Clustering Wikipedia Articles Using BERT and K-Means

Zohar Laskar Koriat 305678476
Achituv Drot 307934455

Introduction

Wikipedia is a free online encyclopedia created and edited by volunteers worldwide and hosted by the Wikimedia Foundation. The platform is structured hierarchically with 13 main categories, each branching into numerous subcategories. Most articles on Wikipedia belong to at least one subcategory, which is listed at the end of the article.

This project aims to group the Wikipedia articles into groups based on the similarity between the content of the Wikipedia articles (excluding its subcategory information) by language model and grouping methods. Next, we assess the connection between the text of an article (excluding its subcategory information) and its corresponding subcategory. Various methods exist to explore this connection, this project focuses on a specific approach.

Dataset

The dataset utilized in this study is sourced from Kaggle and comprises approximately 6 million Wikipedia articles. Each article in the dataset includes an ID number, the title, and the article text. During the preprocessing phase, it was observed that about 10% of Wikipedia articles lack of subcategories, necessitating their removal from the dataset. Additionally, approximately 57% of the Wikipedia articles (around 3.5 million) contain fewer than 300 words, often qualifying them as stub articles. Stub articles are very short, provide minimal information and lack detailed content and structure. Consequently, these Wikipedia articles were also excluded, leaving approximately 2.5 million Wikipedia articles for our research project.

Project Objective

The primary objective of this project is to cluster Wikipedia articles based on the similarity of their content (excluding category information) by using language model (BERT) and grouping methods (K-means). Subsequently, the project seeks to evaluate the correspondence between our clusters and the original subcategories.

Methodology

The ML model:

We used state-of-the-art pre-trained BERT model to create embedding for the text data for each Wikipedia article, capturing deep contextual and semantic relationships within the text and handles polysemy well (Polysemy is the capacity for a sign (e.g. a symbol, a morpheme, a word, or a phrase) to have multiple related meanings).

We use the pre-trained BERT model because this model has been trained on a large corpus of text, which includes the entirety of the English Wikipedia (approximately 2.5 billion words) and the BookCorpus dataset (800 million words of text from books). The combination of these datasets provides a rich and diverse source of information, enabling the BERT model to learn a wide range of language patterns and contextual information and this model has already been trained on Wikipedia data, among other sources, which helps it understand and process Wikipedia-style text effectively.

We prepare the database for use with BERT - we clean the text database to remove any unnecessary characters or formatting issues: Remove HTML tags, Remove URLs, Remove special characters and numbers, Convert to lowercase and strip leading/trailing whitespace.

Clustering with K-means:

The BERT embeddings of the articles were clustered using the K-means algorithm. This project explored 24 different values of k, ranging from 2 to 25. The primary goal of using various k values was to evaluate whether there is a consistent correspondence between our clusters and the original subcategories across different numbers of clusters.

Analysis of Top Categories:

Given the extensive number of subcategories, the analysis focused on the 50 most common ones. The most prevalent subcategory was found only in 1% of the articles, and the 50th subcategory appeared in just 0.1% of the articles. The subcategories such as birth years and unknown birth years were excluded as they did not contribute to meaningful clustering. The Louvain method was used to create a network based on the co-occurrence of category pairs, resulting in clusters derived from these categories.

The clustering analysis identified distinct groups, such as:

- Cluster 0: American Male Actors and Athletes (6 subcategories)
- Cluster 1: Films and Television Shows (6 subcategories)
- Cluster 2: Video Games (3 subcategories)
- Cluster 3: Association Football Players (9 subcategories)
- Cluster 4: American Actresses and Singers (9 subcategories)
- Cluster 5: American Politicians and Writers (12 subcategories)
- Cluster 6: Musicians and Honorees (5 subcategories)

Note: the image of the network is in the appendix.

Correspondence analysis between Wikipedia articles and top subcategories:

After performing K-means clustering on the Wikipedia articles and Louvain clustering on the top 50 subcategories, we examine the relationship between the articles and subcategories using these two clustering methods. As we will demonstrate, there is a significant connection between them.

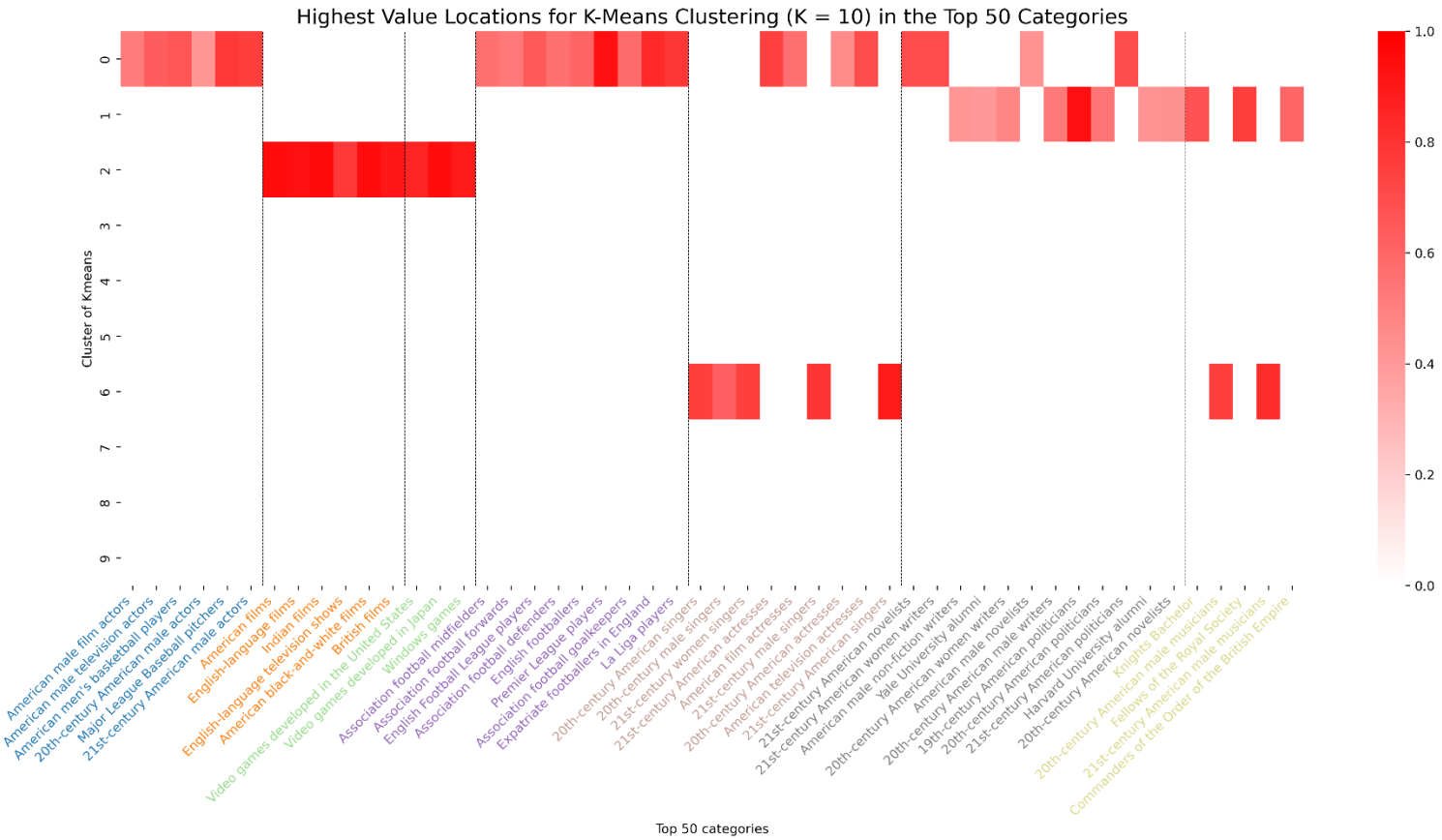
We analyzed how often each subcategory appears in each cluster according to the K-means method (adjusting for the cluster size), then we identify the cluster with the highest percentage for each subcategory.

To visualize this, we created graphs with the top 50 categories on the X-axis and the indexes that represent the clusters (k) according to the K-means method on the Y-axis. The top 50 subcategories are ordered according to their Louvain clustering, with different font colors representing each Louvain cluster. In the graph, the location of the highest adjusted percentage is depicted with a color gradient ranging from absolute red (100%) to absolute white (0%).

We repeated this process for K clusters ranging from 2 to 25. Consistently, the main finding in the top 50 subcategories, is that the subcategories within the same Louvain cluster tend to be found in the one to three clusters in K-means.

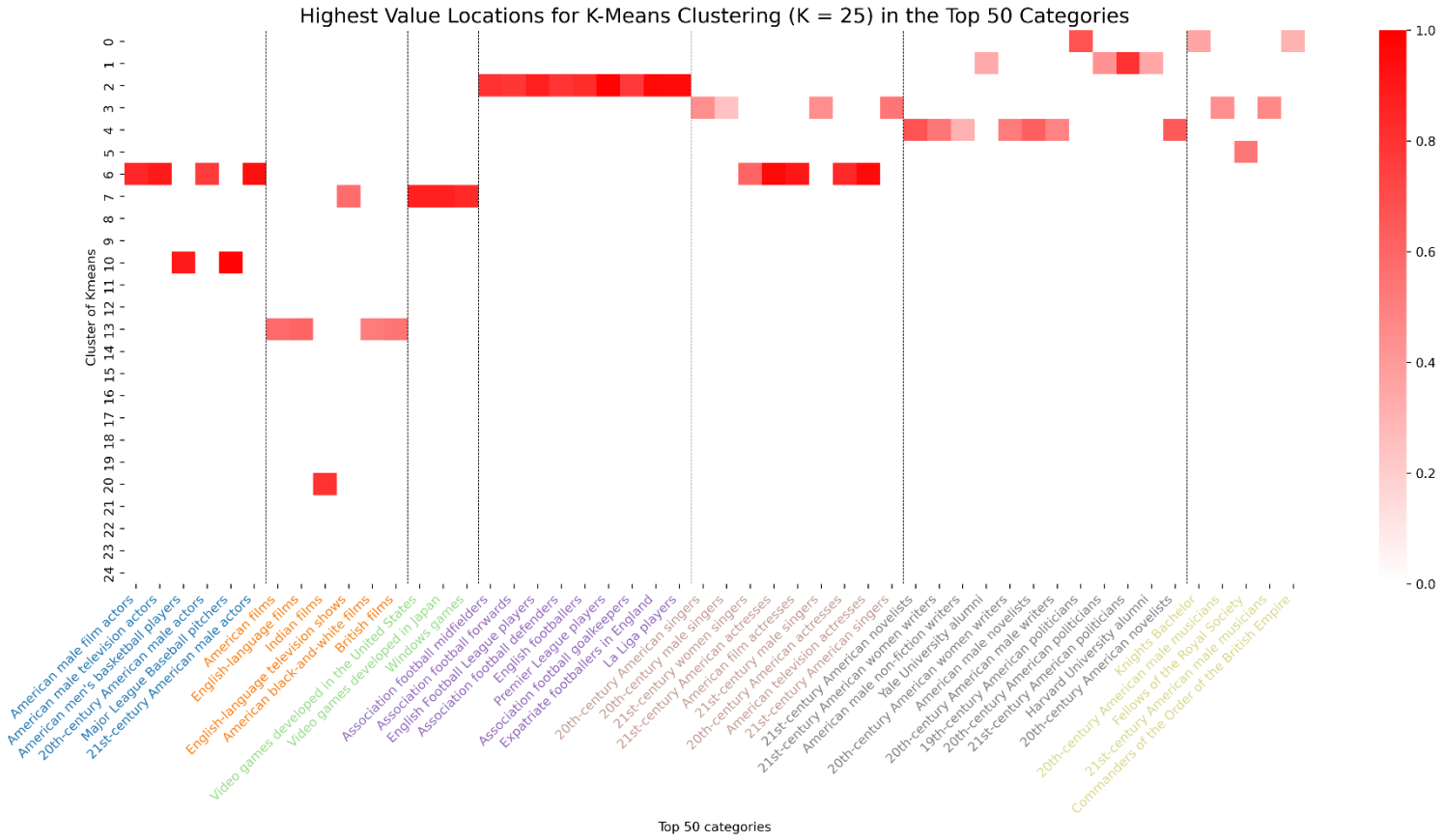
We also observe that Films and Television Shows and Video Games clusters from the Louvain method are predominantly clustered together, in most of the K's we check, indicating a strong correlation among their locations (for k=10 it is the most significant). The remaining clusters also show some degree of correlation in their placement. We will present here the graph for K=10 and for K=25 (we will include the rest in the appendix).

For K=10:



It can be seen that of the 50 most common subcategories in k=10, there is a high rate of appearances of the subcategory in only one cluster, which shows a correct distribution of the k-means method.

For K=25:

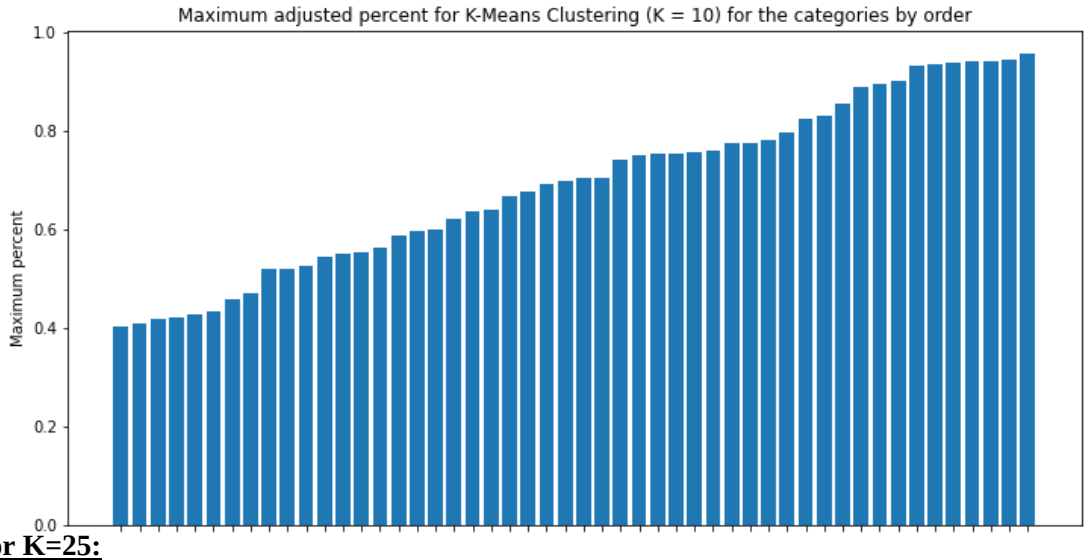


More analyze:

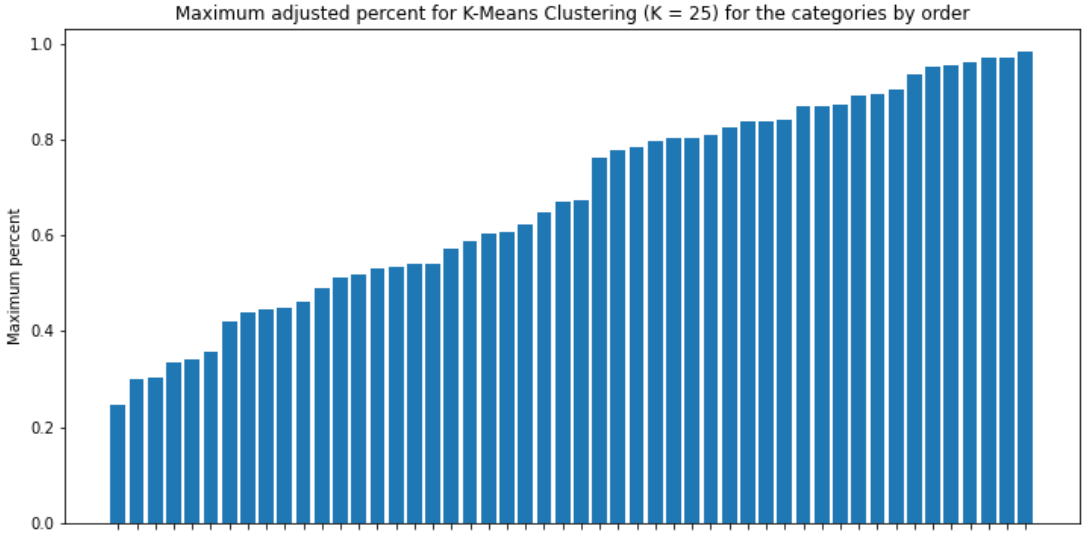
We conducted an additional analysis of the K-means clustering results for the top 50 subcategories.

We analyze the effectiveness of K-means clustering by identifying the cluster in which each of the 50 subcategories has its highest adjusted percentage. The bar plot below illustrates the maximum adjusted percentage for each subcategory within its most representative cluster. We examined these percentages for K-means of K=10 and K=25.

For K=10:



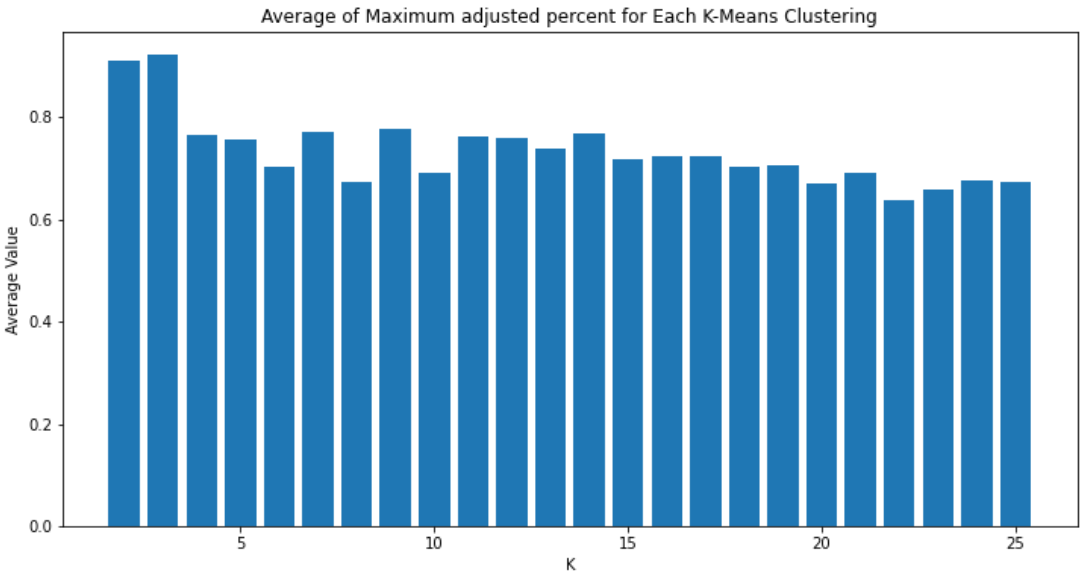
For K=25:



We conduct 24 time K-means for different K values and then we look at the 24 graphs which shows for each K value, the maximum adjusted percentage for each subcategory from the all 50 categories is within its most representative cluster. We found that among the 24 K-means values, the subcategory 'League Players' showed the highest adjusted percentage across 13 different K-means values, while the subcategory 'British Empire' consistently exhibited the lowest adjusted.

We also analyzed for each K value the average of the maximum adjusted percentages within the most representative cluster for all 50 subcategories.

Below is the average of the maximum adjusted percentage for each K-means for each K value that were calculated:



The highest accuracy rates, ranging from 91% to 92%, are observed at K=2 and K=3. Beyond this point, the accuracy declines, with a notable value of 77% at K=9, reaching a minimum of 63% at K=22.

From the above graph it can be seen that the K-means method (with the embeddings of the BERT model) succeeds in dividing the Wikipedia articles into correct clusters at all the K values we tested

Our challenges

The first challenge we encountered was the absence of a target variable in our training data, as this is an unsupervised learning problem. This required us to find an effective method for evaluating the model's performance, which means to determine the accuracy of the k-means clustering applied to the Wikipedia pages. Although we initially considered using the categories of the pages as a metric, each page typically belongs to multiple categories. Therefore, we choose to assess the model’s quality based on the 50 most frequent categories, as outlined in the document above.

The second challenge was selecting an appropriate clustering algorithm for the Wikipedia pages. We explored the HDBSCAN method, which not only clusters the data but also estimates the optimal number of clusters. However, due to several limitations, HDBSCAN did not perform adequately for our dataset, leading us to switch to the k-means algorithm.

The third challenge involved the computational demands of running both the BERT and k-means models. Processing 2.5 million pages required significant computational resources, particularly CPU time. To address this, we utilized the university's infrastructure, which offered access to GPU resources to expedite processing.

Future research

Future research could expand the analysis beyond the top 50 subcategories, potentially considering all subcategories that appear more than a few times. This broader scope would enable a more comprehensive evaluation of whether K-means clustering, when applied to BERT embeddings, consistently aligns with the Louvain clusters and the original categories.

Additionally, research could explore the effectiveness of K-means clustering across different category topics, identifying which topics the BERT-based K-means algorithm excels at clustering and which ones present challenges.

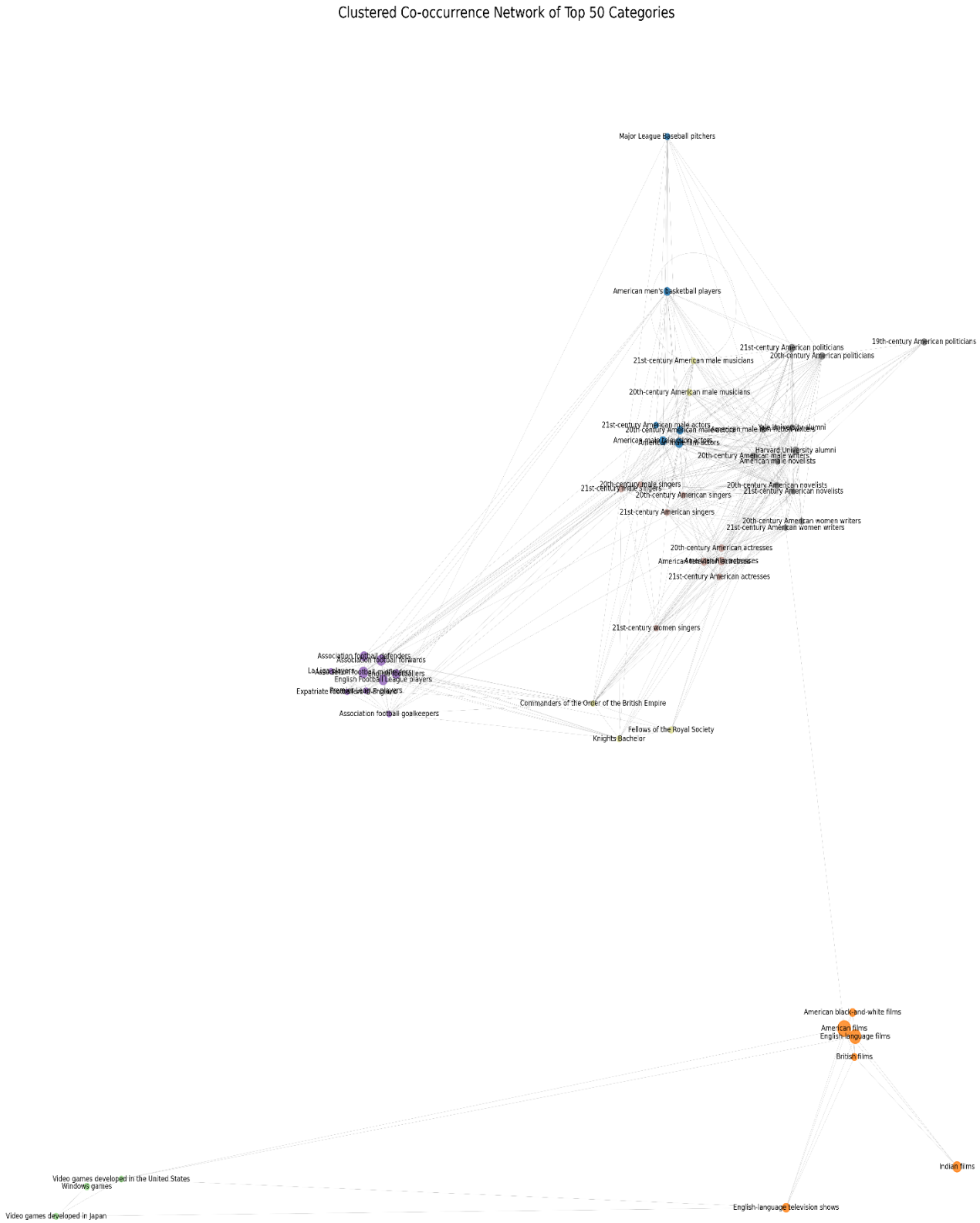
Furthermore, employing more advanced BERT models (e.g., through BERT fine-tuning) could enhance clustering accuracy and provide even more nuanced insights into category relationships.

Conclusion

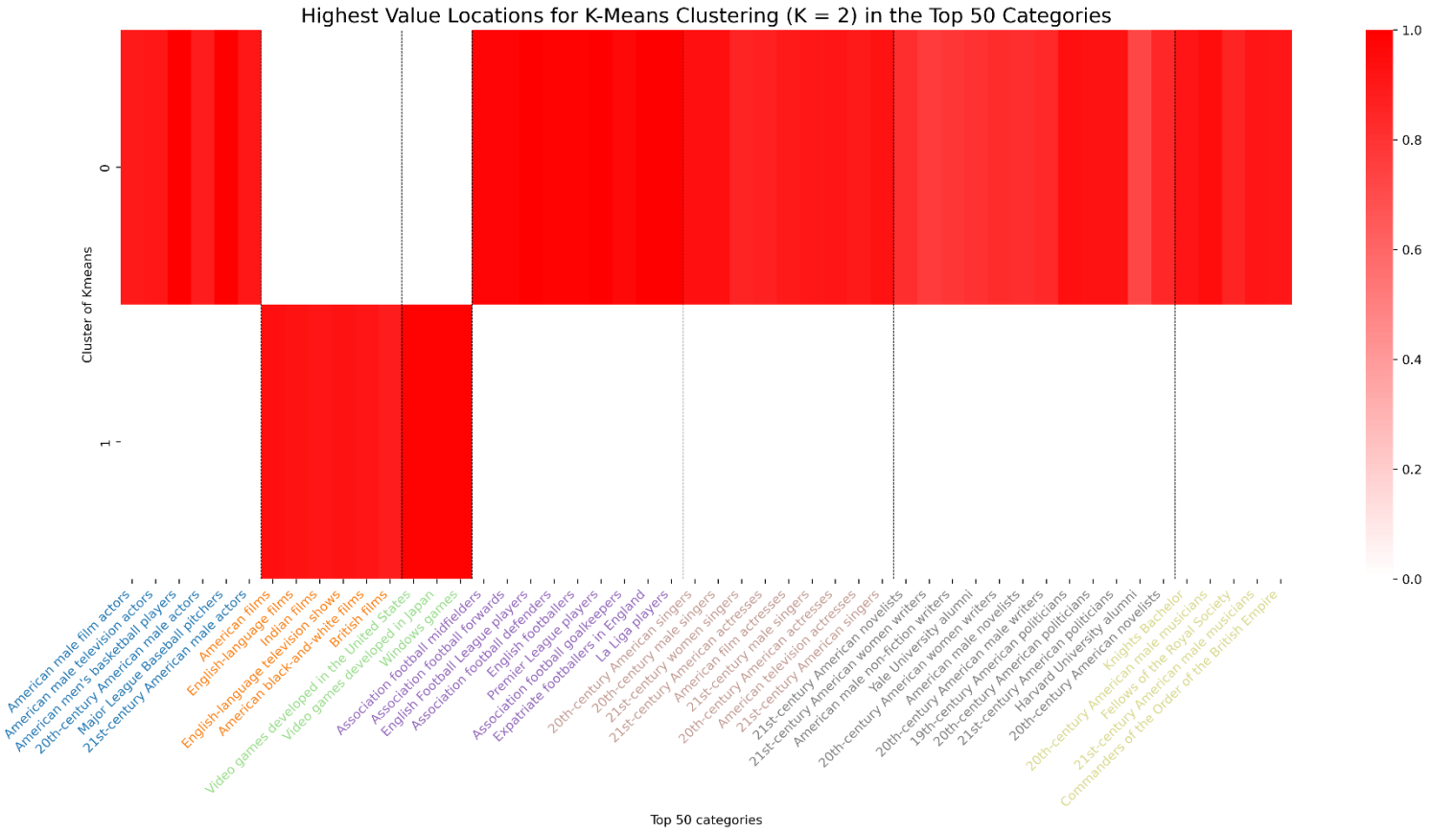
This study successfully demonstrates a significant correspondence between content-based clustering of Wikipedia articles and their original subcategories. The findings suggest that clustering algorithms like K-means, when applied to BERT embeddings, can effectively mirror the hierarchical structure of Wikipedia's categorical organization. Further exploration with a broader range of subcategories could provide additional insights into the robustness of these methods.

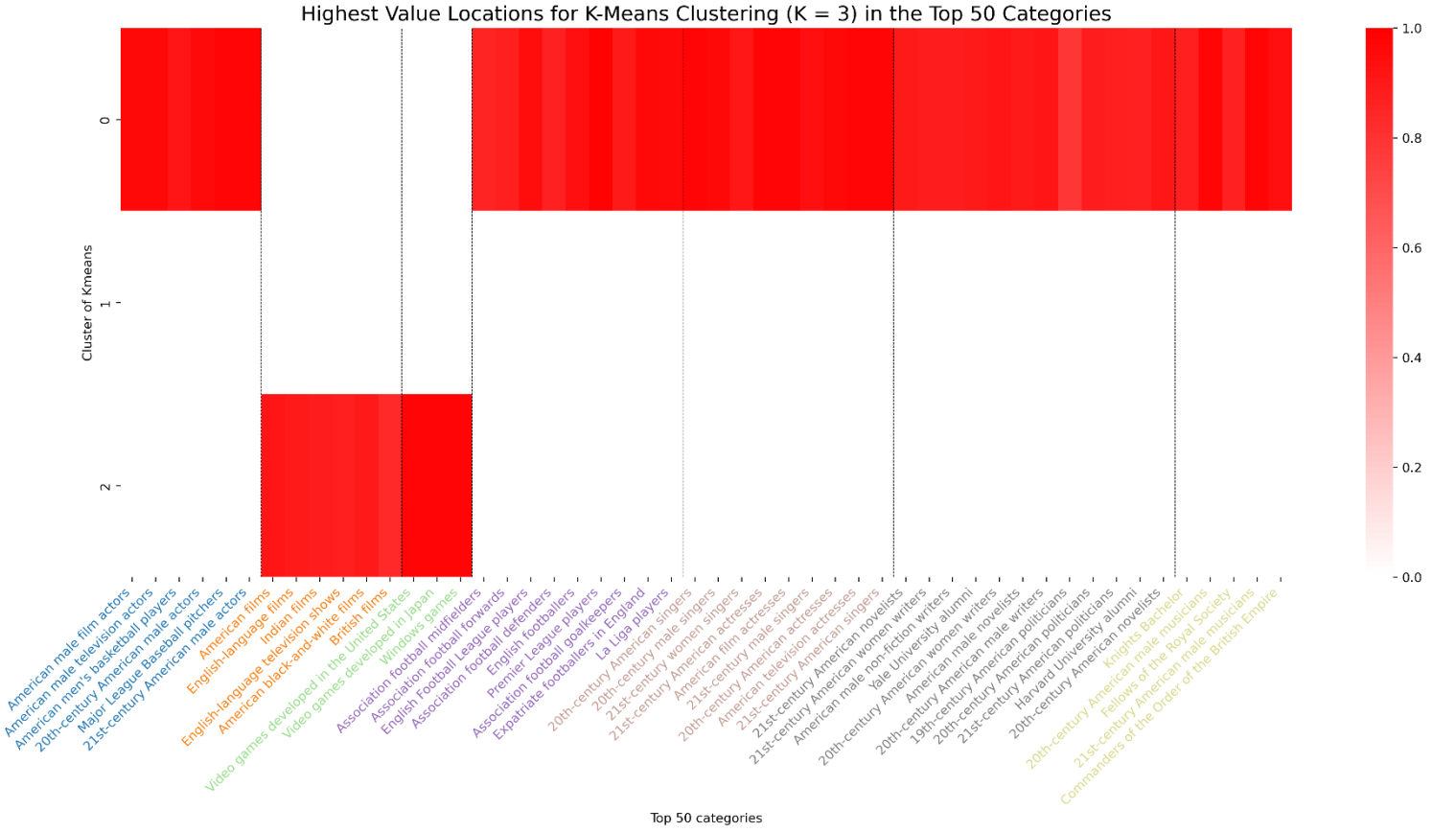
Appendix

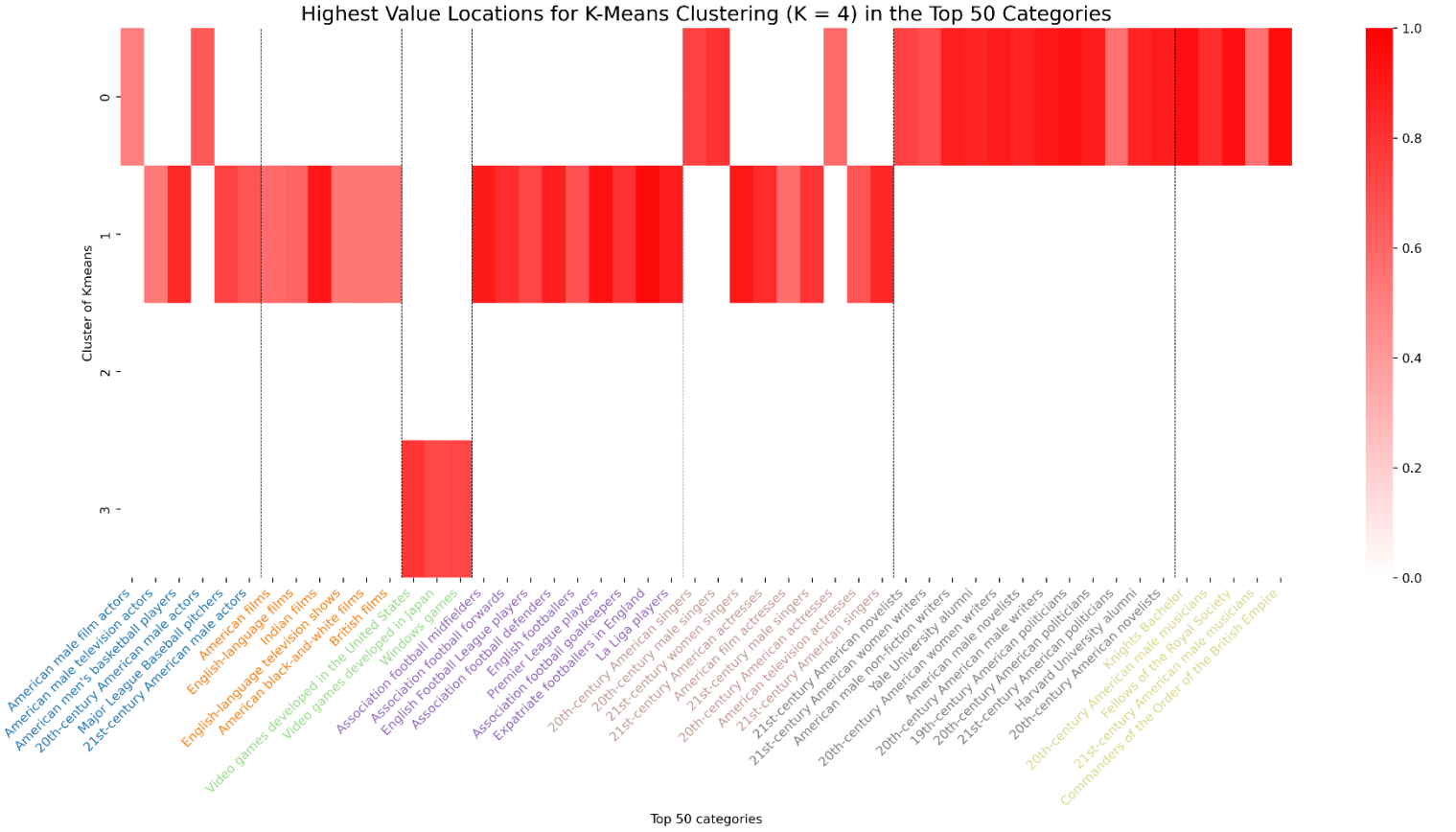
Appendix 1 – Map of the top 50 categories using Louvain clustering:

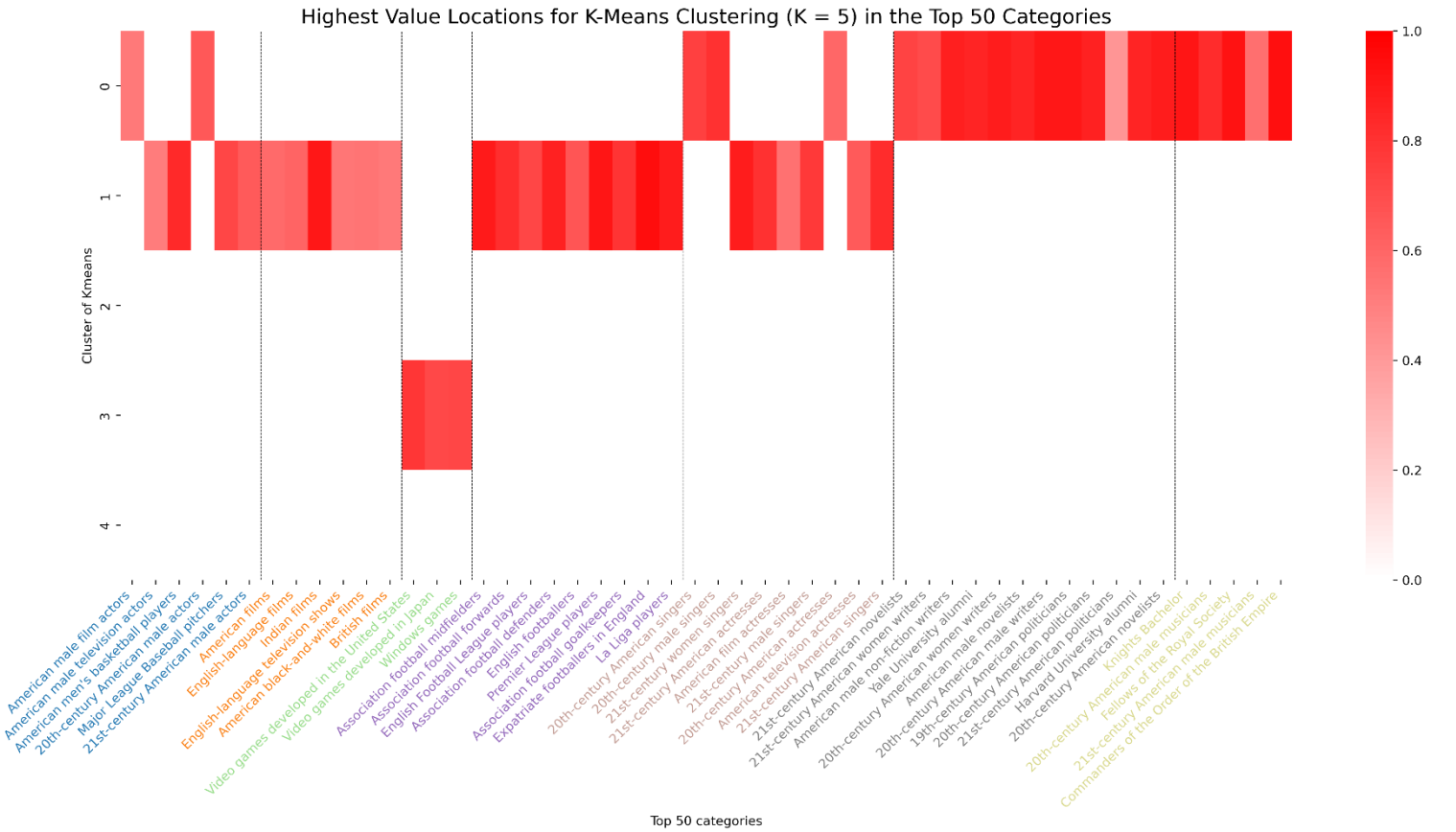


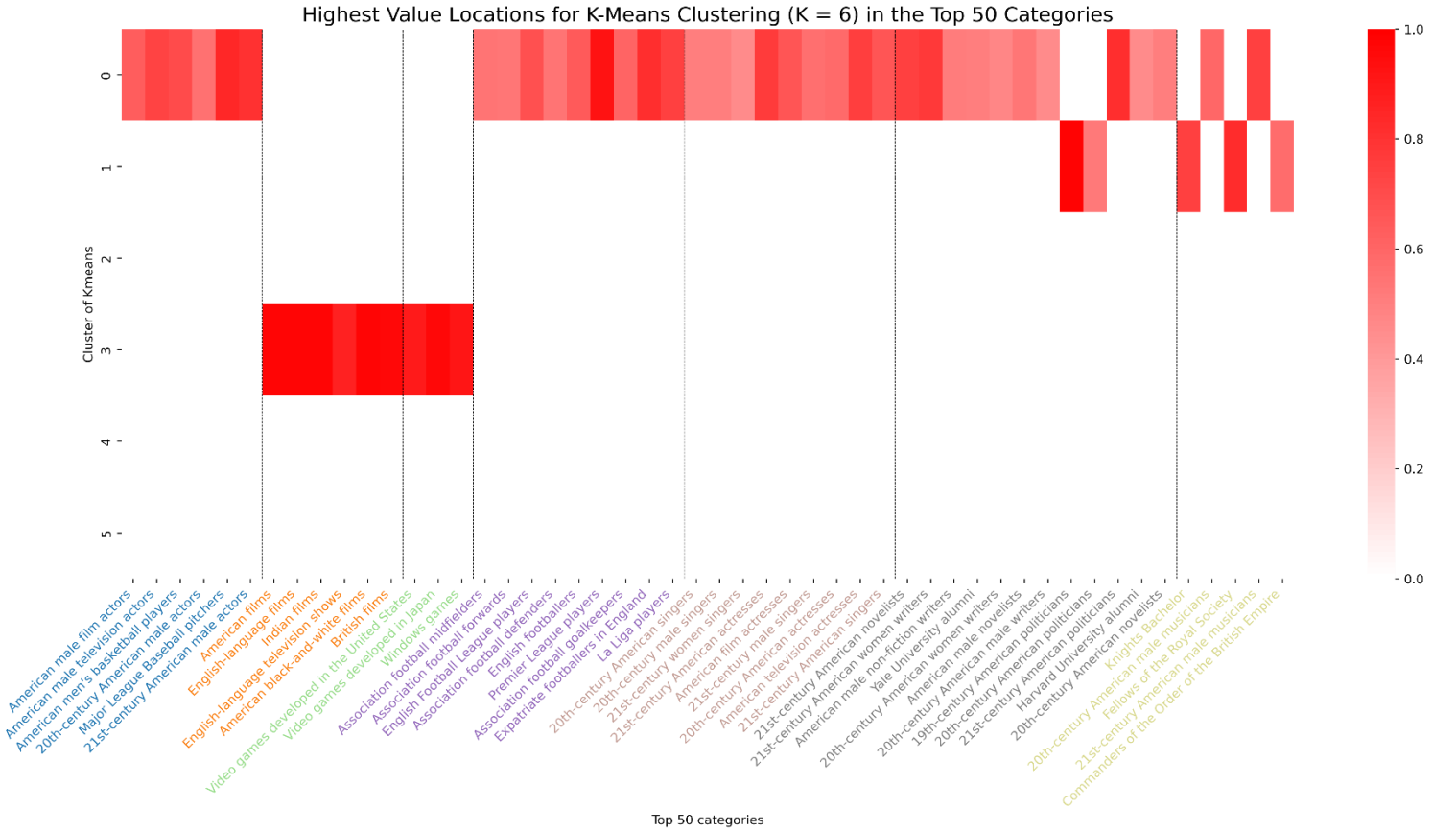
Appendix 2 – The other K means:

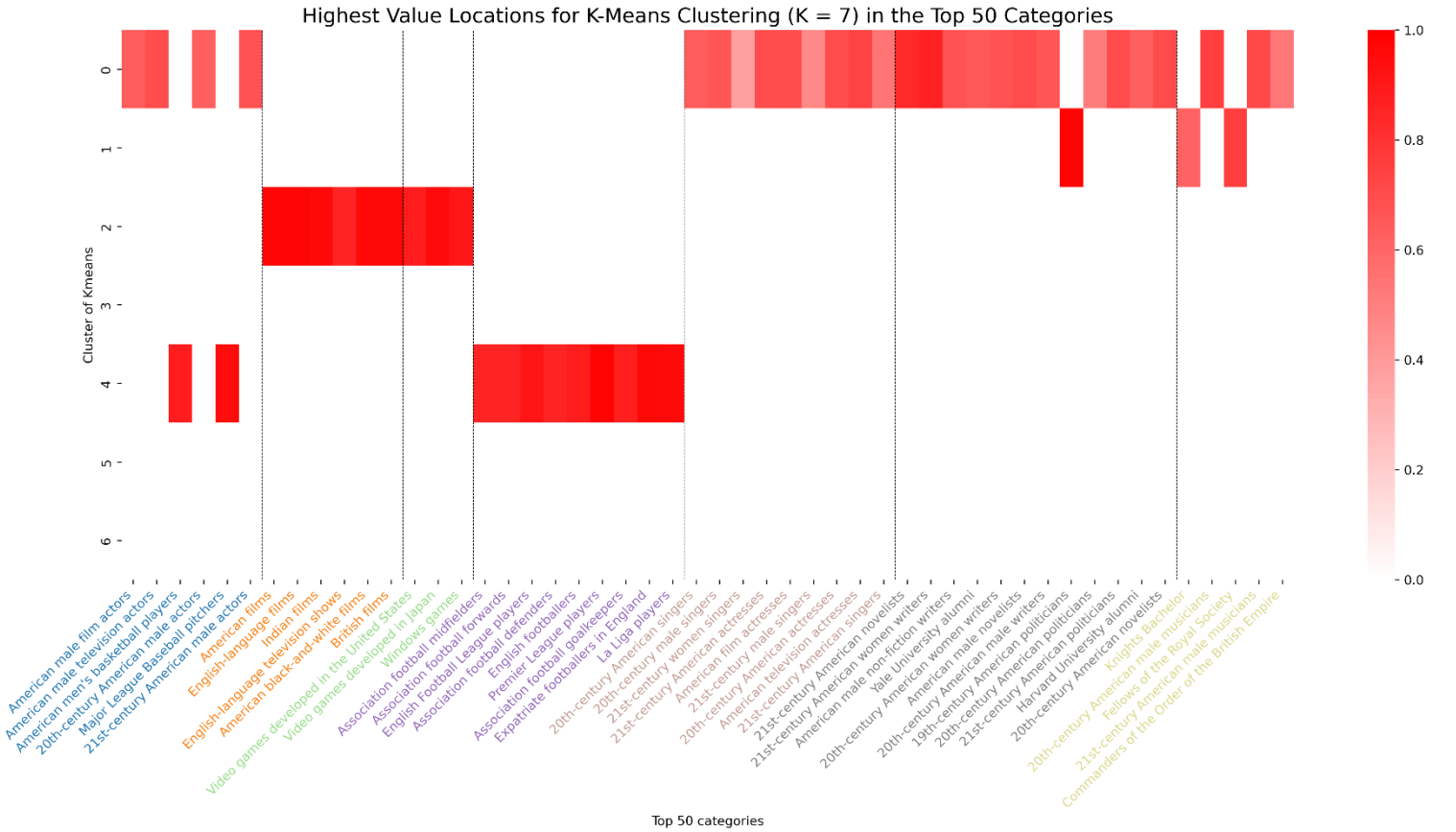


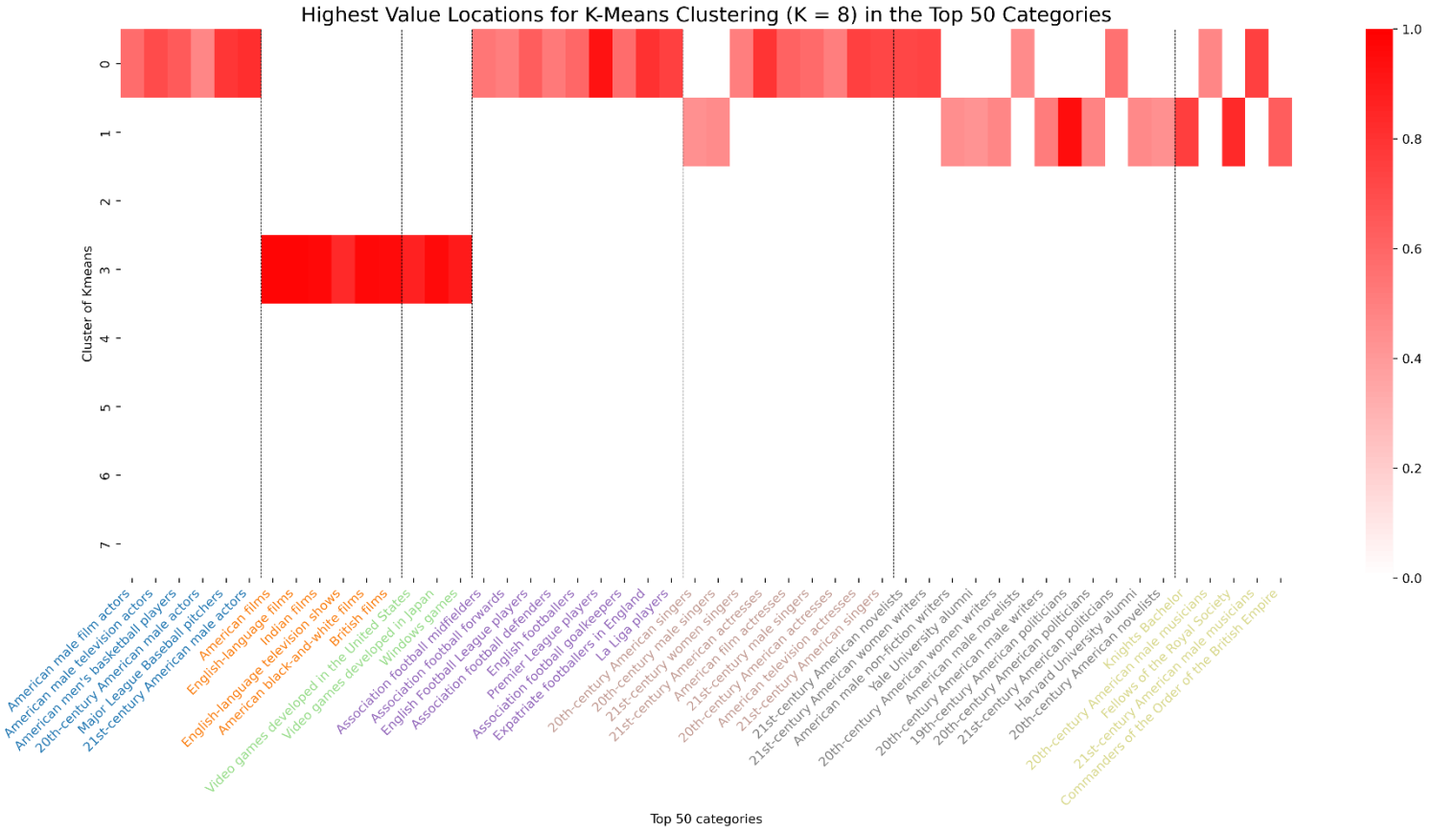


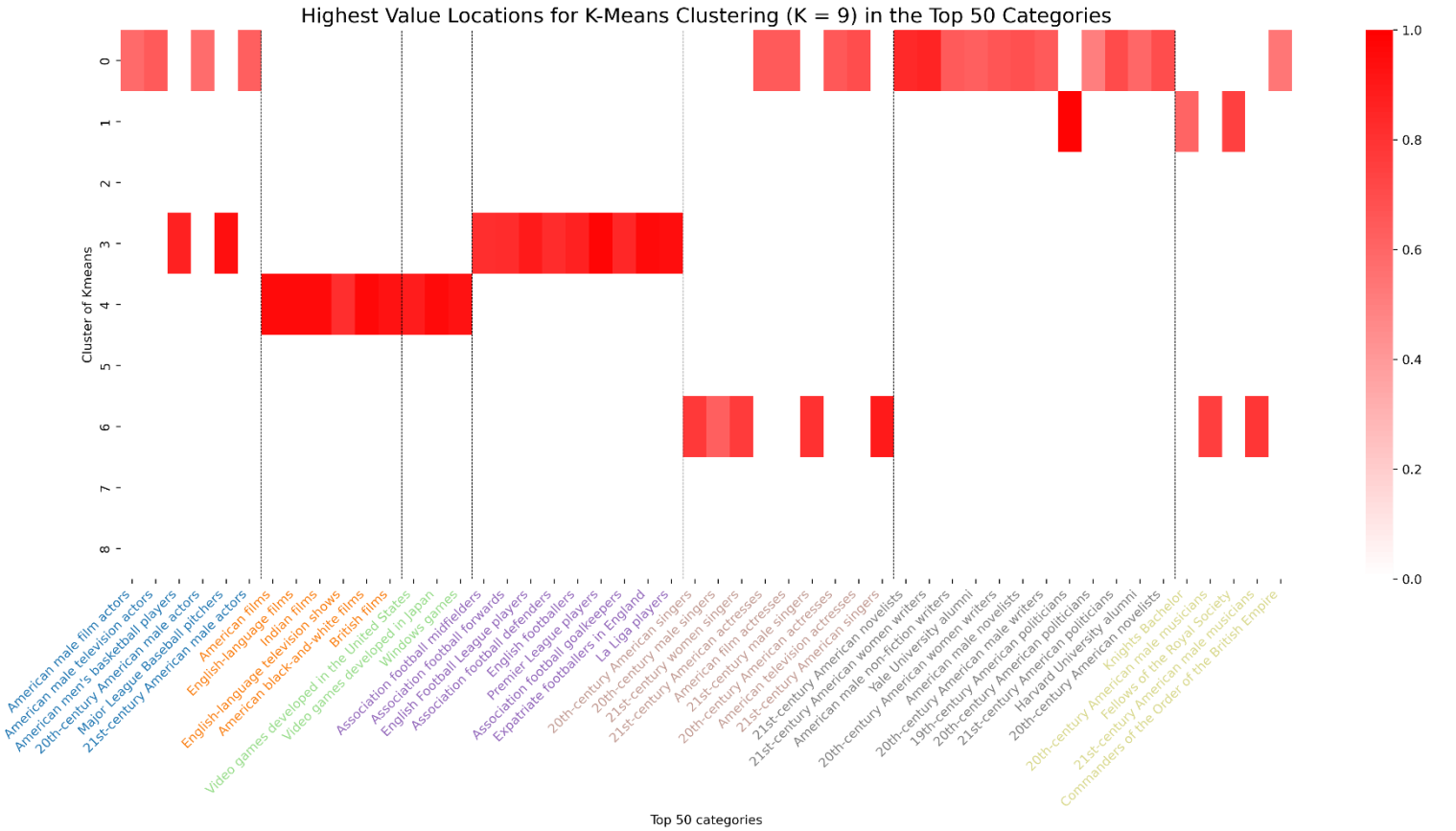

Highest Value Locations for K-Means Clustering (K = 4) in the Top 50 Categories
Cluster of Kmeans
0
1
2
3
1.0
0.8
0.6
0.4
0.2
0.0
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
British films
Video games developed in the United States
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
20th-century women singers
21st-century American singers
21st-century women singers
American film actresses
21st-century male singers
20th-century American actresses
American television actresses
21st-century American actresses
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American women writers
20th-century American male novelists
19th-century American male writers
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire
Top 50 categories

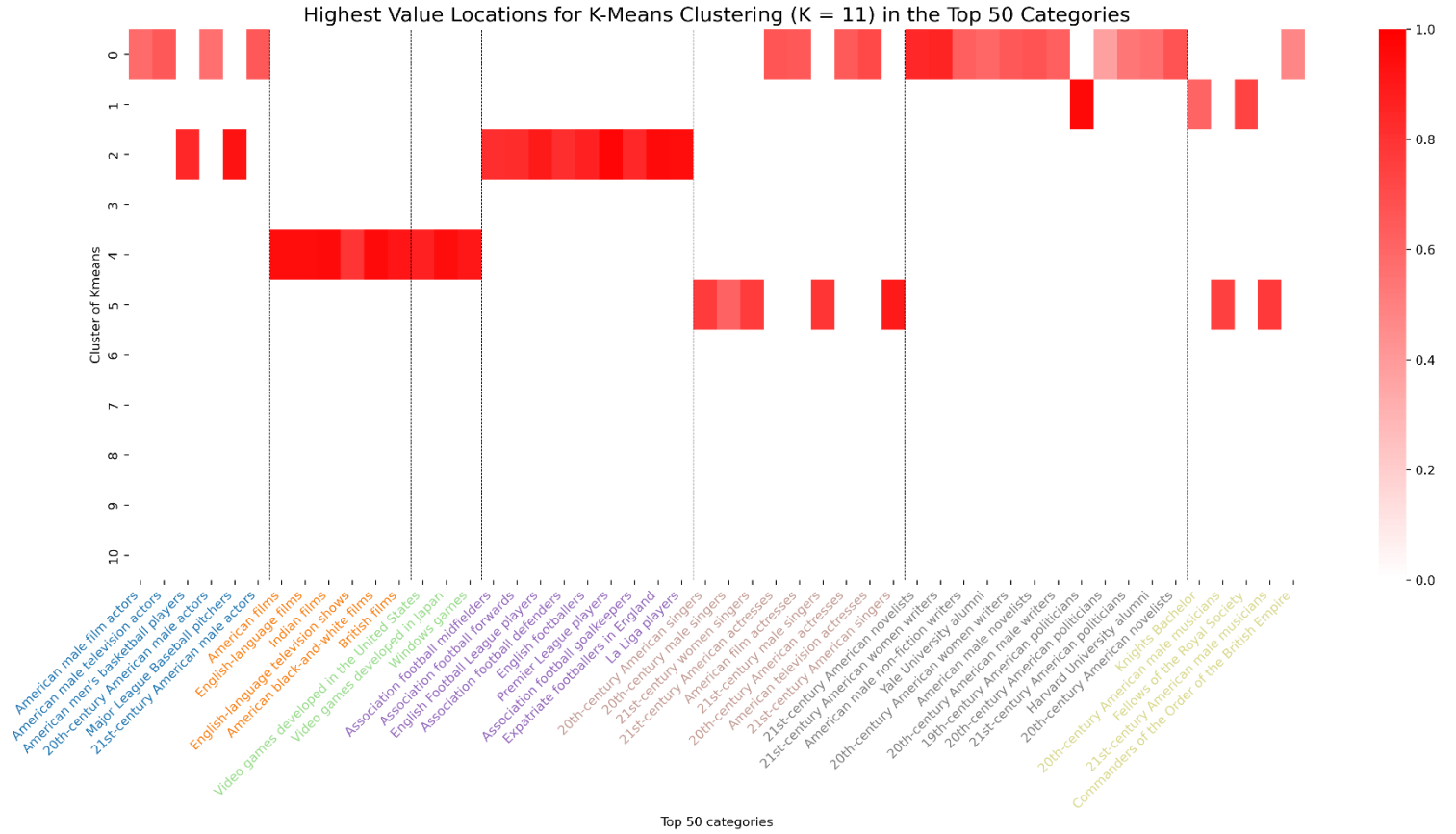

Highest Value Locations for K-Means Clustering (K = 5) in the Top 50 Categories
Cluster of Kmeans
0
1
2
3
4
1.0
0.8
0.6
0.4
0.2
0.0
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
British films
Video games developed in the United States
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
20th-century women singers
21st-century American singers
21st-century women singers
American film actresses
21st-century male singers
20th-century American actresses
American television actresses
21st-century American actresses
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American women writers
20th-century American male novelists
19th-century American male writers
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire
Top 50 categories

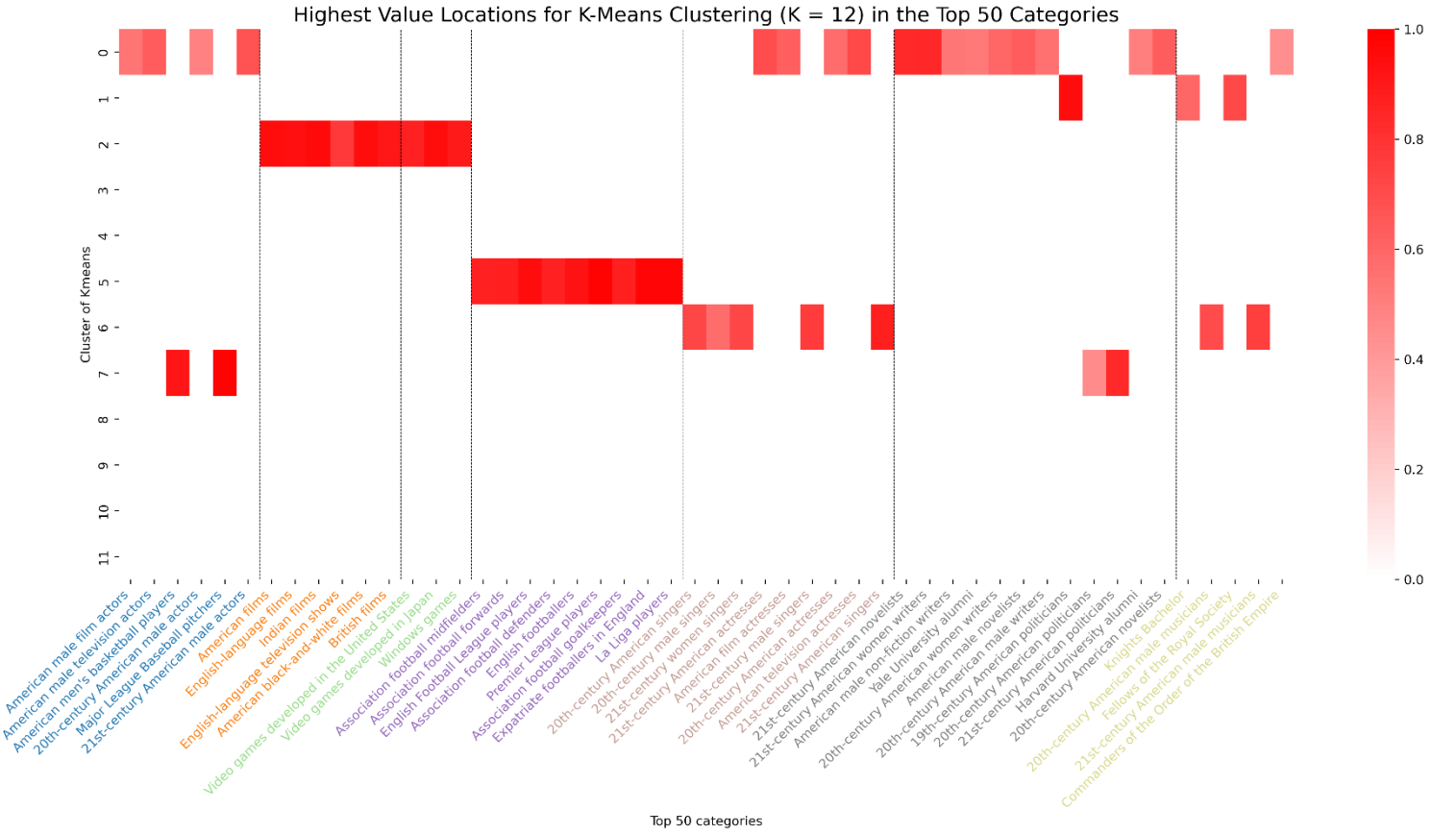

Highest Value Locations for K-Means Clustering (K = 6) in the Top 50 Categories
Cluster of Kmeans
0
1
2
3
4
5
1.0
0.8
0.6
0.4
0.2
0.0
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
British films
Video games developed in the United States
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
20th-century women singers
21st-century American singers
21st-century women singers
American film actresses
20th-century male singers
21st-century male singers
American television actresses
20th-century American actresses
21st-century American actresses
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American male novelists
20th-century American male writers
19th-century American politicians
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire
Top 50 categories

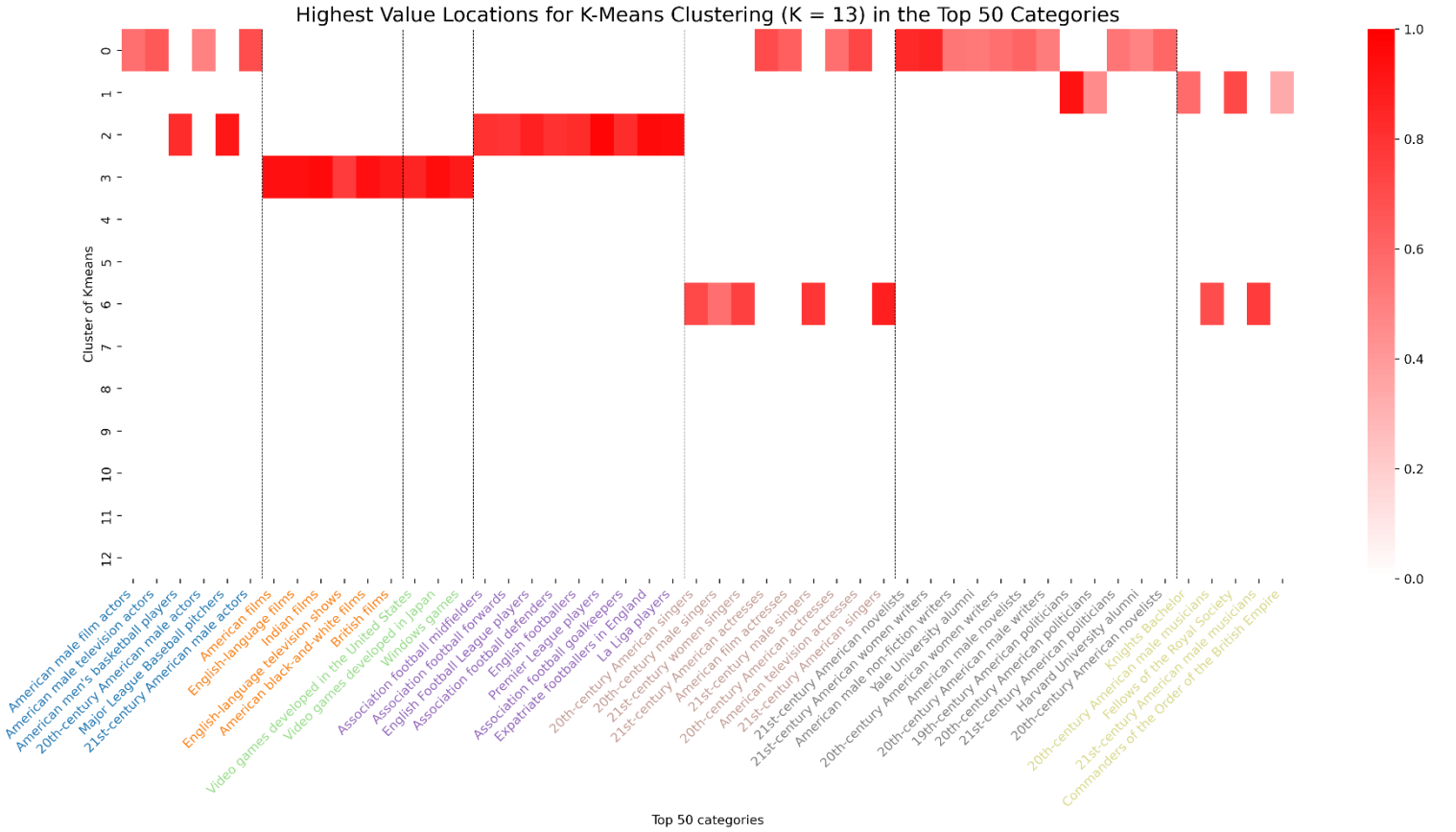

Highest Value Locations for K-Means Clustering (K = 7) in the Top 50 Categories
Cluster of Kmeans
0
1
2
3
4
5
6
1.0
0.8
0.6
0.4
0.2
0.0
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
British films
Video games developed in the United States
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
20th-century women singers
21st-century American singers
21st-century women singers
American film actresses
20th-century male singers
21st-century male singers
American television actresses
20th-century American actresses
21st-century American actresses
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American male novelists
20th-century American male writers
19th-century American politicians
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire
Top 50 categories

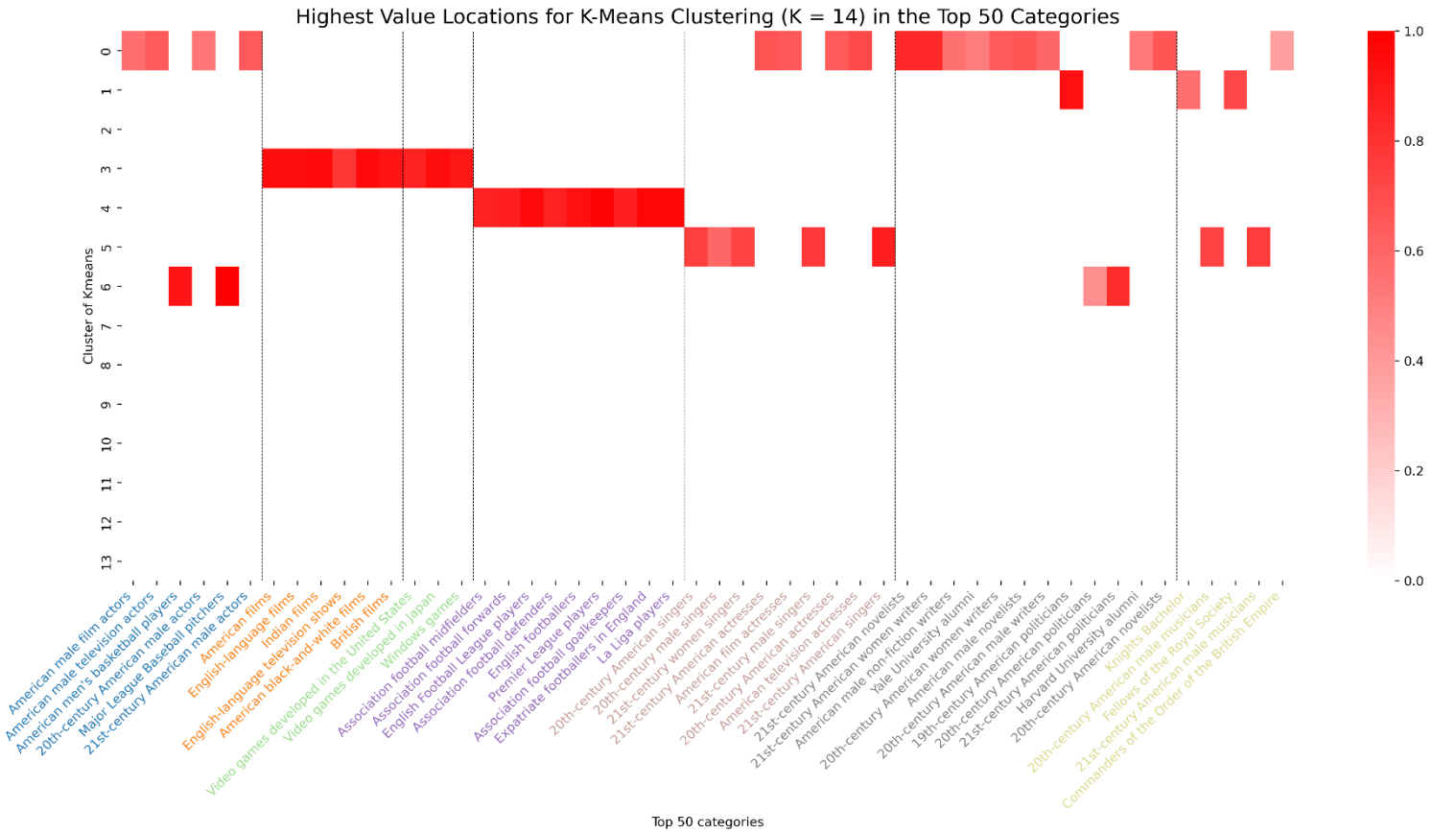

Highest Value Locations for K-Means Clustering (K = 8) in the Top 50 Categories
Cluster of Kmeans
0
1
2
3
4
5
6
7
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
British films
Video games developed in the United States
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
20th-century women singers
21st-century women singers
21st-century American singers
American film actresses
21st-century male singers
20th-century American actresses
American television actresses
21st-century American actresses
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American women writers
American male novelists
19th-century American male writers
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire
Top 50 categories
1.0
0.8
0.6
0.4
0.2
0.0

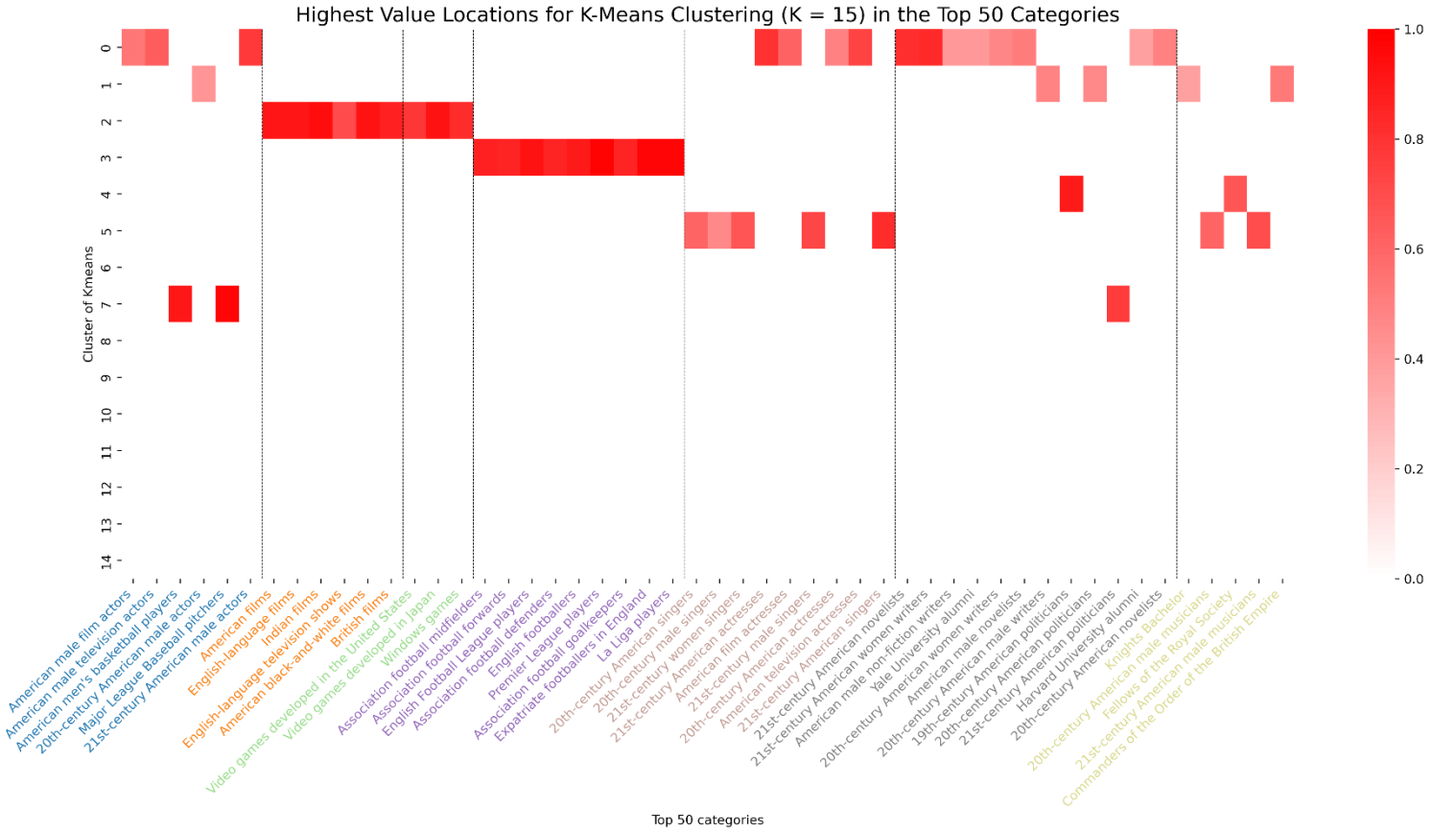

Highest Value Locations for K-Means Clustering (K = 9) in the Top 50 Categories
Cluster of Kmeans
0
1
2
3
4
5
6
7
8
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
British films
Video games developed in the United States
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
20th-century women singers
21st-century women singers
21st-century American singers
American film actresses
21st-century male singers
20th-century American actresses
American television actresses
21st-century American actresses
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American women writers
American male novelists
19th-century American male writers
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire
Top 50 categories
1.0
0.8
0.6
0.4
0.2
0.0

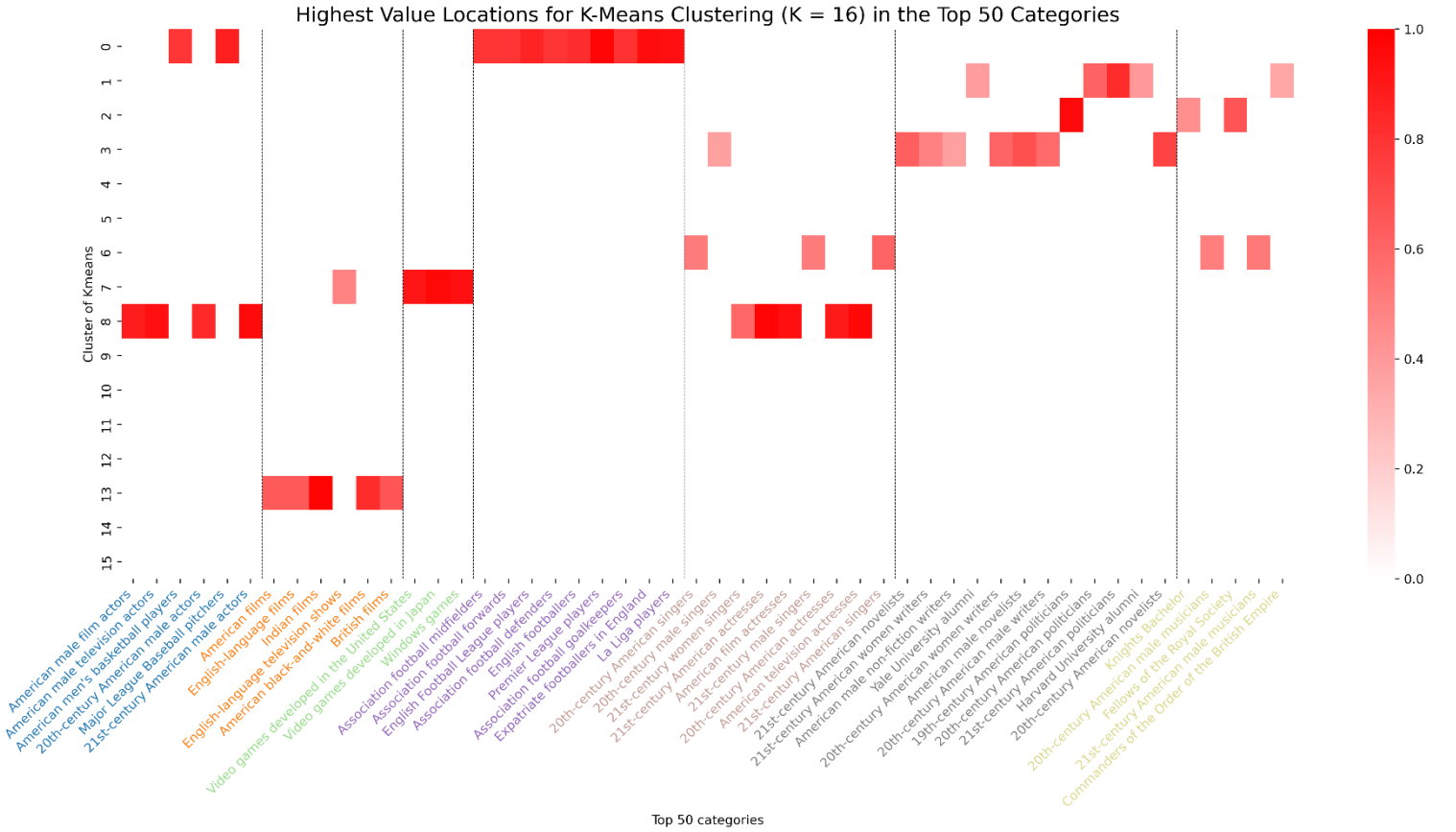

Highest Value Locations for K-Means Clustering (K = 11) in the Top 50 Categories
Cluster of Kmeans
0
1
2
3
4
5
6
7
8
9
10
1.0
0.8
0.6
0.4
0.2
0.0
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
British films
Video games developed in the United States
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
21st-century American singers
21st-century women singers
21st-century male singers
American film actresses
20th-century American actresses
21st-century American actresses
American television actresses
21st-century American singers
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American male novelists
20th-century American male writers
19th-century American politicians
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire
Top 50 categories

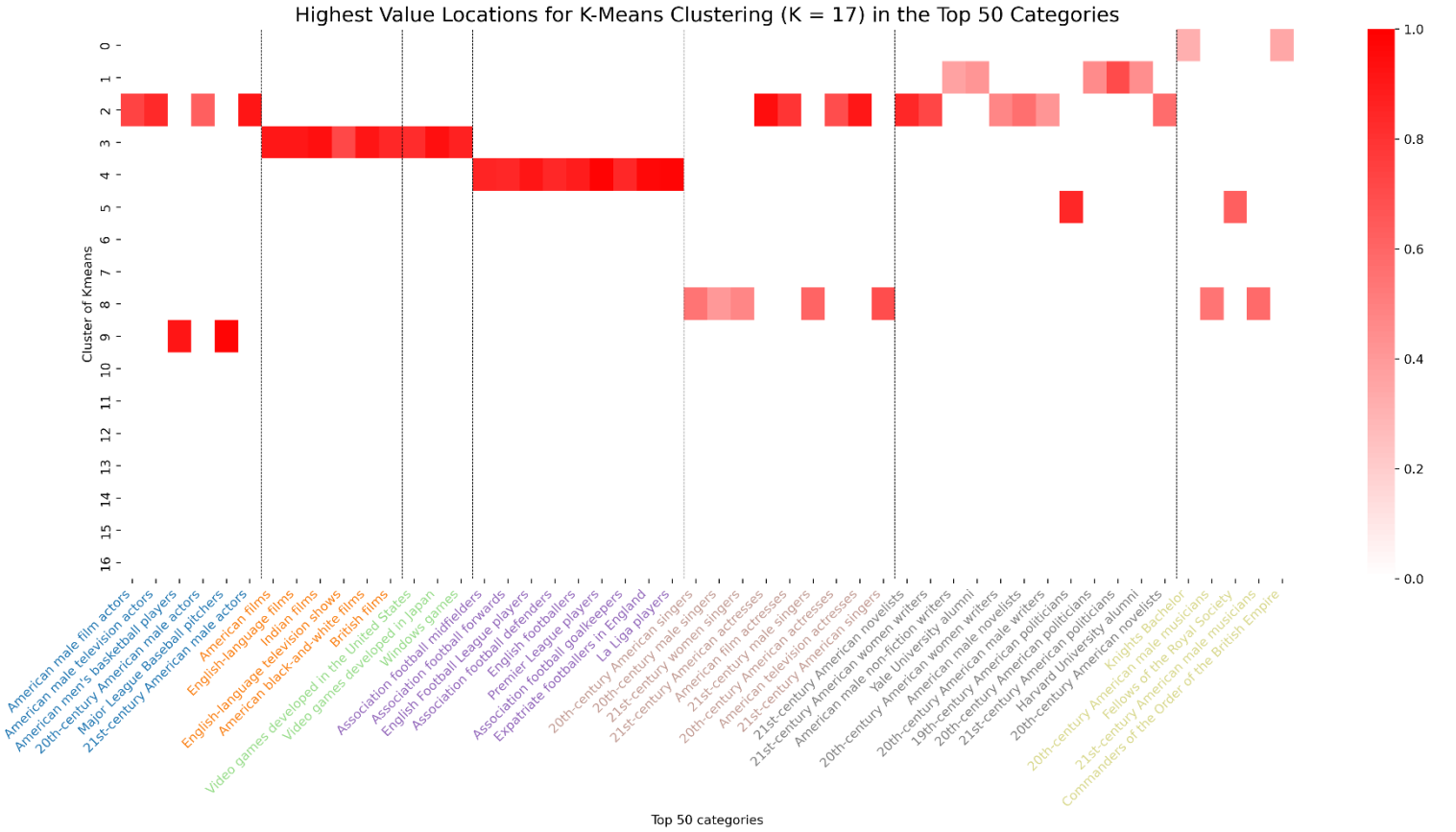

Highest Value Locations for K-Means Clustering (K = 12) in the Top 50 Categories
Cluster of Kmeans
0
1
2
3
4
5
6
7
8
9
10
11
1.0
0.8
0.6
0.4
0.2
0.0
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
British films
Video games developed in the United States
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
21st-century American singers
21st-century women singers
21st-century male singers
American film actresses
20th-century American actresses
21st-century American actresses
American television actresses
21st-century American singers
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American male novelists
20th-century American male writers
19th-century American politicians
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire
Top 50 categories

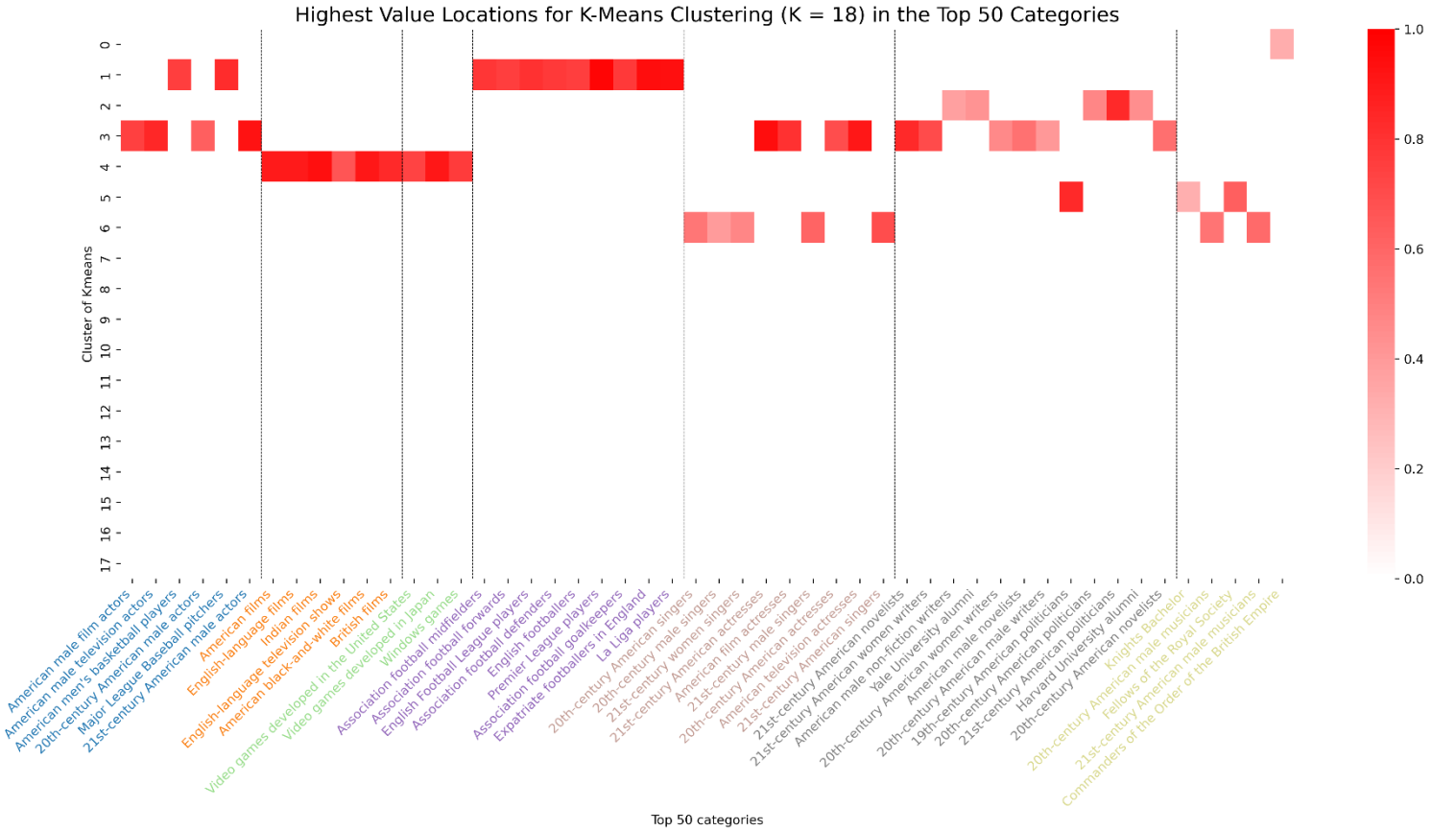

Highest Value Locations for K-Means Clustering (K = 13) in the Top 50 Categories
Cluster of Kmeans
0
1
2
3
4
5
6
7
8
9
10
11
12
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
British films
Video games developed in the United States
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
20th-century women singers
21st-century American singers
21st-century women singers
American film actresses
20th-century American actresses
21st-century American actresses
American television actresses
21st-century American singers
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American male novelists
20th-century American male writers
19th-century American politicians
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire
Top 50 categories
1.0
0.8
0.6
0.4
0.2
0.0

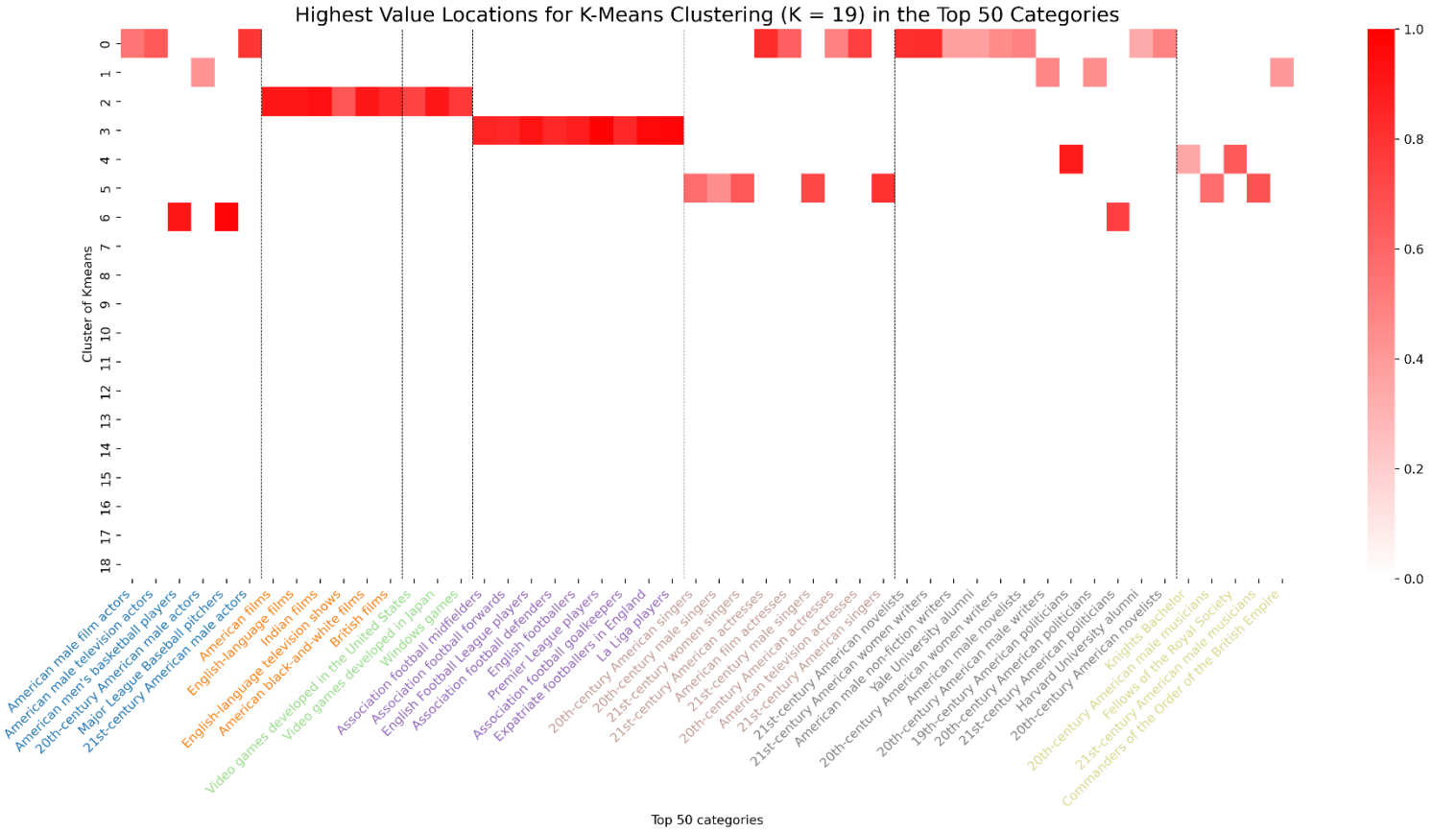

Highest Value Locations for K-Means Clustering (K = 14) in the Top 50 Categories
Cluster of Kmeans
0
1
2
3
4
5
6
7
8
9
10
11
12
13
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
British films
Video games developed in the United States
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
20th-century women singers
21st-century American singers
21st-century women singers
American film actresses
20th-century American actresses
21st-century American actresses
American television actresses
21st-century American singers
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American male novelists
20th-century American male writers
19th-century American politicians
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire
Top 50 categories
1.0
0.8
0.6
0.4
0.2
0.0

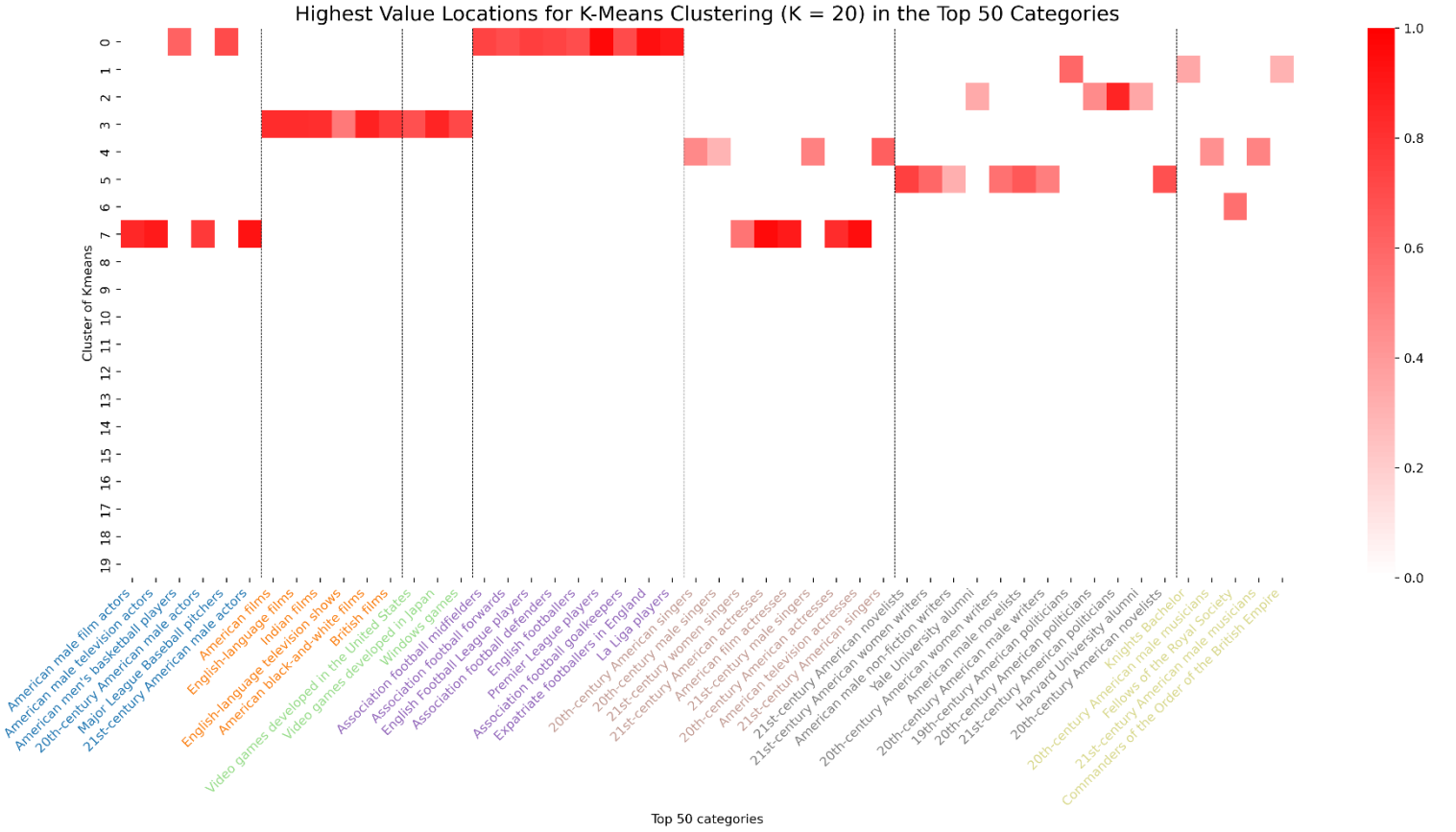

Highest Value Locations for K-Means Clustering (K = 15) in the Top 50 Categories
Cluster of Kmeans
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
1.0
0.8
0.6
0.4
0.2
0.0
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
British films
Video games developed in the United States
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
20th-century male singers
21st-century American singers
21st-century women singers
American film actresses
20th-century American actresses
21st-century male singers
American television actresses
21st-century American actresses
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American male novelists
20th-century American male writers
19th-century American politicians
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire
Top 50 categories

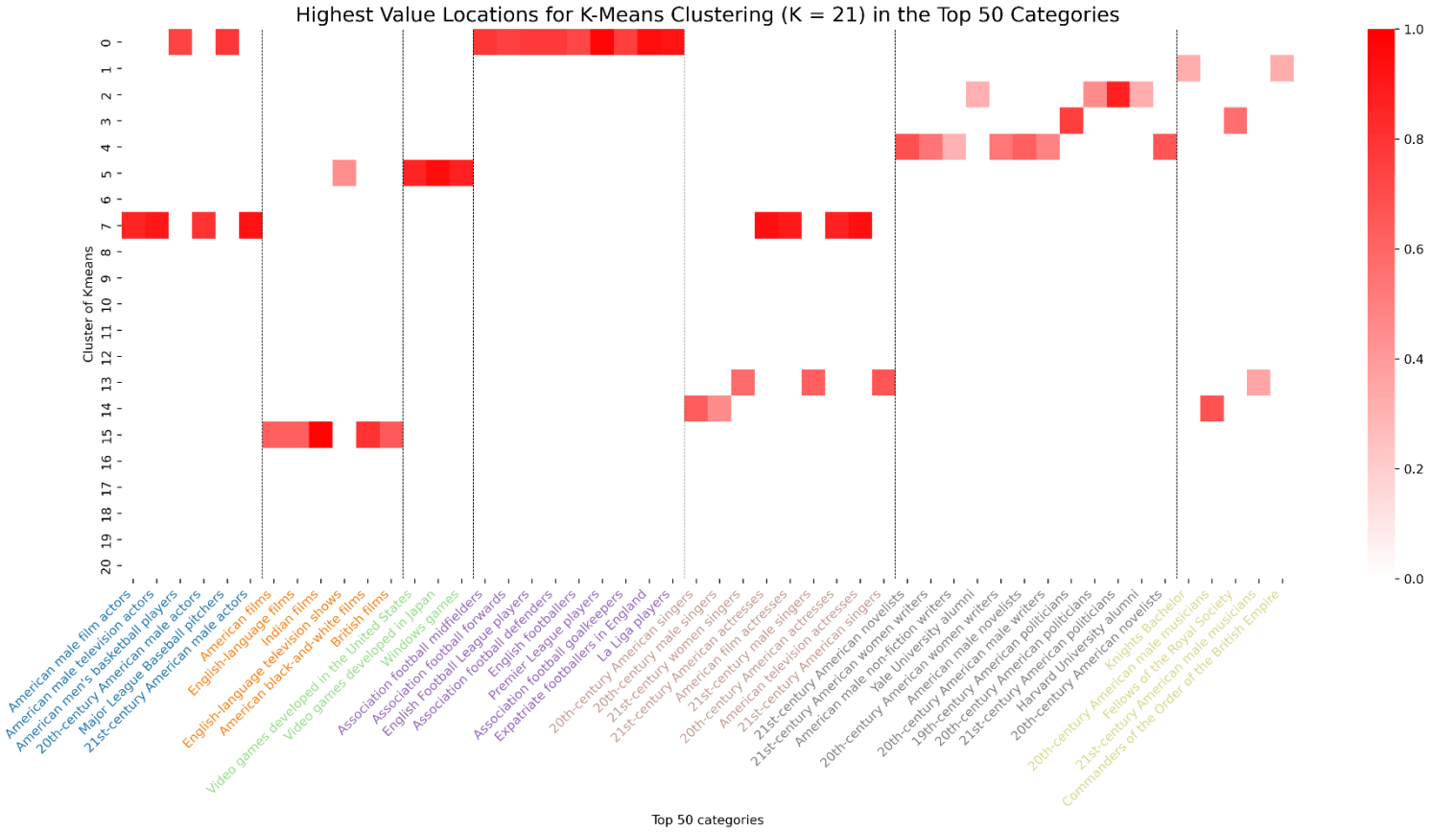

Highest Value Locations for K-Means Clustering (K = 16) in the Top 50 Categories
Cluster of Kmeans
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
1.0
0.8
0.6
0.4
0.2
0.0
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
British films
Video games developed in the United States
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
20th-century male singers
21st-century American singers
21st-century women singers
American film actresses
20th-century American actresses
21st-century male singers
American television actresses
21st-century American actresses
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American male novelists
20th-century American male writers
19th-century American politicians
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire
Top 50 categories

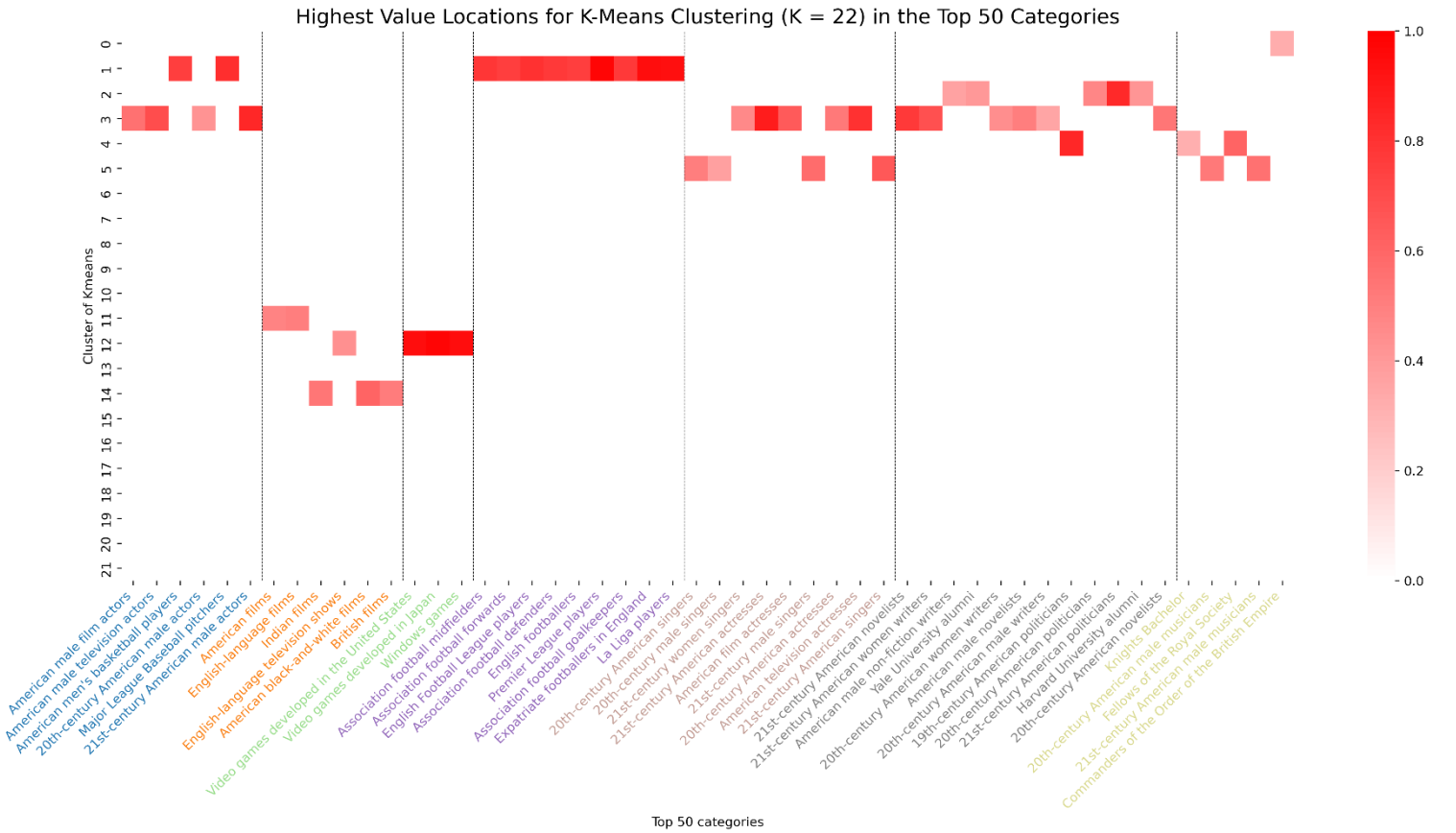

Highest Value Locations for K-Means Clustering (K = 17) in the Top 50 Categories
Cluster of Kmeans
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
British films
Video games developed in the United States
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
Association football League players
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
20th-century male singers
21st-century women singers
21st-century American singers
American film actresses
20th-century American actresses
21st-century male singers
American television actresses
21st-century American actresses
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American male novelists
20th-century American male writers
19th-century American politicians
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire
Top 50 categories
1.0
0.8
0.6
0.4
0.2
0.0

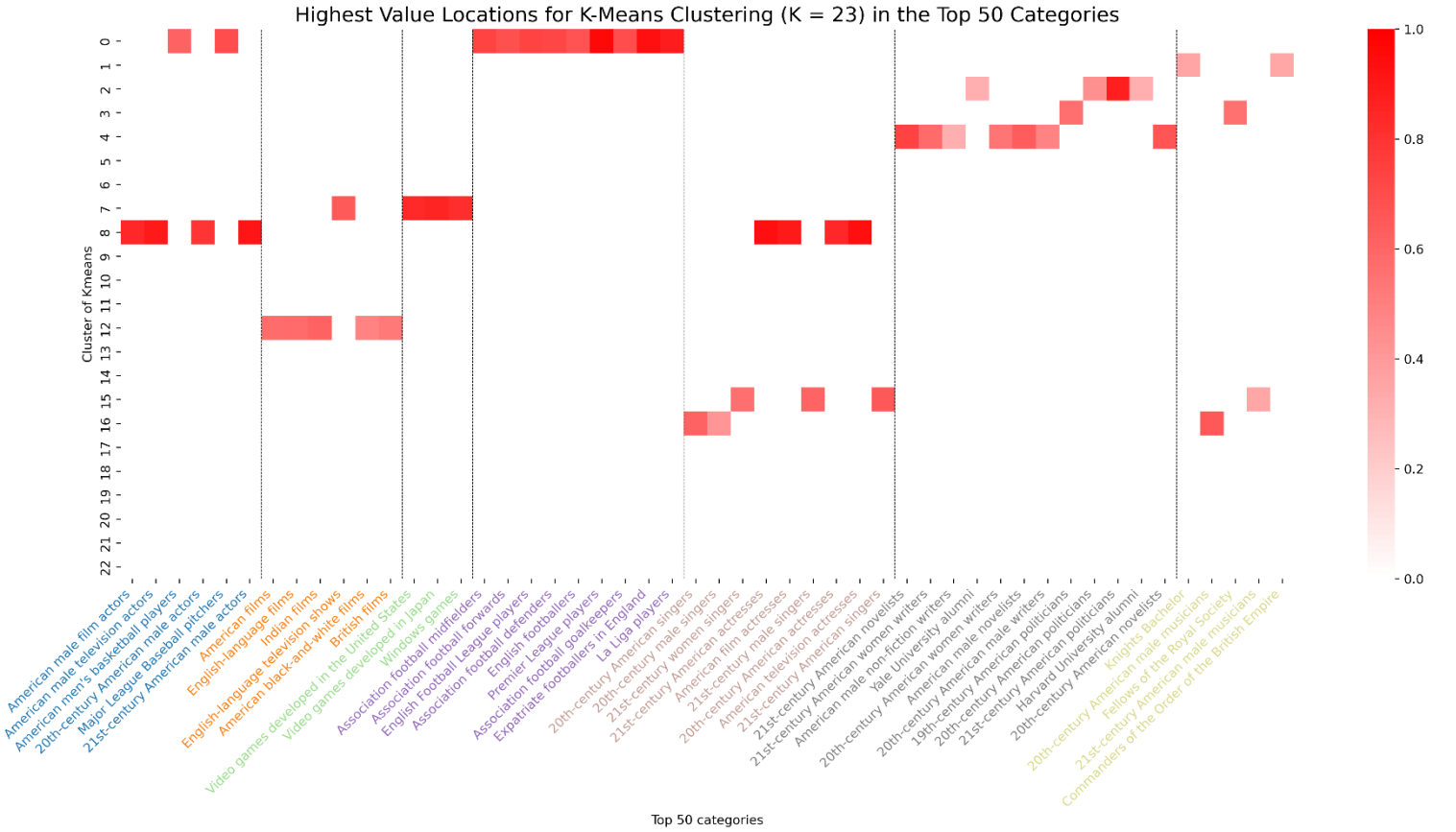

Highest Value Locations for K-Means Clustering (K = 18) in the Top 50 Categories
Cluster of Kmeans
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
British films
Video games developed in the United States
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
Association football League players
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
20th-century male singers
21st-century women singers
21st-century American singers
American film actresses
20th-century American actresses
21st-century male singers
American television actresses
21st-century American actresses
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American male novelists
20th-century American male writers
19th-century American politicians
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire
Top 50 categories
1.0
0.8
0.6
0.4
0.2
0.0

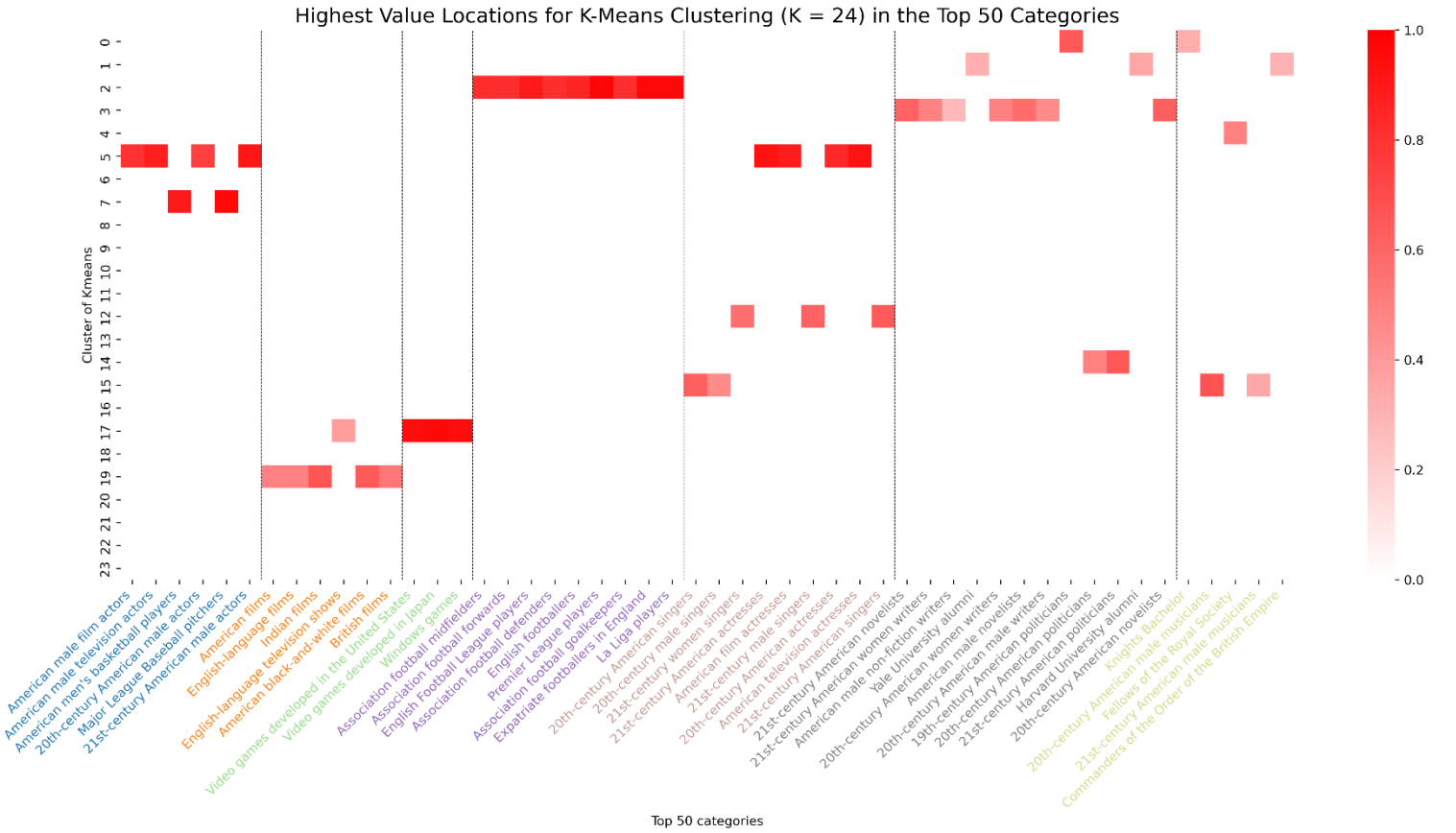

Highest Value Locations for K-Means Clustering (K = 19) in the Top 50 Categories
Cluster of Kmeans
Top 50 categories
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
Video games developed in the United States
British films
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
20th-century women singers
21st-century American singers
21st-century women singers
American film actresses
20th-century American actresses
21st-century American actresses
American television actresses
21st-century American actresses
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American male novelists
20th-century American male writers
19th-century American politicians
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire


Highest Value Locations for K-Means Clustering (K = 20) in the Top 50 Categories
Cluster of Kmeans
Top 50 categories
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
Video games developed in the United States
British films
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
20th-century women singers
21st-century American singers
21st-century women singers
American film actresses
20th-century American actresses
21st-century American actresses
American television actresses
21st-century American actresses
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American male novelists
20th-century American male writers
19th-century American politicians
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire


Highest Value Locations for K-Means Clustering (K = 21) in the Top 50 Categories
Cluster of Kmeans
Top 50 categories
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
British films
Video games developed in the United States
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
20th-century women singers
21st-century women singers
21st-century American actresses
American film actresses
20th-century American actresses
21st-century male singers
American television actresses
21st-century American singers
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American male novelists
20th-century American male writers
19th-century American politicians
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire


Highest Value Locations for K-Means Clustering (K = 22) in the Top 50 Categories
Cluster of Kmeans
Top 50 categories
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
British films
Video games developed in the United States
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
20th-century women singers
21st-century women singers
21st-century American actresses
American film actresses
20th-century American actresses
21st-century male singers
American television actresses
21st-century American singers
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American male novelists
20th-century American male writers
19th-century American politicians
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire


Highest Value Locations for K-Means Clustering (K = 23) in the Top 50 Categories
Cluster of Kmeans
Top 50 categories
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
British films
Video games developed in the United States
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
20th-century women singers
21st-century male singers
21st-century women singers
American film actresses
20th-century American actresses
21st-century American actresses
American television actresses
21st-century American singers
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American male novelists
American male writers
19th-century American politicians
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire


Highest Value Locations for K-Means Clustering (K = 24) in the Top 50 Categories
Cluster of Kmeans
Top 50 categories
American male film actors
American male television actors
American men's basketball players
20th-century American male actors
Major League Baseball pitchers
21st-century American male actors
American films
English-language films
Indian films
English-language television shows
American black-and-white films
British films
Video games developed in the United States
Video games developed in Japan
Windows games
Association football midfielders
Association football forwards
English Football League players
Association football defenders
English footballers
Premier League players
Association football goalkeepers
Expatriate footballers in England
La Liga players
20th-century American singers
20th-century women singers
21st-century male singers
21st-century women singers
American film actresses
20th-century American actresses
21st-century American actresses
American television actresses
21st-century American singers
21st-century American novelists
American male non-fiction writers
20th-century American women writers
Yale University alumni
American male novelists
American male writers
19th-century American politicians
20th-century American politicians
21st-century American politicians
Harvard University alumni
20th-century American novelists
Knights Bachelor
20th-century American male musicians
Fellows of the Royal Society
21st-century American male musicians
Commanders of the Order of the British Empire