**Case Study 2**

Tobias Chin, Ryan Dalton, Jacob Ruppert, Rohan Prasad, and Bishoy Soliman Hanna

**1. Introduction:**

Leveraging the News API, we collated a significant dataset of ESPN articles, capturing pivotal moments in sports journalism from November 11th to 24th. This period, bracketing Thanksgiving, included 681 articles from the week before and 568 during the week of the holiday. We meticulously selected 100 articles from each phase for in-depth analysis in MongoDB. Our objective is to distill these articles to understand prevailing themes and reader interests. This analysis is aimed at empowering advertisers with strategic insights, guiding them to focus on specific demographics. Ultimately, this will enhance their advertising efficacy in terms of reach and impact, particularly during a culturally and commercially significant time.

**2. Motivation:**

Our team chose to focus on ESPN articles during the Thanksgiving period, recognizing the deep-rooted significance of sports, particularly football, in American culture. This time of year, when family and sports traditions intertwine, presents a unique opportunity to explore how content shapes and reflects public interests. ESPN, with its vast readership, stands at the forefront of sports journalism, where the crafting of article titles plays a crucial role in captivating millions. These titles are more than just words; they are strategically designed hooks that not only draw readers in but also encapsulate the essence of the stories.

Moreover, ESPN's coverage extends beyond mere reporting – it serves as a potent promotional tool for the NFL. The articles play a pivotal role in building excitement for upcoming games, creating a buzz that transcends the sports arena. They influence fan engagement, drive merchandise sales, and enhance the overall fan experience. By delving into this rich trove of content, we aim to unravel the nuances of how ESPN articles resonate with and galvanize their audience, particularly during a season synonymous with heightened emotions and communal spirit.

**3. Analyzing the Data:**

**Article Counts Analysis:**

Article Counts Analysis serves as a barometer for the volume and distribution of content across different periods or categories. It is a quantitative approach to measure the amount of published material on various topics, which can reflect public interest, media focus, or the prominence of certain themes.

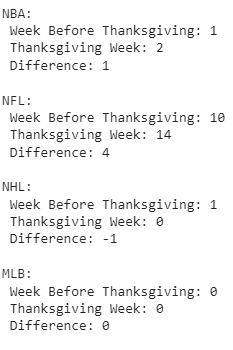
Algorithm & Explanation:

* The process involves collecting metadata from a set of articles, such as publication date, title, and content source.
* Articles are then categorized based on predefined criteria, which could include topics, events, or any relevant taxonomy.
* Counting the articles in each category yields a quantitative understanding of the distribution of content.

Application in Project:

In our advertising-focused project, Article Counts Analysis was pivotal in two ways:

* Temporal Analysis: We compared the number of articles published during the week of Thanksgiving to the week prior. This temporal view helped us understand shifts in content volume related to the holiday season, which is crucial for timing marketing campaigns.
* Thematic Analysis: By examining the counts of articles per theme or subject, we identified what topics were most covered. This insight allowed us to tailor our advertising content to align with the most discussed subjects, ensuring higher relevance and engagement.
* Strategic Planning: The analysis informed our content calendar, guiding when to ramp up advertising efforts. For example, a spike in sports-related articles suggests a prime opportunity for sports-related marketing.



**Keyword Analysis:**

Keyword Analysis in the context of content and marketing involves identifying and examining the frequency and relevance of specific words or phrases within a dataset. This analysis helps to uncover the topics that are capturing public attention and are of interest to readers, which is particularly valuable for SEO and content marketing strategies.

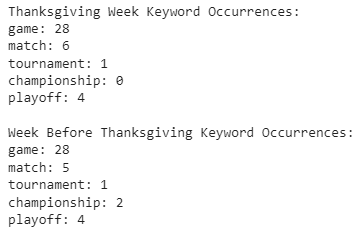
Algorithm & Explanation:

* A comprehensive scan of a body of text is performed to extract prevalent keywords and phrases.
* These terms are then quantified to determine their frequency and subsequently, their significance in the dataset.
* Advanced techniques, such as Natural Language Processing (NLP), can further analyze the context in which these keywords are used to discern their relevance and sentiment.

Application in Project:

Our project leveraged Keyword Analysis in several strategic ways:

* Content Strategy Development: By identifying trending keywords in articles during different weeks, we tailored our content to better align with the interests and searches of our target audience, increasing the likelihood of engagement.
* SEO Optimization: Understanding the most frequent keywords allowed us to optimize our online content for search engines, making our articles more discoverable and improving web traffic.
* Advertising Targeting: Keywords provided insight into audience interests, enabling us to develop targeted advertising campaigns that speak directly to the current conversations and needs of our consumers.



**Sentiment Analysis:**

Sentiment Analysis is a technique used in computational linguistics to identify and categorize opinions expressed in a piece of text, especially to determine whether the writer's attitude towards a particular topic, product, or service is positive, negative, or neutral.

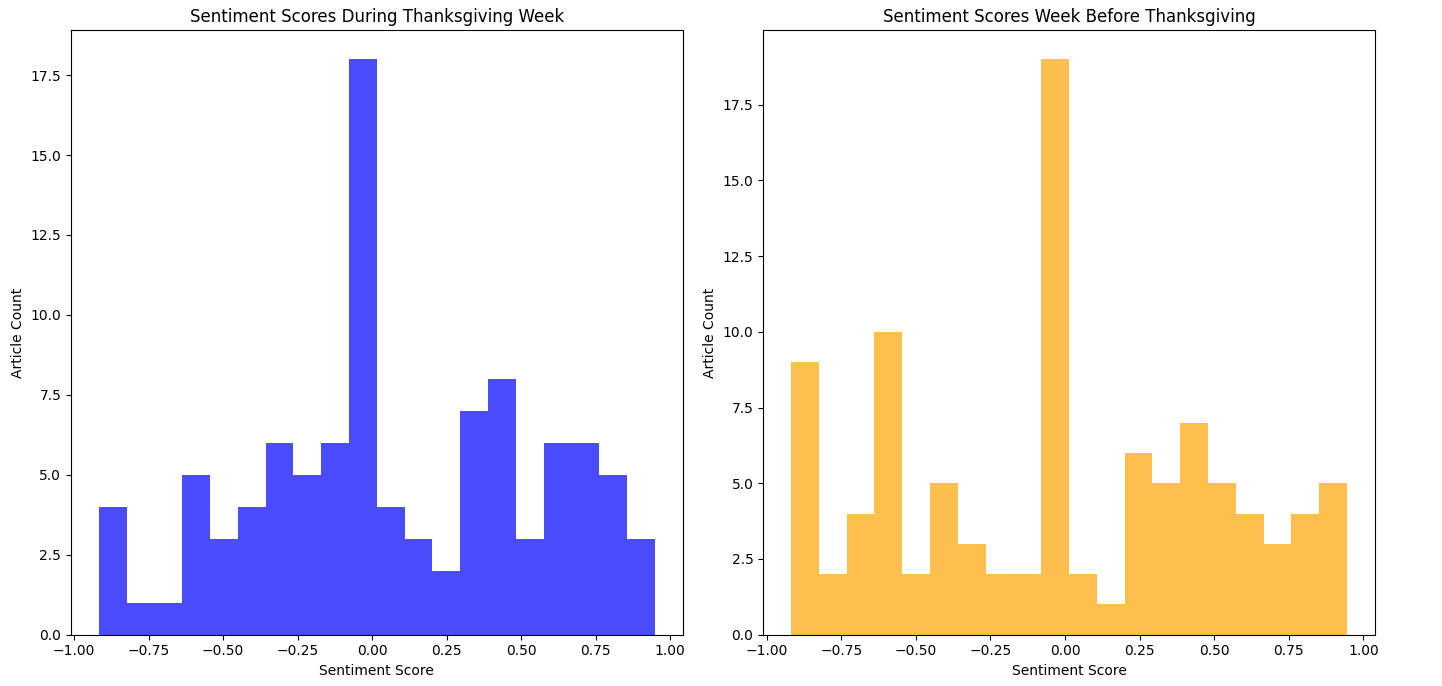
Algorithm & Explanation:

* The method typically involves processing textual data through NLP algorithms that can assess and assign sentiment scores to pieces of text.
* These scores are often calculated using sentiment lexicons (lists of words and their associated sentiment values) or machine learning models that have been trained on annotated datasets.
* The output is usually a sentiment score that reflects the overall sentiment of the text or categorization into sentiment classes.

Application in Project:

In our project, we applied Sentiment Analysis to:

* Content Analysis: By evaluating the sentiment of articles published before and during Thanksgiving, we assessed the overall emotional tone of the content. This helped in understanding the sentiment trends that could affect reader engagement and response.
* Trend Monitoring: Monitoring changes in sentiment over time provided insights into how public opinion shifted, allowing for timely adjustments to content and communication strategies.
* Editorial Guidance: Sentiment scores informed editorial decisions, helping to steer the narrative of published content towards a more positive or realistic tone depending on the desired impact on the audience.



**Trending Topics :**

Trending Topics analysis is pivotal in understanding what subjects or themes are currently popular or gaining momentum within a given dataset. This can include social media, news articles, or any substantial corpus where public discourse is reflected.

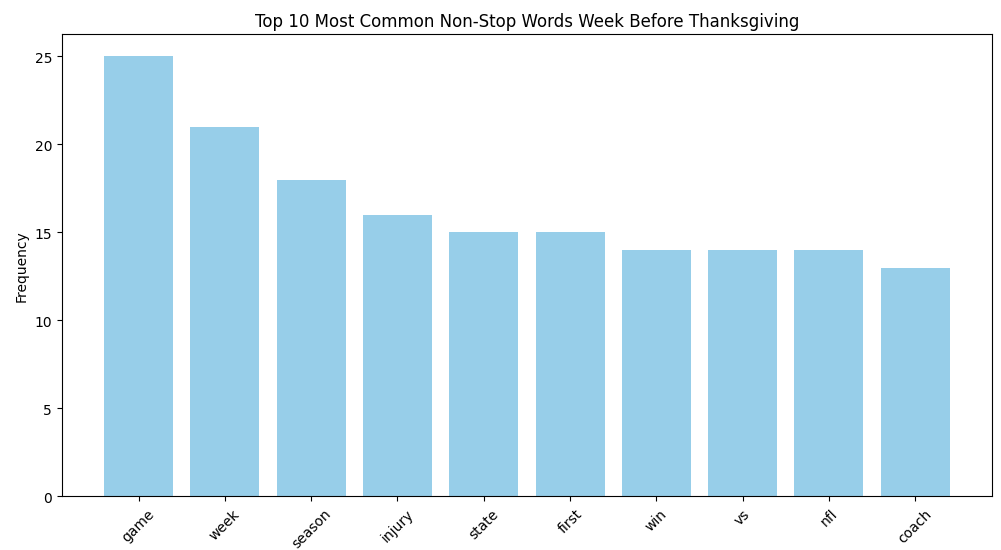
Algorithm & Explanation:

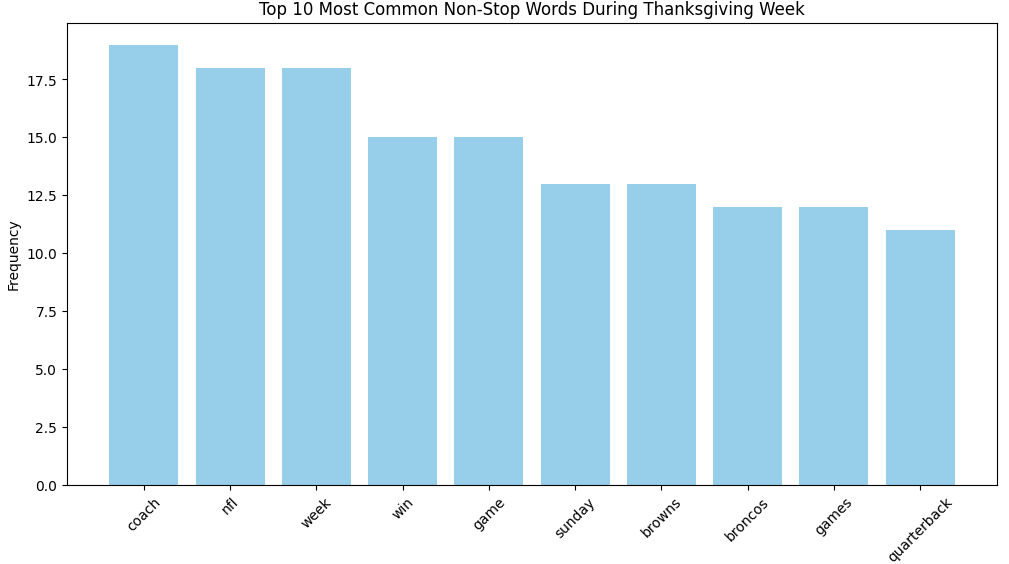
* Textual data is processed to identify frequently occurring words and phrases, often using NLP techniques to ensure that variations of the same word are counted together.
* Common stopwords (e.g., "the", "is", "at") are usually filtered out to focus on more meaningful terms that are indicative of specific topics.
* Advanced algorithms can also track the rise and fall of these terms over time to pinpoint emerging trends.

Application in Project:

For our project, Trending Topics analysis was leveraged to:

* Content Relevance: By pinpointing what topics are trending during different timeframes, we could align our content creation with these trends to remain relevant and interesting to our audience.
* Audience Engagement: Understanding trending topics allowed us to engage with our audience by contributing to the conversations they are most interested in, fostering a stronger connection with them.
* Strategic Planning: Recognizing trending topics provided insights that informed our long-term content and product development strategies, ensuring that we stay ahead of the curve and maintain a competitive edge.

****

****

**Enriching the data with Machine learning**

1. **Named Entity Recognition (NER):**

* Algorithm & Explanation: Named Entity Recognition is an NLP technique used to identify and classify key information (entities) in text into predefined categories such as names of people, organizations, locations, etc. NER typically employs models like Conditional Random Fields (CRF), Bidirectional LSTM (Bi-LSTM), or Transformer-based architectures like BERT.
* Application in Project: In our analysis, NER was used to extract specific entities from the text of articles. This process helped in identifying significant names, places, and brands, which are crucial for understanding the context and focus of the content. By tagging these entities, we gained insights into prevalent themes and subjects in the news coverage.

2. **Sentiment Analysis Enhancement:**

* Algorithm & Explanation: Sentiment Analysis involves determining the emotional tone behind a body of text. This is typically done using algorithms ranging from simple rule-based systems to advanced deep learning models. Recently, Transformer-based models like BERT and RoBERTa have been popular for their effectiveness in understanding context and nuances in language.
* Application in Project: For sentiment analysis, we enhanced traditional methods by employing advanced models that are adept at capturing subtle nuances, sarcasm, and complex expressions. This enhancement allowed for a more accurate and granular understanding of the sentiments expressed in the articles, providing deeper insights into public opinion and reaction.

3. **Extract Embeddings:**

* Algorithm & Explanation: Extracting embeddings involves converting words into vector representations, where semantically similar words have similar vector representations. Techniques like Word2Vec, GloVe, and FastText are common, with more recent approaches using deep learning models like BERT to generate contextual embeddings.
* Application in Project: In our project, embeddings were extracted to understand the semantic relationships between words in our dataset. This process provided a nuanced view of how certain topics or themes are discussed in the context of Thanksgiving. By analyzing these embeddings, we could identify trends, similarities, and differences in content, aiding in content categorization and theme identification.

# **Entity Frequency Analysis:**

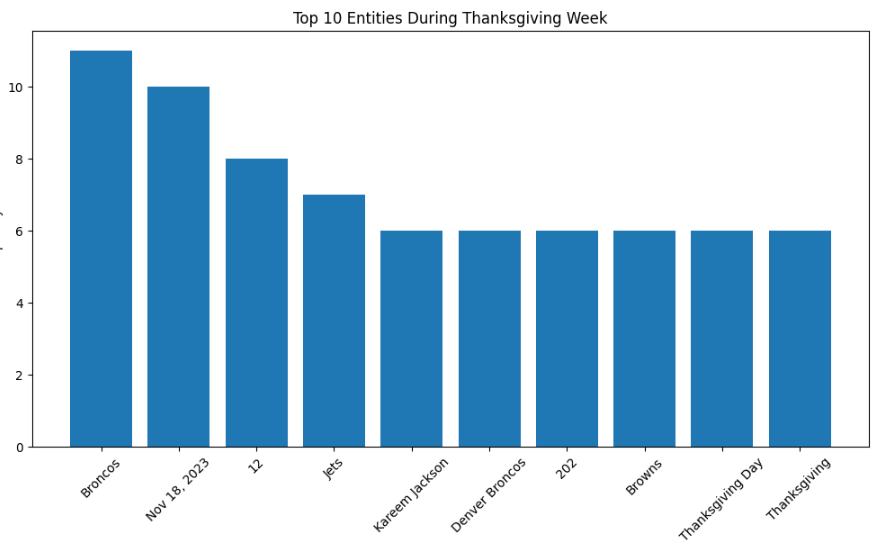
Entity Frequency Analysis is a technique used in Natural Language Processing (NLP) to quantify the occurrence of named entities within a dataset. It helps in identifying the most frequently mentioned entities such as people, organizations, locations, and more. This analysis is particularly useful for understanding key focal points or topics within large volumes of text.

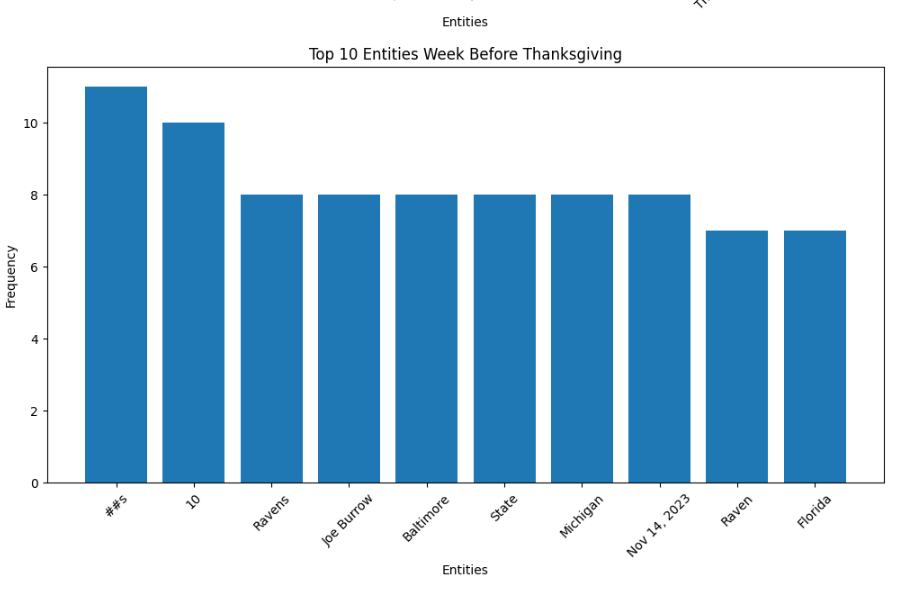
Algorithm & Explanation:

* The process typically begins with Named Entity Recognition (NER), where entities are identified and classified in the text.
* Once entities are extracted, their occurrences are counted and analyzed. This can be as straightforward as a frequency count or involve more complex statistical methods.

Application in Project:

* In our project, Entity Frequency Analysis was utilized to identify the most commonly mentioned entities in articles published during and before Thanksgiving week.
* This analysis provided insights into what topics were most prevalent during these periods. For example, if a particular sports team or player was frequently mentioned, it indicated heightened public interest or relevance.
* The frequency data helped in mapping the landscape of news coverage and understanding thematic shifts – essential for contextualizing our sentiment analysis and clustering results.
* This analysis was instrumental in revealing dominant topics and entities that can be leveraged for targeted content strategies and advertising opportunities.





### **Entity Co-occurrence Analysis:**

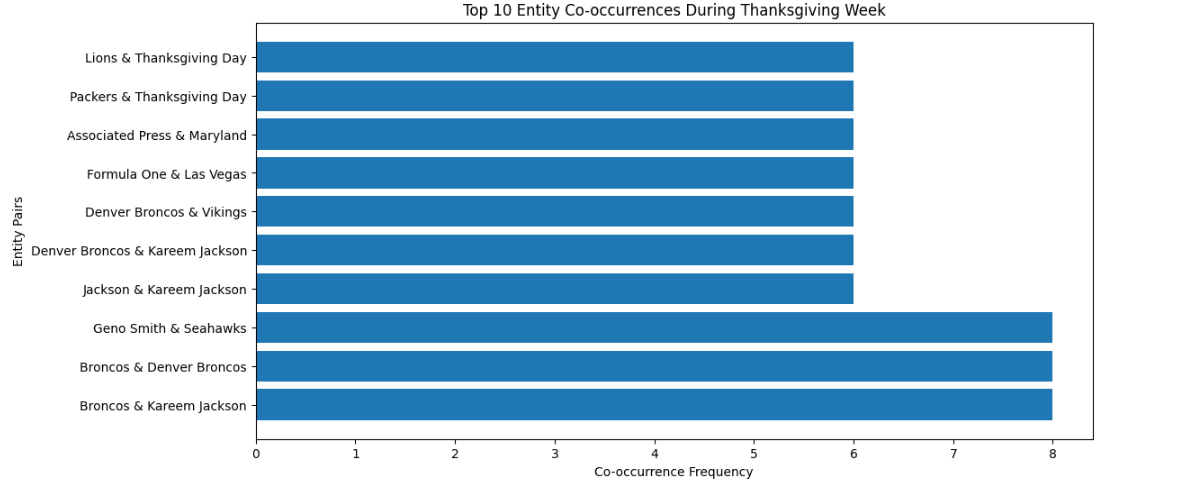
Entity Co-occurrence Analysis is a method used in Natural Language Processing (NLP) to identify entities that frequently appear together within a dataset. This analysis uncovers relationships and associations between different entities such as names, places, organizations, etc., revealing patterns and networks within the text data.

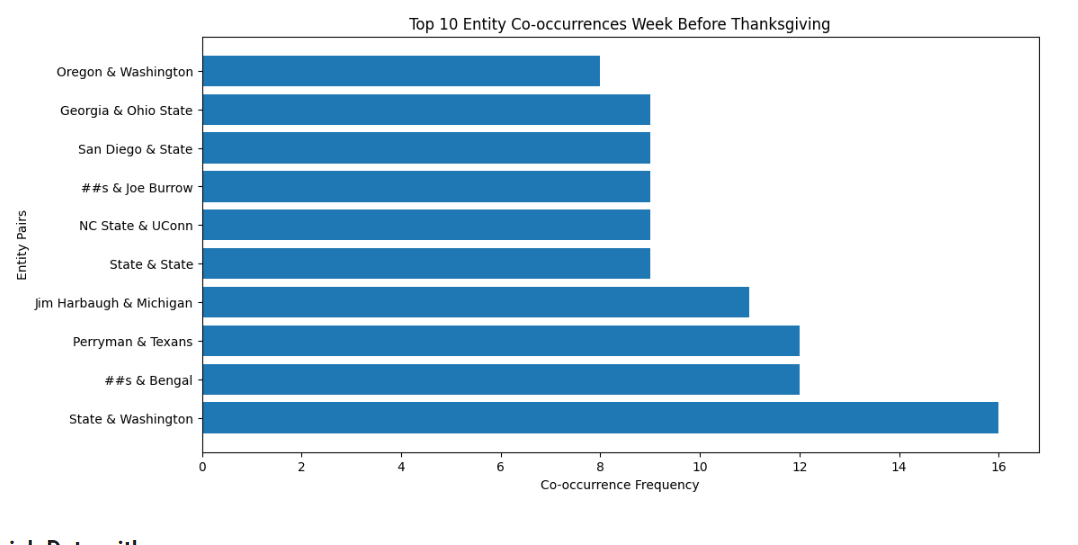
Algorithm & Explanation:

* The process begins with the extraction of entities from the text, often using Named Entity Recognition (NER).
* After identifying these entities, the analysis focuses on measuring how often pairs or groups of entities occur together within certain contexts or within a defined proximity in the text.
* Algorithms used may range from simple frequency counts of entity pairs to more complex probabilistic models that can capture deeper relational patterns.

Application in Project:

* In our project, Entity Co-occurrence Analysis was employed to explore the relationships between different entities mentioned in articles during the Thanksgiving period.
* This approach helped in identifying interconnected themes and subjects. For instance, if certain sports teams and events were often mentioned together, it indicated a shared context or relevance.
* Understanding these co-occurrence patterns was crucial in developing a nuanced view of the thematic structure of the content. It helped in identifying key narrative threads and topics that are interconnected, offering a more holistic understanding of the content landscape.
* The insights gained from this analysis were instrumental in enhancing content strategy, particularly in tailoring content and advertising to reflect the interconnected nature of topics that resonate with the audience.





### **Overall Sentiment Distribution Analysis:**

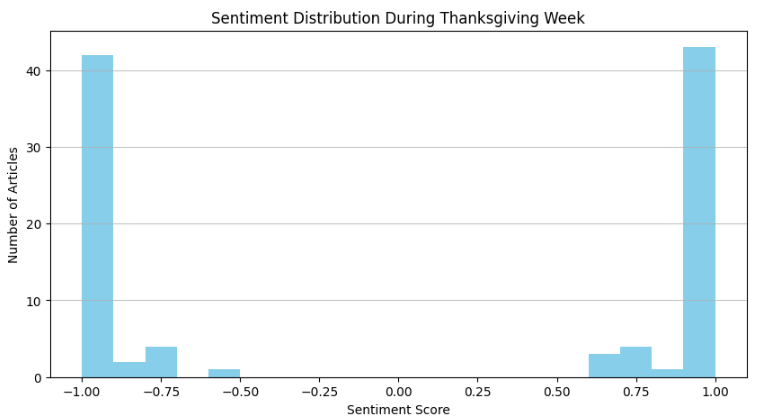
Overall Sentiment Distribution Analysis in NLP is a process used to understand the general sentiment orientation (positive, negative, neutral) across a large body of text. This analysis provides a macro view of the emotional tone in a dataset, offering insights into public mood and opinions.

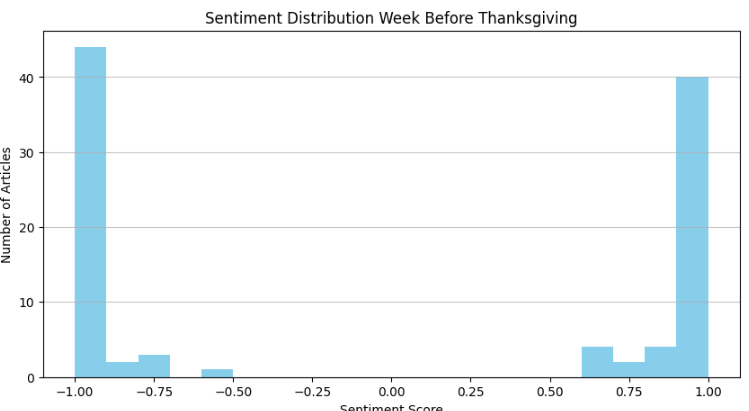
Algorithm & Explanation:

* Sentiment analysis algorithms, ranging from simple lexicon-based methods to complex machine learning models, evaluate the emotional tone of text.
* The overall sentiment distribution is then calculated by aggregating individual sentiment scores across the entire dataset.
* This can involve plotting sentiment scores on a scale (e.g., -1 to 1 for negative to positive) and analyzing the frequency distribution of these scores.

Application in Project:

* In our project, Overall Sentiment Distribution Analysis was applied to articles from the week before Thanksgiving and during Thanksgiving week.
* By analyzing the sentiment distribution, we could identify whether the general tone of the content was more positive, negative, or neutral during these periods.
* This analysis revealed patterns in how sentiments fluctuated, providing insights into how public mood or opinions might have shifted due to the Thanksgiving holiday.
* For instance, an increase in positive sentiments during Thanksgiving week could indicate more upbeat or favorable coverage, which is valuable information for aligning advertising campaigns or content strategies to match the public mood.

****

****

### **Clustering using K-Means:**

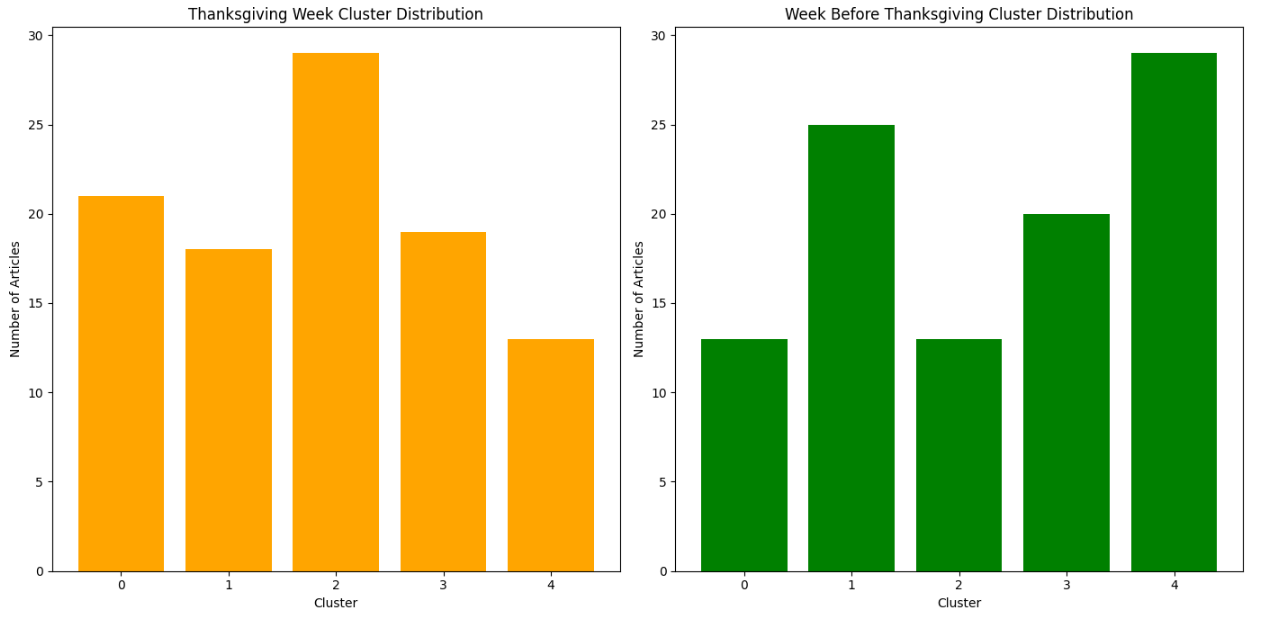
Clustering using K-Means is a powerful unsupervised machine learning technique used to group similar data points together. In the context of text analysis, it helps in categorizing articles, or text content, into clusters based on their similarities.

Algorithm & Explanation:

* K-Means clustering partitions the dataset into K distinct, non-overlapping subgroups (clusters) where each data point belongs to only one group.
* It works by assigning data points to the nearest cluster center (centroid) and iteratively adjusting the position of the centroids for optimal grouping.
* The algorithm is particularly effective in identifying inherent groupings within data when the number of clusters (K) is specified correctly.

Application in Project:

* For our analysis, K-Means was used to cluster articles from the week before Thanksgiving and during Thanksgiving week.
* By applying this method, we were able to identify distinct thematic groupings in the news coverage, with each cluster representing a unique topic or theme.
* Analyzing these clusters helped us understand the predominant themes in news coverage during these periods, providing valuable insights for targeted advertising and content strategies.
* For example, identifying a cluster heavily focused on a specific sports event or holiday sale could guide advertisers to place relevant ads in articles within that cluster.



### **Identifying Themes/True Labels for Clusters:**

Identifying the theme or the true label of each cluster in a dataset is a crucial step in understanding the underlying patterns and groupings discovered through clustering algorithms like K-Means.

Algorithm & Explanation:

* After clustering, each group of data points (or articles, in text analysis) is examined to determine its dominant theme or characteristic.
* This process often involves analyzing the most frequent or central terms, entities, or sentiments within each cluster.
* Advanced NLP techniques like Topic Modeling or NER can be employed to assist in uncovering these key themes or labels.

Application in Project:

* In our project, after applying K-Means clustering to group articles, we sought to identify the primary theme of each cluster.
* Techniques like analyzing the frequency of key terms or entities, or applying topic modeling algorithms, were used to derive a 'label' or 'theme' that best represents the content of each cluster.
* This allowed us to understand what each cluster was predominantly about – whether it was focused on a particular event, personality, sentiment, etc.
* Knowing the true label of each cluster helped in tailoring advertising and content strategies to align with specific themes that are resonant with our audience segments during and around the Thanksgiving period.



**4. Findings:**

* When viewing the most frequently used words from article titles and descriptions we can identify sport specific words, terms relating to teams,players, and games. In this case we noticed that from articles before Thanksgiving that the more popular words are more vague and can be applied to multiple sports meaning that there isn’t really a specific sport Espn is primarily talking about. Compared to words being used during Thanksgiving week we can see they are all primarily focused around football such as (NFL, Sunday, Browns, Broncos, Quarterback). With this in mind we see a spike in NFL related content giving us some insight in how ESPN wants to promote Thanksgiving football games. We can see a big promotion on the Browns and Broncos game being mentioned, which we concluded to be how the Broncos upsetted the Browns.
* The bar charts depicting the "Top 10 Entities During Thanksgiving Week" and the "Top 10 Entities Week Before Thanksgiving" illustrate the predominant topics of discussion in the respective periods. During Thanksgiving week, entities such as "Broncos," "Nov 23, 2023," and "Jets" featured prominently, suggesting a strong focus on sports, with specific attention to football events around that date. Notably, "Thanksgiving Day" and "Thanksgiving" appear as key entities, underscoring the thematic relevance of the holiday itself in the week's discourse. In contrast, the week before Thanksgiving saw entities like "#5," "10," "Ravens," and "Joe Burrow" leading the frequency count, again indicating sports, particularly football, as a central theme, but with a different set of trending topics and personalities. Locations like "Baltimore," "State," "Michigan," and "Florida" also emerged, possibly pointing to significant games or events in these areas. The shift in entity prominence from one week to the next reflects the dynamic nature of news coverage, with specific sports teams and personalities gaining attention in the lead-up to and during Thanksgiving, providing insights into strategic content alignment for those weeks.
* The bar charts depicting "Top 10 Entity Co-occurrences During Thanksgiving Week" and "Top 10 Entity Co-occurrences Week Before Thanksgiving" illuminate the most common entity pairings within each respective period. During Thanksgiving week, pairings such as "Broncos & Kareem Jackson" and "Broncos & Denver Broncos" were most frequent, indicating focused discussions likely around specific NFL games or players. The entities "Lions & Thanksgiving Day" and "Packers & Thanksgiving Day" suggest a thematic connection to NFL Thanksgiving Day games, a staple of the holiday. In contrast, the week before Thanksgiving showed a different trend with "State & Washington" and "#4 & Bengal" among the top co-occurrences, which may highlight college football discussions, reflecting anticipation for end-of-season matches or commentary on rankings and playoffs. "Jim Harbaugh & Michigan" points to coverage on college football coaching, while "Perryman & Texans" might refer to NFL narratives. These charts reveal not only the prominent discussions of the time but also suggest the shift from a broader football dialogue before Thanksgiving to a narrowed focus on specific NFL events during the week, providing rich context for content and advertising alignment with audience interests.
* The sentiment distribution graphs for the Thanksgiving week and the week before Thanksgiving reveal notable patterns in the emotional tone of the articles. In both weeks, there is a bimodal distribution of sentiment scores, with high frequencies at the extremes of the sentiment scale and fewer articles having a neutral sentiment. This suggests that the articles tend to have either a strongly positive or strongly negative sentiment, with little ambivalence. The Thanksgiving week graph shows a slight increase in positive sentiment compared to the week before, indicating a potential uplift in positive sentiment during the holiday. These patterns suggest that while the overall sentiment extremes are consistent, there is a holiday influence that slightly shifts the distribution towards positivity during Thanksgiving. This information could be invaluable for advertisers, indicating that audiences might be more receptive to positive messaging during the holiday period.
* The cluster distribution bar charts for Thanksgiving week and the week before Thanksgiving illustrate the grouping of articles based on thematic content, as determined by K-Means clustering. During Thanksgiving week, the largest cluster is cluster 2, which might represent articles focused on a specific theme such as 'receiver'—a term likely related to football, given the context of Thanksgiving. This is followed closely by cluster 0, potentially related to 'night' events or games. Cluster 1, labeled 'season', and cluster 3, 'miami', indicate additional focal points, while cluster 4, 'coach', has the fewest articles. In contrast, the week before Thanksgiving shows a different distribution. The most significant cluster here is cluster 4, associated with 'loss', which could suggest a focus on games where losses were a central narrative. Cluster 1, representing 'coach', and cluster 0, 'texas', also have a substantial number of articles, indicating these as prevalent topics. Clusters 2, 'baltimore', and 3, 'quarterback', complete the distribution. These distributions not only reflect the dominant sports-related topics during these times but also the shift in thematic focus from the week before Thanksgiving to the holiday week itself. The charts provide a clear visual comparison between the two periods, showing how certain themes become more or less prominent, information that can be leveraged for targeted content creation and advertising strategies.

**5. Conclusion of Analysis:**

**1. Key Takeaways:**

* Our comprehensive analysis of ESPN articles provided valuable insights into the patterns and preferences in sports journalism around Thanksgiving.
* We discovered that certain sports themes, particularly football, dominated the content, reflecting their significance in American culture during the holiday season.
* The strategic use of engaging titles in articles and the timing of their release highlighted the importance of content strategy in capturing audience attention.
* Our findings on entity frequency and sentiment analysis suggest a potential shift in public interest and mood during Thanksgiving, offering a window for targeted advertising and content strategies.
* The application of advanced NLP techniques, including sentiment analysis and entity recognition, proved crucial in deriving these insights, underscoring the importance of data-driven approaches in content analysis.

**2. Concerns (in terms of questions):**

* How might the observed trends in article themes and sentiments vary across different sports seasons or major events?
* To what extent does the focus on specific sports or teams influence broader audience engagement and readership patterns?
* What are the potential limitations of our analysis in capturing the complete picture of audience interests, especially considering the diverse demographics of ESPN’s readership?
* How can advertisers ensure that their strategies, informed by our findings, remain adaptable to sudden shifts in sports trends or unexpected events?
* In what ways might the integration of real-time data analytics further enhance the accuracy and relevance of content and advertising strategies in the future?

**6. Business questions and ideas:**

**Business Question:**

How can a company leverage MongoDB and NewsAPI data from ESPN articles to identify emerging sports trends and capitalize on advertising and marketing opportunities in the evolving landscape of sports content consumption?

**Description of the Business Case:**

The business case involves using NewsAPI data, specifically from ESPN, to proactively identify emerging sports trends and leverage this information for strategic advertising planning. The aim is to stay ahead of the curve, understand shifting consumer interests, and capitalize on advertising opportunities in the dynamic landscape of sports content.

**Making Money:**

We can expect companies and advertisers to make money because based on our data we can see a dramatic increase in the publicity of NFL games during Thanksgiving week compared to the week before. With our findings we can help advertisers with analytics and measurements tools to fund their campaigns. From this they can see how much their viewership data, audience demographics, and engagement metrics change from before and during Thanksgiving. Advertisers can align their messages with the context of the content. For instance, if there's a Thanksgiving football game airing, advertisers can create ads related to football gear, sports jerseys, or other relevant products. This contextual relevance enhances the effectiveness of the advertising by making it more relatable to the viewer in the given moment.

Sentiment analysis: Helps us to understand the mood behind a specific article to cater to a target audience.

Example:

* Brand monitoring can be used based on this sentiment analysis. This can be used as a service for major sports teams in order to monitor the sentiment of their players throughout social media, and allow them to focus on a specific player on their own social media based on their sentiment rating. This model can be applied to merchandise sales as well as the players with the highest sentiment will have the highest jersey sales thus the team should stock that players jersey more frequently.

Entity Co-occurrence Analysis: Reveals associations between different entities (brands, people, locations), allowing for a form of indication that the entities have a specific relationship.

Example:

* This strategy can be applied to sponsorship and endorsement ideas between athletes, teams, and sponsors based on the strength of their co - occurrence analysis. If throughout the articles there are certain brands that have a strong positive co - occurrence with a team or athlete you can align those brands as they will have a better fit and ultimately make more money.

Clustering using K means: This machine learning technique groups similar data points into clusters based on certain patterns or features. For this project, we had two clusters for articles before the week of Thanksgiving, and the week of Thanksgiving. This strategy when implemented in the business realm can allow companies to make more data driven decisions.

Example:

* Using K-means clustering can allow for targeted advertising based on the identified themes in the clusters. These themes can vary from specific interests to regions and allow for ad campaigns to tailor their advertisements based on themes. For instance, one cluster can have a theme of football, so the advertisements directed towards that cluster will be tailored towards football, and this can be in the form of more NFL commercials or even fantasy football ads.

**Summary of Business Strategy:**

The insights gleaned from Entity Frequency and Co-occurrence Analysis offer a myriad of business opportunities by aligning marketing strategies with audience interests. During peak periods like Thanksgiving, when certain topics like sports and holiday events dominate the conversation, there's a golden opportunity for businesses to engage in real-time marketing, launching thematic campaigns that resonate with current events. The frequent mention of specific teams and players signals a chance for partnerships and sponsorships, where businesses can align themselves with prominent figures to gain visibility. Moreover, understanding entity co-occurrence allows for the strategic placement of ads within related articles, ensuring higher relevance and engagement. This analysis also informs content creators on trending topics, guiding them to produce tailored content that captures audience interest, increasing the likelihood of virality. For e-commerce, these trends can inform stock inventory and promotions, anticipating and catering to the surge in demand. Overall, by harnessing the power of entity analysis, businesses can create a more personalized user experience, increase brand affinity, and drive sales during critical periods of consumer activity.

**7. Conclusions:**

Based on our findings…

For many Americans Thanksgiving sports are just as important to the holiday as turkey and cranberry sauce, specifically these are NFL games. The NFL sees a notable jump in representation and positive increase in sentiment in ESPN articles during the holiday which reflects a greater focus on professional football by the average sports’ consumer. This increased presence in the cultural zeitgeist is especially relevant considering the significant increase in consumer spending during the week of Thanksgiving which, of course, corresponds with Black Friday. Advertisers and sports memorabilia based company’s should take advantage of the increased consumer market during the Thanksgiving holiday. The frequent mentions of certain teams and players gives these companies an opportunity to partner with these individuals to further bolster their campaigns. In conclusion, Thanksgiving offers an important chance for companies who wish to target the same demographic as ESPN to notably improve sales and ad campaign reception.

**8. Citations:**

“News API – Search News and Blog Articles on the Web.” News API: Search News and Blog Articles on the Web, [News API – Search News and Blog Articles on the Web](https://newsapi.org/). Accessed 1 Nov. 2023.

OpenAI. (2023). ChatGPT (November 6 version) [Large language model]. <https://chat.openai.com>

MongoDB. (2023). Home.<https://www.mongodb.com/>