaylesong rock blahblahblahalbum 8 27 investigation money andyhayman system june johnlittlechild doctor popmusic talkinginyoursleepcrystalgaylesong blahblahalbum Crime andyhayman system martensvillesatanicsexscandal johnlittlechild unitedstatesvlibby					
songs teenpop album britishpopmusic chart talkinginyoursleepcrystalgaylesong rock blahblahblahalbum 8 27 investigation digitalmonetarytrust money andyhayman system martensvillesatanicsexscandal june johnlittlechild doctor unitedstatesvlibby	7	44			Pop Music
album britishpopmusic talkinginyoursleepcrystalgaylesong rock blahblahblahalbum 8 27 investigation digitalmonetarytrust money andyhayman system martensvillesatanicsexscandal june johnlittlechild doctor unitedstatesvlibby					1
chart talkinginyoursleepcrystalgaylesong rock blahblahblahalbum 8 27 investigation digitalmonetarytrust money andyhayman system martensvillesatanicsexscandal june johnlittlechild doctor unitedstatesvlibby					
8 27 investigation digitalmonetarytrust Crime money andyhayman system martensvillesatanicsexscandal june johnlittlechild doctor unitedstatesvlibby					
8 27 investigation digitalmonetarytrust Crime money andyhayman system june johnlittlechild doctor unitedstatesvlibby			rock		
money andyhayman system martensvillesatanicsexscandal june johnlittlechild doctor unitedstatesvlibby	8	27	investigation	digitalmonetarytrust	Crime
june johnlittlechild doctor unitedstatesvlibby					
june johnlittlechild doctor unitedstatesvlibby			system		
doctor unitedstatesvlibby			-		
/ 22 mais nosagemm Wats	9	22	mars	hostagefilm	Mars
crater marsrapper			crater		
astronomical entomopter					
expedition mars2					
planets carbonatesonmars					

References

- [1] MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* . 1. University of California Press. pp. 281297. MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07.
- [2] Arthur, D.; Vassilvitskii, S. (2007). "k-means++: the advantages of careful seeding" (PDF). *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 10271035.