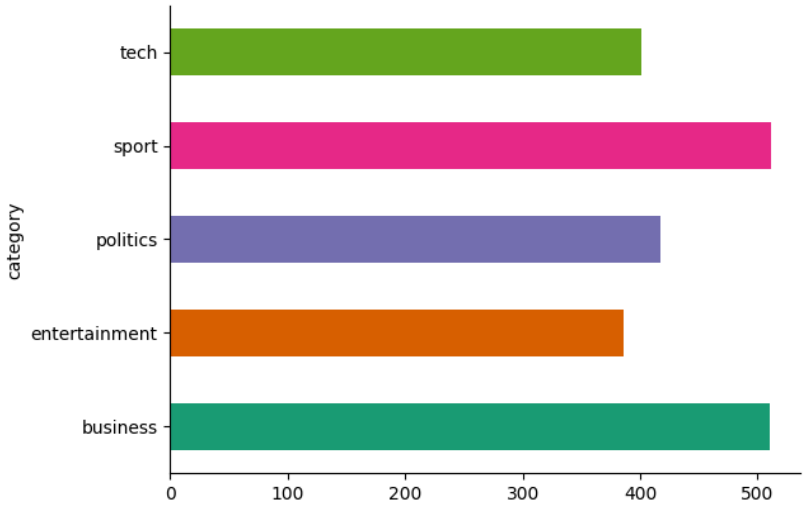
Napassorn Litchiowong (Pleng)
diluxed@gmail.com

Report

BBC Full Text Unsupervised Classification

1. Dataset Selection

Taking all the constraints into account, I have chosen the [BBC Full Text Document Classification] dataset by SHIVAM KUSHWAHA on Kaggle. The dataset contains 2,225 news documents or articles across 5 news categories.



The data is provided as a CSV file with two columns (text, category). However, in the classification pipeline, we will only use the sentence column as we plan to perform unsupervised learning.

The reason why I chose this dataset is based on two main reasons:

1. The task is supposed to be completed in 2–4 hours; therefore, I decided to use a dataset that is reasonably sized. Using open-source data with a logical distribution should be a good starting point for this project, as I do not need to do extensive data cleaning.
2. I chose a news dataset because it provides a fascinating opportunity to explore how machine learning can discover latent topics and thematic groupings within a large body of text. News articles are inherently rich in keywords and contextual information that we can gain insights into how the model "perceives" topics, not only with the same number of clusters as the actual categories, but also with both a broader and a more specific scope than the actual categories (e.g. science & art as border scopes).

Napassorn Litchiowong (Pleng)
diluxed@gmail.com

2. Unsupervised Classification

Before choosing the model or algorithm or designing the pipeline, let's perform some exploratory data analysis (EDA) to get an overall look at the data. And here are the results:

Overall word cloud

After I performed some basic cleaning, such as removing URLs, symbols, lowercasing the text, and removing stop words using NLTK.

From the overall word cloud, we can see that the word “said” is at the top. These also appear alongside common support or common words such as “new”, “one”, “say”,and etc. However, the word u is there which is so weird on how the library does the segmentation.

<u>Algorithm/Techniques Choice</u>

1. TF-IDF and repetitive, unrelated words concern

There was some concern that frequently occurring words might shift the model’s focus. For example, one might ask, “Will these common words cause the model to concentrate on them?” However, in a more advanced, context-aware classification model, such words (like “said” and its variations) can actually carry important contextual information about how news is reported.

In many cases, these words capture the journalist’s narrative or indicate a particular reporting style. Removing them might risk losing key linguistic markers of news tone. Moreover, if a word appears extremely frequently across all classes, it tends to act as a stopword and provides little discriminative power. This issue is effectively managed by TF-IDF, which automatically down-weights very common words. As a result, even if “said” is frequent, it should have a low TF-IDF score if it appears ubiquitously, thus addressing our concern.

Napassorn Litchiowong (Pleng)
diluxed@gmail.com

2. Dimensionality Reduction with Truncated SVD

Given the high-dimensional nature of the TF-IDF representation (even with a limited vocabulary), applying Truncated SVD is essential. This technique not only alleviates memory concerns by reducing the number of dimensions but also helps in capturing the core semantic variance present in the text. By reducing the dimensions to, i.e., 100, we preserve the most informative components, which in turn allows the clustering algorithm to focus on the primary themes in the news articles rather than getting lost in noisy, sparse data.

3. K-means

I then applied K-Means clustering on the reduced features. I experimented with different numbers of clusters—using k = 2 (around half the expected categories), k = 3 (around half the expected categories), k = 5 (the original number of categories), k = 10 (double the expected categories), and k = 15 (triple the expected categories). For each value of k, I computed the clusters and visualized the results using UMAP to project the high-dimensional feature space into two dimensions. This visualization allows us to observe whether the clusters capture coherent groupings of articles (for example, grouping articles that are predominantly about war, guns, or other topics) and how they relate to the actual news categories.

Analysis choice

1. Visualization with UMAP and t-SNE

I chose UMAP and t-SNE for visual interpretation. UMAP often provides a clearer global structure of the data, while t-SNE can highlight local groupings and subtle distinctions. The resulting plots help validate whether the chosen features and clustering approach are able to reveal distinct topic groupings such as clusters that might be predominantly focused on war, politics, or business; thus confirming the effectiveness of our feature engineering process.

2. Inter-Cluster Distance Analysis

A heatmap of pairwise distances between cluster centroids (computed from K-Means) offers additional insight into how distinct the clusters are. Smaller distances between certain clusters might indicate thematic similarities, while larger distances suggest more pronounced differences. This quantitative measure aids in understanding the overall cohesion and separation of the clusters.

3. Network Graph of Cluster Relationships

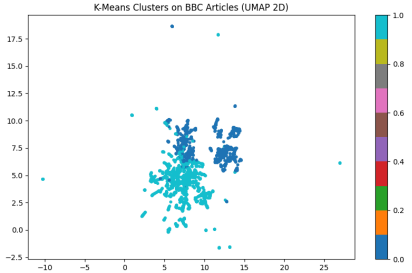
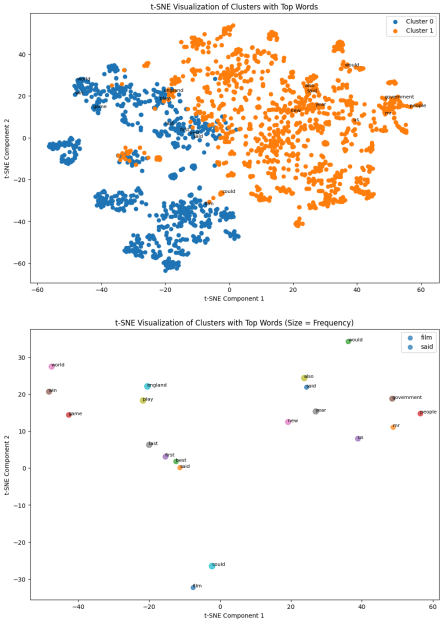
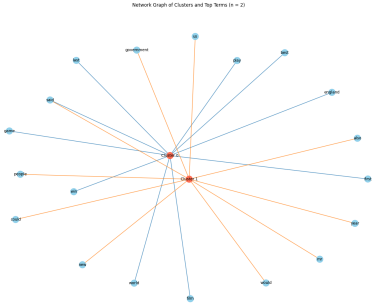
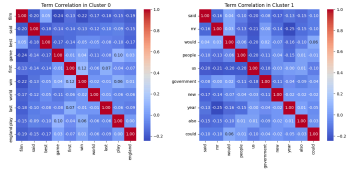
We can gain an intuitive visualization of the relationships between clusters and key vocabulary by constructing a network graph where cluster nodes are connected to their top TF-IDF terms. This not only highlights which words are most representative of each cluster but also how similar or overlapping the clusters are in terms of language use. Additionally, I run a per-cluster term correlation matrix further exploring the internal relationships among top words to see the cohesion of linguistic signals within each cluster.

Napassorn Litchiowong (Pleng)
diluxed@gmail.com

3. Result Analysis

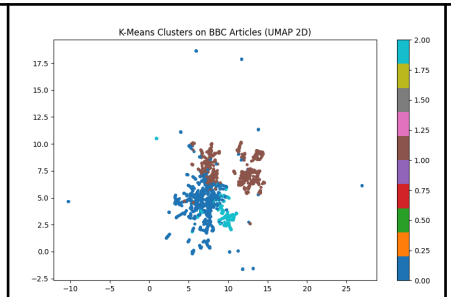
First, let's take a look at visualizations created from each number of k we set during the training process. Then, I will elaborate some important findings and begin the discussion.

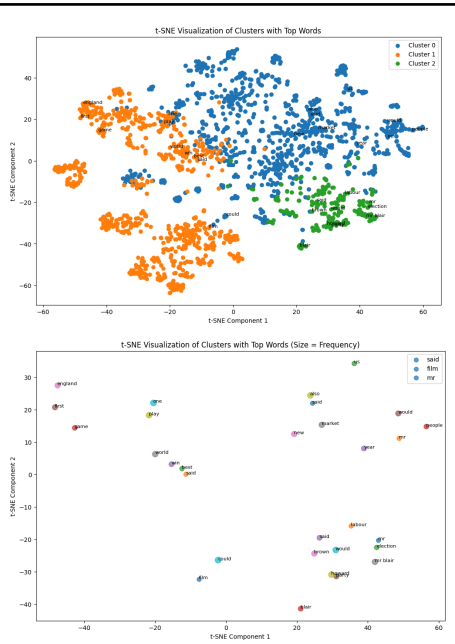
** For a better view of the graphs, you can explore it in the file folder I submitted.

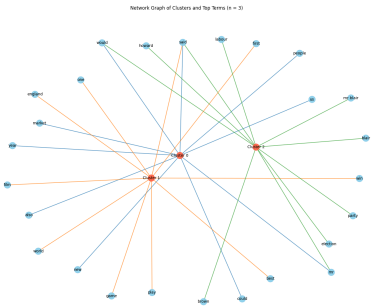
No. K	UMAP 2D	t-SNE with Top Words	Network Graph	Term Correlation
2				

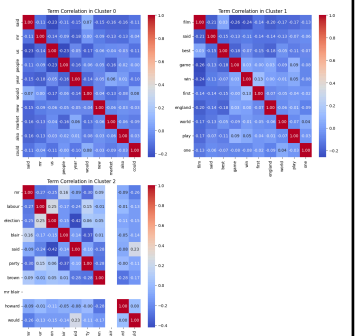
Napassorn Litchiowong (Pleng)
diluxed@gmail.com

3



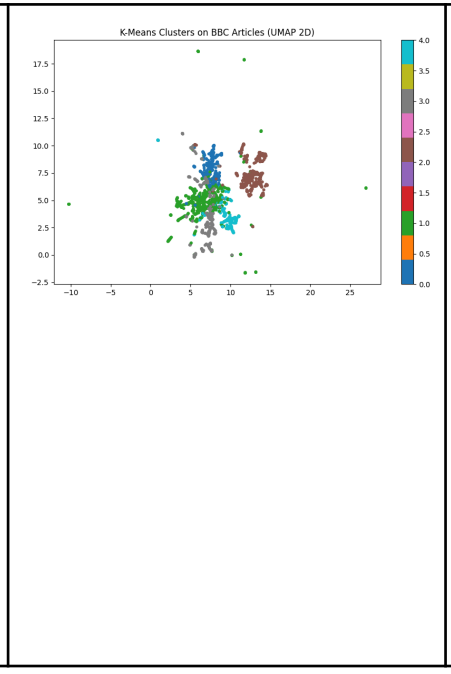


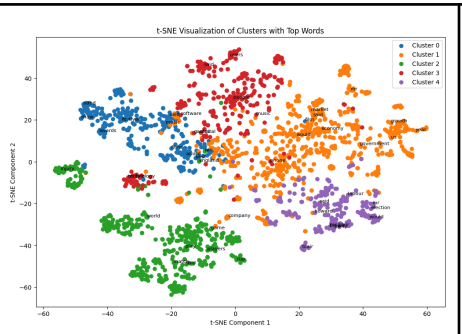


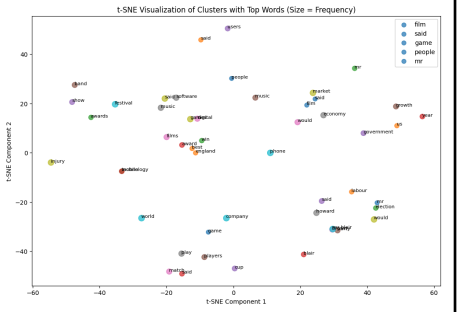


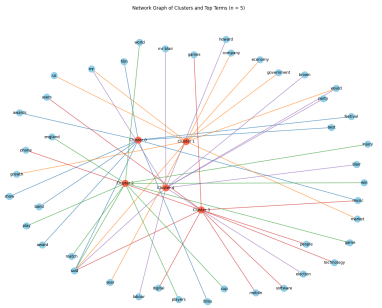
Napassorn Litchiowong (Pleng)
diluxed@gmail.com

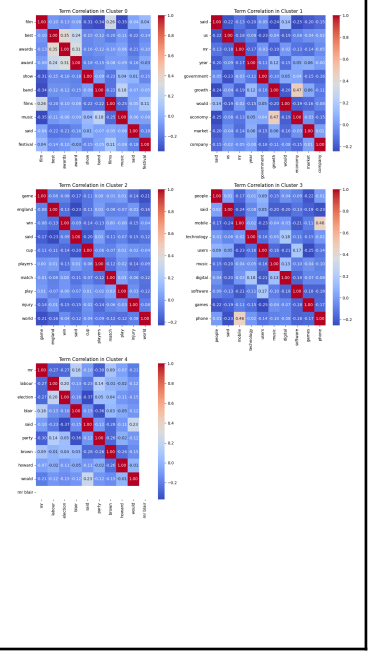
5





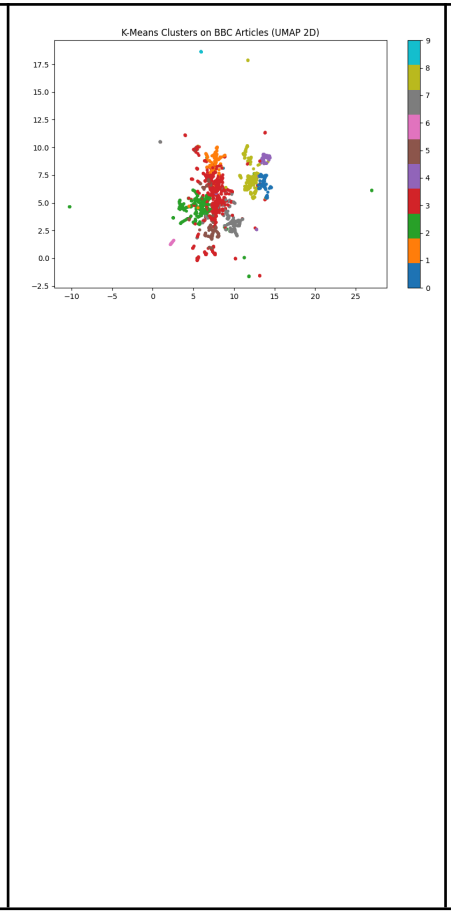


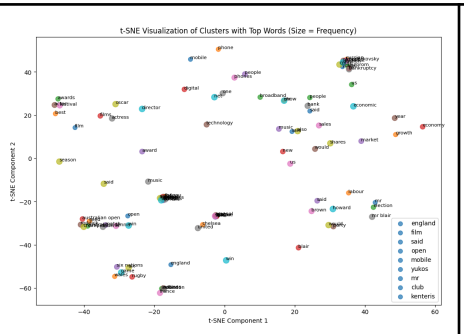


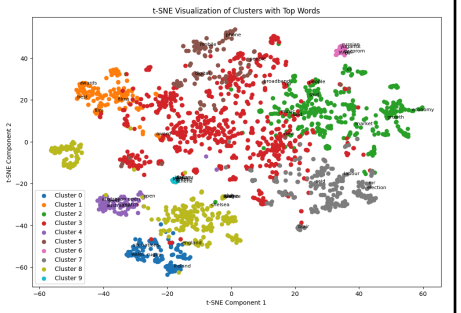


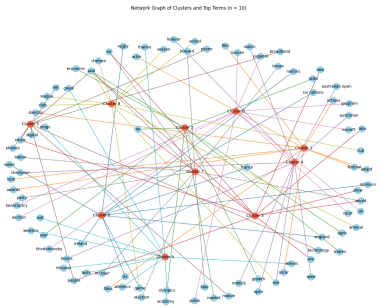
Napassorn Litchiowong (Pleng)
diluxed@gmail.com

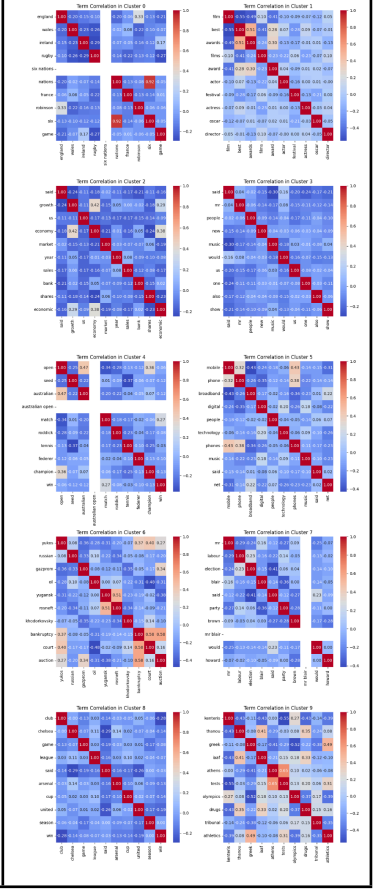
10





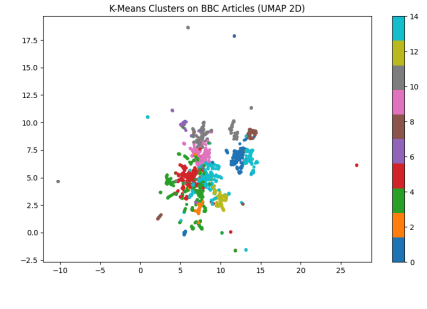


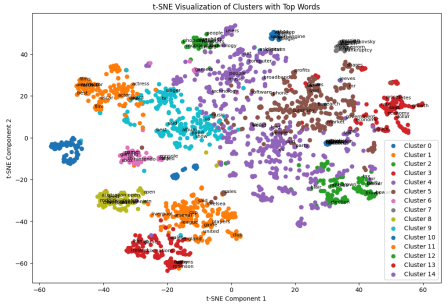


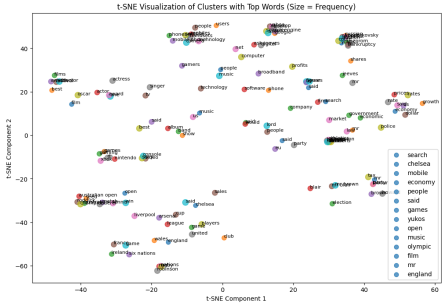


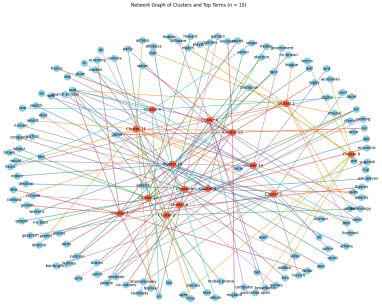
Napassorn Litchiowong (Pleng)
diluxed@gmail.com

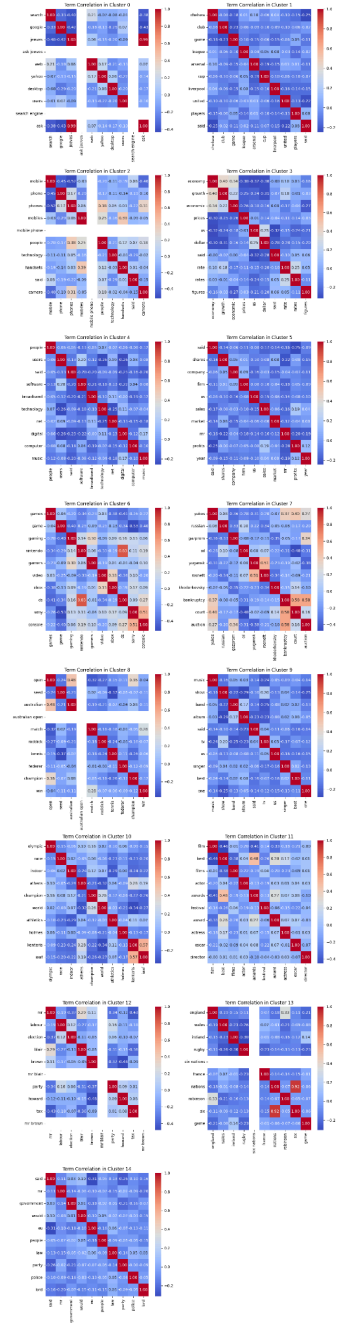
15











Napassorn Litchiowong (Pleng)
diluxed@gmail.com

Overall Analysis

From the clustering results and visualizations, we found that at a higher-level scope (i.e., k = 2), the model roughly splits the data into two broad groups. I would name these groups “entertainment” and “business.” The “entertainment” group is characterized by top words related to sports, film, or other entertainment topics, whereas the “business” group is on the opposite side, with top words such as government, people, or “us” (derived from the U.S.).

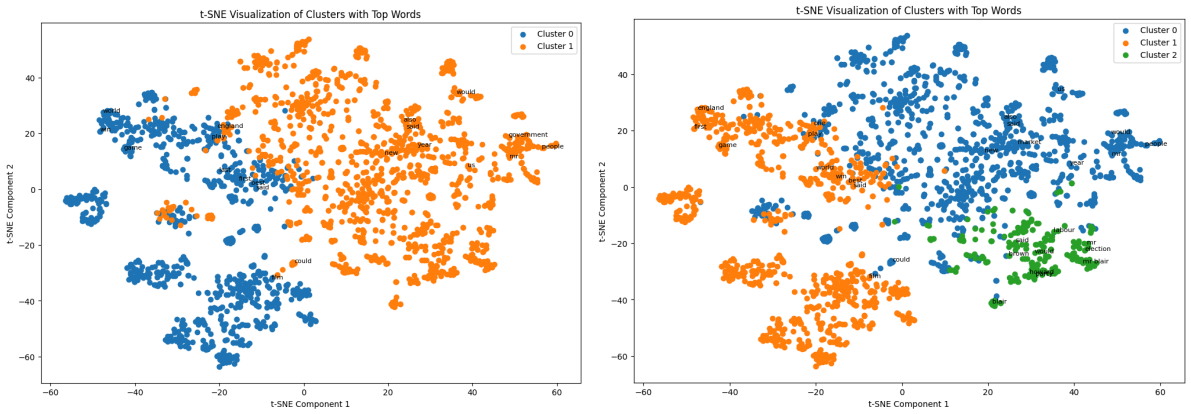
However, when we delve into a lower, more granular scope—such as when k = 15—the clusters split into smaller, more specific groups, while the clusters corresponding to a larger umbrella (e.g., entertainment-related groups) remain near each other. For example, there are two sports clusters located next to each other on the bottom left of the k = 15 t-SNE top words graph. It makes sense that both sports clusters are situated near each other because they pertain to sports. But how do they differ? One cluster’s top words include “open” alongside country-related terms like tennis, Australian, win, and champion; we can interpret this as mostly tennis-related news, with the word “open” referring to tournaments such as the Australian Open (one of the biggest Grand Slam events). In contrast, the other sports cluster is clearly mostly football-related, particularly about the Premier League, based on words like Liverpool, Chelsea, club, and England. These two mainstream sports exhibit distinct patterns, which is how the model distinguishes between them.

In conclusion, this analysis demonstrates that the unsupervised clustering approach effectively captures both broad and fine-grained thematic distinctions within the news data. At higher levels, the clusters reflect major domains, while at lower levels, the model distinguishes specific subtopics based on nuanced language differences. Technically, this method leverages state-of-the-art feature extraction and dimensionality reduction techniques to produce interpretable and actionable insights, underscoring its potential for real-world applications in automated news categorization.

Other Key findings

1. K = 2 vs K = 3

One question that arose during the analysis is: when k differs by only one (for example, comparing k = 2 versus k = 3), what significant insights can we derive?

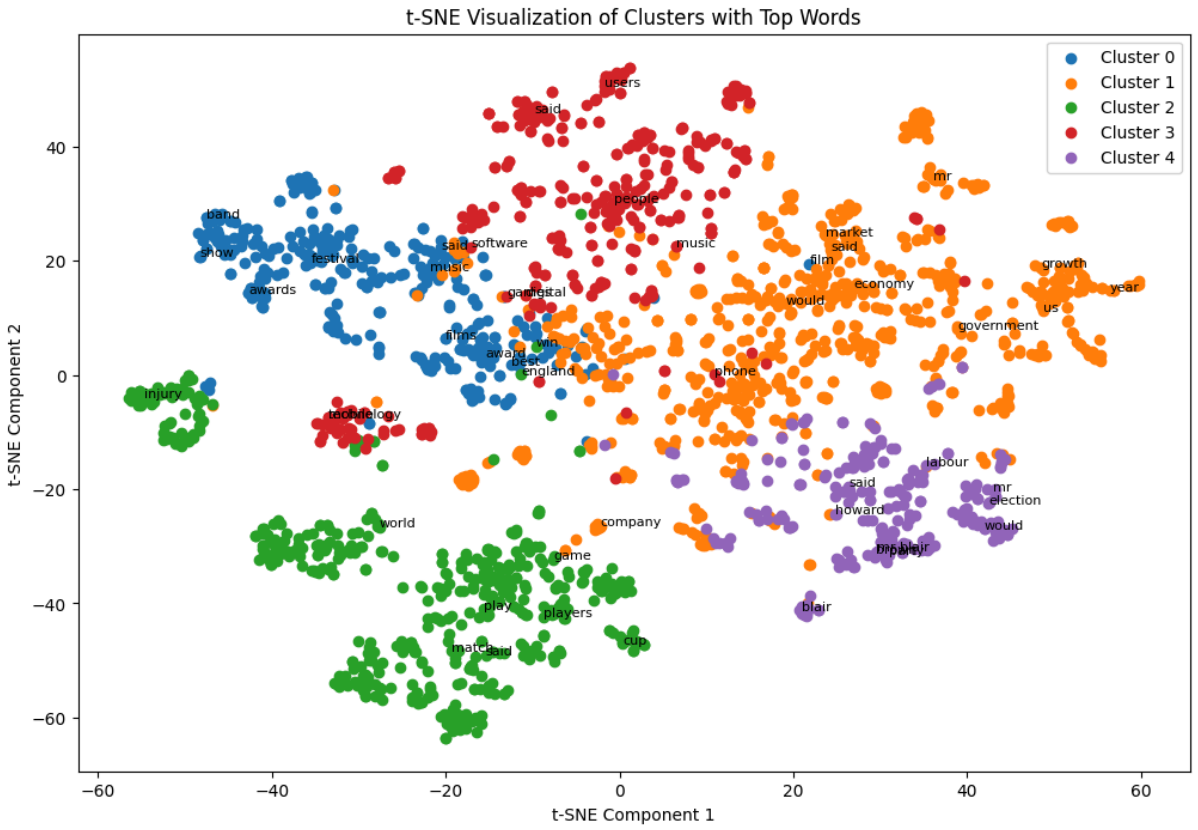


First, when examining the t-SNE graph (left: k = 2 and right: k = 3), we see that the left cluster (the “entertainment” cluster) remains unchanged. However, when k = 3, a new third cluster emerges from a split within the “business” group. A closer look reveals that, based on the colors in the k = 3 graph, the blue cluster displays top words such as “market,” “year,” and “the U.S.,” suggesting that this cluster represents news related to the economy and markets. Meanwhile, the green cluster exhibits top words like “election,” “labour,” “party,” and even names associated with renowned politicians (for instance, “Blair,” likely referring to Tony Blair, the former UK Prime Minister, and “Howard,” referring to Michael Howard, a former British political leader). I would label this new group “government.”

What can we derive from this? It indicates that the “government” group is distinct enough that, with a slight increase in k, it separates from the larger “business” umbrella—rather than, for example, the “entertainment” group splitting into subgroups like sports and film. This suggests that within the business domain, there is a strong thematic distinction that allows the model to differentiate between economic/market news and political (government) news, whereas in the entertainment domain, there is more overlapping vocabulary, which prevents further splitting.

2. K is set equal to the original number of labels

When k is set equal to the original number of labels, we can directly compare the unsupervised clusters with our intuitive categorization to evaluate whether the model "thinks" like us. For instance, look at the clusters:



Napassorn Litchiowong (Pleng)
diluxed@gmail.com

Cluster 0: ['film', 'best', 'awards', 'award', 'show', 'band', 'films', 'music', 'said', 'festival']

- This cluster clearly aligns with the entertainment domain, particularly focused on cinema and related events.

Cluster 1: ['said', 'us', 'mr', 'year', 'government', 'growth', 'would', 'economy', 'market', 'company']

- Here, the dominant terms such as "government," "economy," and "market" indicate a cluster focused on business and economic news.

Cluster 2: ['game', 'england', 'win', 'said', 'cup', 'players', 'match', 'play', 'injury', 'world']

- The presence of words like "game," "win," and "cup" suggests this cluster is centered on sports news, likely related to football or other competitive events.

Cluster 3: ['people', 'said', 'mobile', 'technology', 'users', 'music', 'digital', 'software', 'games', 'phone']

- This cluster is driven by technological terms such as "mobile," "technology," and "software," pointing to news related to technology and digital innovation.

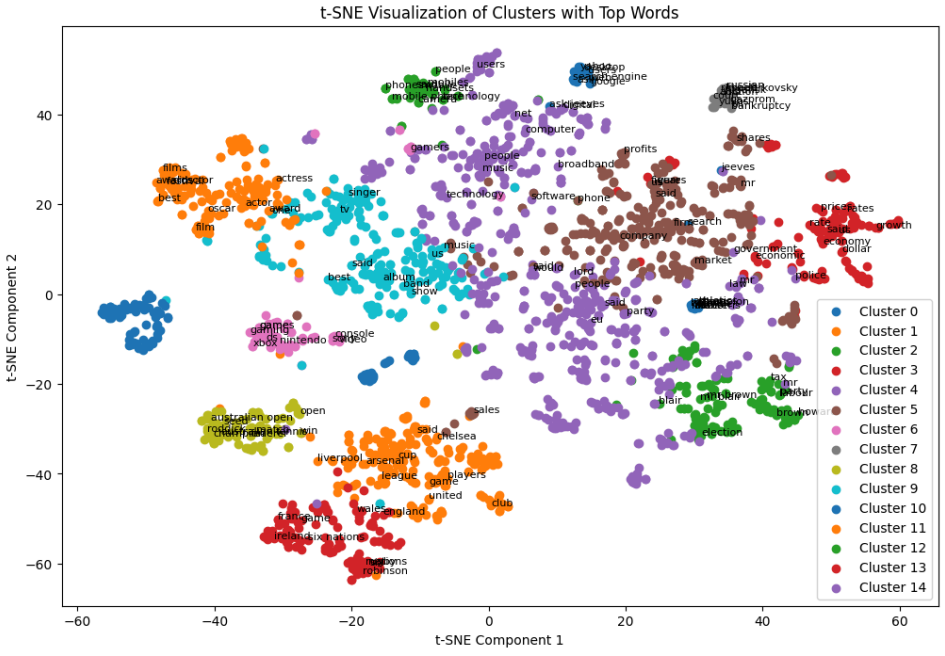
Cluster 4: ['mr', 'labour', 'election', 'blair', 'said', 'party', 'brown', 'howard', 'would', 'mr blair']

- The top words here clearly indicate a focus on politics, with references to political parties and prominent political figures.

Overall, these clusters demonstrate that the unsupervised model is able to uncover themes that are consistent with human categorization. The entertainment, business, sports, technology, and political domains emerge naturally from the data, showing that the model captures key thematic distinctions in a manner that aligns with our expectations.

3. The Bigger K

From the clustering results and visualizations for k = 10 and k = 15, we observe that the clusters split into more granular, specific groups, revealing subtle thematic differences. For example, in k = 15:



Napassorn Litchiowong (Pleng)
diluxed@gmail.com

Cluster 0: ['search', 'google', 'jeeves', 'ask jeeves', 'web', 'yahoo', 'desktop', 'users', 'search engine', 'ask']
- This cluster appears to capture articles related to search engines and internet usage.

Cluster 1: ['chelsea', 'club', 'game', 'league', 'arsenal', 'cup', 'liverpool', 'united', 'players', 'said']
- This cluster is clearly focused on football.

Cluster 2: ['mobile', 'phone', 'phones', 'mobiles', 'mobile phone', 'people', 'technology', 'handsets', 'said', 'camera']
- This cluster is mobile technology.

Cluster 3: ['economy', 'growth', 'economic', 'prices', 'us', 'dollar', 'said', 'rate', 'rates', 'figures']
- This cluster indicates a focus on business and financial news.

Cluster 4: ['people', 'users', 'said', 'software', 'broadband', 'technology', 'net', 'digital', 'computer', 'music']
- This cluster suggests a blend of technology, network, and digital cultural topics.

Cluster 5: ['said', 'shares', 'company', 'firm', 'us', 'sales', 'market', 'mr', 'profits', 'year']
- This cluster focuses on market performance

Cluster 6: ['games', 'game', 'gaming', 'nintendo', 'gamers', 'video', 'xbox', 'ds', 'sony', 'console']
- This cluster focuses on gaming.

Cluster 7: ['yukos', 'russian', 'gazprom', 'oil', 'yugansk', 'rosneft', 'khodorkovsky', 'bankruptcy', 'court', 'auction']
- This cluster clearly represents Russian economic and energy news

Cluster 8: ['open', 'seed', 'australian', 'australian open', 'match', 'roddick', 'tennis', 'federer', 'champion', 'win']
- This cluster points to tennis-related coverage.

Cluster 9: ['music', 'show', 'band', 'album', 'said', 'tv', 'us', 'singer', 'best', 'one']
- This cluster clearly distinguishes music-related news.

Cluster 10: ['olympic', 'race', 'indoor', 'athens', 'champion', 'world', 'athletics', 'holmes', 'kenteris', 'iaaf']
- This cluster suggests a focus on Olympic and athletics coverage

Cluster 11: ['film', 'best', 'films', 'actor', 'awards', 'festival', 'award', 'actress', 'oscar', 'director']
- This cluster is clearly about film-related topics.

Cluster 12: ['mr', 'labour', 'election', 'blair', 'brown', 'mr blair', 'party', 'howard', 'tax', 'mr brown']
- This cluster represents political news, specifically in the UK.

Cluster 13: ['england', 'wales', 'ireland', 'rugby', 'six nations', 'france', 'nations', 'robinson', 'six', 'game']
- This cluster clearly focuses on rugby, particularly the Six Nations Championship.

Cluster 14: ['said', 'mr', 'government', 'would', 'eu', 'people', 'law', 'party', 'police', 'lord']
- This cluster appears to cover general political and governmental topics.

Napassorn Litchiowong (Pleng)
diluxed@gmail.com

Some Key Insights

- **Granular Sports Clusters**

Within the sports domain, the model is able to differentiate distinct subtopics. For example, one sports cluster (Cluster 8) focuses on tennis—evidenced by terms such as "open," "australian open," "match," "tennis," "federer," and "champion." Meanwhile, Cluster 1 groups football-related news with terms like "chelsea," "club," "game," "league," and "arsenal." Additionally, Cluster 13 captures rugby news, primarily with terms like "england," "wales," "ireland," and "rugby." Interestingly, although the tennis, football, and rugby clusters are positioned close to each other in the visualization (indicating some shared vocabulary or thematic overlap), the Olympic cluster (Cluster 10) is located among the political groups. This suggests that Olympic coverage might share more common ground with political discourse, possibly due to discussions of national representation, funding, or geopolitical implications.

- **Cohesiveness within Specific Clusters**

Clusters such as Cluster 7 (focused on Russian energy and economic news) exhibit very tight grouping in the visualizations, indicating that the vocabulary used in these articles is highly consistent and distinct. Similarly, the Olympic cluster (Cluster 10) also shows strong cohesion, reflecting a well-defined topic with specialized terminology.

- **Distinct Entertainment Categories**

It is noteworthy that the model distinctly separates film (Cluster 11) from music (Cluster 9). Although both fall under the broader entertainment umbrella, their unique vocabularies are clearly delineated, demonstrating that the model can capture fine-grained semantic differences even within closely related domains.

Napassorn Litchiowong (Pleng)
diluxed@gmail.com

Suggestion for current approach

Although the current pipeline utilizes NLTK's tokenization and basic cleaning, there is room for enhancement in sentence segmentation. While basic tokenization with NLTK is effective, incorporating a more advanced sentence segmentation tool such as spaCy or HuggingFace's tokenizers could further improve text segmentation accuracy, especially for news articles that may include informal or conversational content. This improvement would help ensure that each sentence is correctly identified, leading to more precise feature extraction.

In addition, expanding the pipeline to extract more linguistic features (for example, bigrams, dependency parse features, and detailed part-of-speech patterns) could enable us to explore more nuanced groupings within the data. These extra features might reveal subtle distinctions in how news topics are reported, such as differences between political commentary and business analysis.

It is also important to note that some common words such as "said" still appear frequently across many clusters. This may indicate that these words, despite their ubiquity, might not add significant discriminative value or could even skew the clustering process. Therefore, it would be prudent to double-check whether such tokens are providing meaningful insights or if they should be removed or further down-weighted in the analysis.

Suggestion as an Alternative Approach

An interesting alternative approach would be to combine DEC (Deep Embedded Clustering) with BERT-based embeddings for semantic representation. Using BERT would allow us to capture rich, context-aware features from the text, while DEC could learn a cluster-friendly latent space by jointly optimizing the representation and the cluster assignments. Another alternative is to employ an LLM-based zero-shot classification approach, where the LLM is prompted with the text without explicit category guidance, letting it infer categories based solely on its pre-trained knowledge. Although this method might yield clusters that capture subtle thematic nuances, it may trade off some interpretability and computational efficiency compared to traditional methods.

Napassorn Litchiowong (Pleng)
diluxed@gmail.com

Appendix

Top words

K = 2

Cluster 0: ['film', 'said', 'best', 'game', 'first', 'win', 'world', 'last', 'play', 'england']
Cluster 1: ['said', 'mr', 'would', 'people', 'us', 'government', 'new', 'year', 'also', 'could']

K = 3

Cluster 0: ['said', 'mr', 'us', 'people', 'year', 'would', 'new', 'market', 'also', 'could']
Cluster 1: ['film', 'said', 'best', 'game', 'win', 'first', 'england', 'world', 'play', 'one']
Cluster 2: ['mr', 'labour', 'election', 'blair', 'said', 'party', 'brown', 'mr blair', 'howard', 'would']

K = 5

Cluster 0: ['film', 'best', 'awards', 'award', 'show', 'band', 'films', 'music', 'said', 'festival']
Cluster 1: ['said', 'us', 'mr', 'year', 'government', 'growth', 'would', 'economy', 'market', 'company']
Cluster 2: ['game', 'england', 'win', 'said', 'cup', 'players', 'match', 'play', 'injury', 'world']
Cluster 3: ['people', 'said', 'mobile', 'technology', 'users', 'music', 'digital', 'software', 'games', 'phone']
Cluster 4: ['mr', 'labour', 'election', 'blair', 'said', 'party', 'brown', 'howard', 'would', 'mr blair']

K = 10

Cluster 0: ['england', 'wales', 'ireland', 'rugby', 'six nations', 'nations', 'france', 'robinson', 'six', 'game']
Cluster 1: ['film', 'best', 'awards', 'films', 'award', 'actor', 'festival', 'actress', 'oscar', 'director']
Cluster 2: ['said', 'growth', 'us', 'economy', 'market', 'year', 'sales', 'bank', 'shares', 'economic']
Cluster 3: ['said', 'mr', 'people', 'new', 'music', 'would', 'us', 'one', 'also', 'show']
Cluster 4: ['open', 'seed', 'australian', 'australian open', 'match', 'roddick', 'tennis', 'federer', 'champion', 'win']
Cluster 5: ['mobile', 'phone', 'broadband', 'digital', 'people', 'technology', 'phones', 'music', 'said', 'net']
Cluster 6: ['yukos', 'russian', 'gazprom', 'oil', 'yugansk', 'rosneft', 'khodorkovsky', 'bankruptcy', 'court', 'auction']
Cluster 7: ['mr', 'labour', 'election', 'blair', 'said', 'party', 'brown', 'mr blair', 'would', 'howard']
Cluster 8: ['club', 'chelsea', 'game', 'league', 'said', 'arsenal', 'cup', 'united', 'season', 'win']
Cluster 9: ['kenteris', 'thanou', 'greek', 'iaaf', 'athens', 'tests', 'olympics', 'drugs', 'tribunal', 'athletics']

Napassorn Litchiowong (Pleng)
diluxed@gmail.com

K = 15

Cluster 0: ['search', 'google', 'jeeves', 'ask jeeves', 'web', 'yahoo', 'desktop', 'users', 'search engine', 'ask']
Cluster 1: ['chelsea', 'club', 'game', 'league', 'arsenal', 'cup', 'liverpool', 'united', 'players', 'said']
Cluster 2: ['mobile', 'phone', 'phones', 'mobiles', 'mobile phone', 'people', 'technology', 'handsets', 'said', 'camera']
Cluster 3: ['economy', 'growth', 'economic', 'prices', 'us', 'dollar', 'said', 'rate', 'rates', 'figures']
Cluster 4: ['people', 'users', 'said', 'software', 'broadband', 'technology', 'net', 'digital', 'computer', 'music']
Cluster 5: ['said', 'shares', 'company', 'firm', 'us', 'sales', 'market', 'mr', 'profits', 'year']
Cluster 6: ['games', 'game', 'gaming', 'nintendo', 'gamers', 'video', 'xbox', 'ds', 'sony', 'console']
Cluster 7: ['yukos', 'russian', 'gazprom', 'oil', 'yugansk', 'rosneft', 'khodorkovsky', 'bankruptcy', 'court', 'auction']
Cluster 8: ['open', 'seed', 'australian', 'australian open', 'match', 'roddick', 'tennis', 'federer', 'champion', 'win']
Cluster 9: ['music', 'show', 'band', 'album', 'said', 'tv', 'us', 'singer', 'best', 'one']
Cluster 10: ['olympic', 'race', 'indoor', 'athens', 'champion', 'world', 'athletics', 'holmes', 'kenteris', 'iaaf']
Cluster 11: ['film', 'best', 'films', 'actor', 'awards', 'festival', 'award', 'actress', 'oscar', 'director']
Cluster 12: ['mr', 'labour', 'election', 'blair', 'brown', 'mr blair', 'party', 'howard', 'tax', 'mr brown']
Cluster 13: ['england', 'wales', 'ireland', 'rugby', 'six nations', 'france', 'nations', 'robinson', 'six', 'game']
Cluster 14: ['said', 'mr', 'government', 'would', 'eu', 'people', 'law', 'party', 'police', 'lord']