EVALUATING GOLF COURSE EXCELLENCE

A Comparative Analysis of Natural Language Processing Techniques in Extracting Insights from Top-Ranked U.S. Golf Course Reviews

Stefan Jenss [1, 2,†]

1 Northwestern University School of Profession Studies

2 https://www.linkedin.com/in/stefanjenss/

† stefanjenss@gmail.com

Abstract

The aim of this study is to compare various Natural Language Processing (NLP) techniques and their performance of term extraction and topic modeling from a unique and novel corpus of golf course reviews. Specifically, this study will compare two feature extraction methods—Term Frequency-Inverse Document Frequency (TF-IDF) and Doc2Vec—and two topic modeling methods—K-Means clustering and Latent Dirichlet Allocation (LDA). The proprietary corpus used for this study consists of 60 golf course reviews for the top 30 ranked golf courses in the United States. This study's main objective is to evaluate the best NLP techniques for analyzing text data in the domain of golf course reviews. Additionally, the secondary objective is to extract valuable insights from the topic modeling results to understand better the distinguishing characteristics and themes associated with these top courses. This research's broader goal is that these insights could guide golf course designers and other industry stakeholders to develop and enhance the design and experience of golf courses of all levels.

Table of Contents

Abstract ii

Introduction and Problem Statement 1

- Problem Statement 1
- Introduction 1
- Study Objectives 2

Literature Review 2

- Background and Context 2
- Natural Language Processing for the Analysis of Sports-Related Content 2
- A Background of the NLP Techniques Used in this Study 3
  - Feature Extraction Methods 3
  - Topic Modeling Methods 4

Data 4

- Overview of the Golf Course Review Text Data 4
- Exploratory Data Analysis of the Golf Course Review Data 5

Research Design and Model Methods....8
Text Data Preprocessing....8
Feature Extraction....9
Feature Extraction Using TF-IDF and Parameter Tuning Using GridSearchCV....9
Feature Extraction Using Doc2Vec (including Parameter Tuning)....10
Evaluation of TF-IDF and Doc2Vec Vector Clusters (Topic Modeling)....11
Visualization of TF-IDF and Doc2Vec Vectors Using Multidimensional Reduction Techniques....11
Evaluation of the Top Features (words) of TF-IDF and Doc2Vec Vector Clusters....11
Topic Modeling Using Latent Dirichlet Allocation (LDA)....11
Results....12
Visualization of Feature Extraction Clusters (TF-IDF-K-Means vs. Doc2Vec-K-Means)....12
Evaluation of Clustering....13
Discussion....14
Conclusions....18
Directions for Future Work....20
Acknowledgments....21
Data Availability....21
Code Availability....21
References....21
Appendix A....23
Appendix A.i: Full-Page Visualization of PCA TF-IDF/K-Means Clustering....23
Appendix A.ii: Full-Page Visualization of t-SNE TF-IDF/K-Means Clustering....24
Appendix B....25
Appendix B.i: Full-Page Visualization of PCA Doc2Vec/K-Means Clustering....25
Appendix B.ii: Full-Page Visualization of t-SNE Doc2Vec/K-Means Clustering....26

Introduction and Problem Statement

Problem Statement

Despite the extensive availability of textual data on the United States' premier golf courses through online reviews, there is a notable gap in any form of systematic analysis of these reviews to extract meaningful insights regarding the features and player experiences of these golf courses. Golf course designers and industry stakeholders could significantly benefit from understanding the attributes that distinguish top-ranked courses, enabling them to replicate these exceptional player experiences more effectively.

This study aims to address this void by employing advanced Natural Language Processing (NLP) techniques to analyze a collection of 60 golf course reviews from the top 30 ranked golf courses in the United States. The research not only seeks to identify key features and themes within these reviews but also aims to lay a foundation for future studies and practical applications of NLP techniques within the golf domain. Ultimately, this research aims to enhance course design and player experience on a global scale.

Introduction

Golf is a sport recognized for its unique mixture of skill, strategy, and natural beauty. Unlike many other sports, the field – the golf field – plays an essential role in shaping the player's experience. With about 15,500 courses in the United States, only a small number of courses are recognized as the best in the country (Lomas 2024). These top courses are characterized by their exceptional design, balanced difficulty (risks versus rewards), meticulous maintenance, and rich history. However, the subjective nature of these criteria makes the task of replicating and adopting the characteristics of these elite courses particularly difficult.

Understanding the distinctive themes and features of top golf courses can provide valuable information to enthusiasts, designers, and industry stakeholders. The aim of this study is to bridge this gap in knowledge by systematically analyzing golf course reviews using advanced natural language processing techniques (NLPs). By extracting and modeling the subject matter from these reviews, we hope to identify the qualities that contribute to the high reputation of these courses. In the end, it is the aim of this study that the insights gained from this research could guide the design and improvement of golf courses, promoting a superior player experience at all levels.

Study Objectives

1. **Perform a Comparative Analysis of NLP Techniques for Analyzing Golf Course Reviews:**
   - **Evaluate and Compare Feature Extraction Methods**: Assess the performance of Term Frequency-Inverse Document Frequency (TF-IDF) and Doc2Vec in extracting relevant features from golf course review texts.
   - **Assess Topic Modeling Techniques**: Analyze the effectiveness of K-Means clustering and Latent Dirichlet Allocation (LDA) in modeling topics within the corpus of golf course reviews.
2. **Extract Meaningful Insights from Topic Modeling Results:**
   - **Identify Common Themes and Features**: Determine key themes and characteristics related to the golf course and player experience present in the golf course reviews.
   - **Provide Actionable Insights**: Offer practical recommendations for golf course designers and industry stakeholders to enhance the design and experience of both existing and future golf courses based on the extracted insights.

Literature Review

Background and Context

The domain of sport-related text data provides a valuable and crucial opportunity for the development of Natural Language Processing (NLP) due to the richness of strategic and dynamic context (Baca et al. 2023). The language surrounding sports—including golf—includes valuable information about the analysis and interpretation of complex game strategy, as well as sentiments about player performance and overall experience. Thus, by enhancing the performance of language processing techniques to understand sport-related text data, we can not only achieve deeper insights but also help meet the demand for language models to provide answers to more complex, scenario-based, and context-specific sport reasoning questions.

Natural Language Processing for the Analysis of Sports-Related Content

The extensive amount of sports content available on the internet, encompassing publications, social media posts, and comments, serves as an ideal data source for Natural Language Processing (NLP). This volume of content is far too vast for any individual to fully

consume and interpret; however, appropriate NLP techniques can efficiently deconstruct and analyze this content for purposes such as topic modeling and sentiment analysis.

Research and application of NLP techniques in sports to elucidate the sentiments of large text datasets have already begun to penetrate various sports sub-domains. For instance, Wahid, Hasan, and Alom (2019) conducted a study employing Long Short-Term Memory (LSTM) networks combined with Recurrent Neural Networks (RNN) to analyze sentiments in cricket-related comments and the feelings of fan bases regarding specific events in the sport. Their work illustrates the potential of NLP techniques to extract meaningful insights from sports data, which can be extended to other sports contexts, including golf.

A Background of the NLP Techniques Used in this Study

In this study, various NLP methods are being deployed, analyzed, and compared regarding their ability to effectively cluster and identify topics and themes within the corpus of golf reviews on the Top 30 golf courses in the United States. These methods include Term Frequency time Inverse Document Frequency (TF-IDF) paired with K-Means clustering, Doc2Vec paired with K-Means clustering, and Latent Dirichlet Allocation (LDA).

Feature Extraction Methods

This study employs two prominent feature extraction methods including Term Frequency-Inverse Document Frequency (TF-IDF) and Doc2Vec.

Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF, as described by Hapke, Howard, and Lane (2019), is a method that calculates the importance of a word by considering its frequency across documents, thus highlighting the significance of the word in the context of the entire corpus. The formula for TF-IDF, for a specific word, i, is as follows:

$$\text{tfidf}_i = \frac{n_i}{N} \times \left(\log \frac{D}{1 + d_i}\right)$$

Where n_i is the frequency of word i . N is the total number of words, D is the total number of documents, and d_i is the number of documents in which word i appears (Jin et al. 2024).

Doc2Vec

Doc2Vec, the second feature extraction method used, is an extension of Word2Vec. Word2Vec is a word embedding model that uses a shallow two-layer neural network to represent the linguistic context of words in dense, low-dimensional vectors. It operates by

analyzing individual words and their surrounding context within a sliding window, systematically processing the text corpus to produce meaningful representations (Jin et al. 2024). Doc2Vec uses this same conceptual methodology but instead converts whole sentences, paragraphs, or documents into numerical vectors that capture the semantic meaning of works within the context of a document (Hapke, Howard, and Lane 2019).

Topic Modeling Methods

For topic modeling, this research compares K-Means clustering and Latent Dirichlet Allocation (LDA).

K-Means Clustering

K-Mean clustering is the first topic modeling technique deployed in this study to analyze the resulting data from the TF-IDF and Doc2Vec feature extraction methods. K-Mean clustering is a well-established method of partitioning bodies of text into, *k*, number of clusters based on the similarity of their content. K-Mean clustering has also been shown to be especially effective in the context of analyzing reviews (Bhukya and Sheshikala 2023).

Latent Dirichlet Allocation (LDA)

Additionally, LDA will be used for this research. LDA is a topic modeling technique that assigns documents to a number of pre-defined topics by attributing each word in the document to one of the document's topics. Similar to K-Means clustering, previous studies have demonstrated LDA's effectiveness in analyzing online reviews. For instance, a group of researchers used LDA to analyze online hospital reviews to help improve patient experiences (Le et al. 2024).

Data

Overview of the Golf Course Review Text Data

The data for this research experiment is a collection of reviews for the top 30 golf courses according to Golf.com's 202 rankings. Golf.com was selected as the source of the ranking due to its well-established reputation as a trusted source for golf information; additionally, each course mentioned in the list also included details about the course, such as the year it was built, architecture, and location. All of the information for the top 30 courses was web-scraped from https://golf.com/travel/courses/top-100-golf-courses-in-the-us-2022-23-ranking. The Python script for this web-scrapping is included in the Code Availability section.

For each of the top 30 courses, two reviews were collected. These reviews were selected by completing a Google.com search query that followed the format "{course name} golf course review." Then, one of the reviews that appeared on the first results page was selected. A variety of authors were targeted for this corpus creation; however, achieving a wide variety of authors proved difficult as there are a limited number of golf reviewers who are both popular enough to appear on the first page of Google.com search results and have access to play such coveted and often time exclusive courses. Table 1 shows a preview of the first 5 entries of the dataset.

Review ID	Course Name	Location	Architect	Year Built	Review Title	Review Author	File Name	Review Text
1	Pine Valley	Pine Valley, NJ	George Crump / Harry S. Colt	1918	PINE VALLEY GOLF CLUB - 19 POINTS	David Jones	rev1_pine_valley_1	There's not much point trying to do a hole-by-...
2	Pine Valley	Pine Valley, NJ	George Crump / Harry S. Colt	1918	Pine Valley Golf Club (Clementon, New Jersey)	Bill Satterfield	rev2_pine_valley_2	What to Expect: Pine Valley is the finest gol...
3	Cypress Point	Pebble Beach, CA	Alister MacKenzie	1928	CYPRESS POINT REVIEW	Graylyn Loomis	rev3_cypress_point_1	"No one but a poet should be allowed to write ...
4	Cypress Point	Pebble Beach, CA	Alister MacKenzie	1928	Cypress Point Golf Club (Pebble Beach, Califor...	Bill Satterfield	rev4_cypress_point_2	What to Expect: I don't even feel worthy to w...
5	Shinnecock Hills	Southampton, NY	William Flynn	1931	Review: Shinnecock Hills Golf Club	Andrew Harvie	rev5_shinnecock_1	There's not many courses as acclaimed, sought ...

Table 1: Preview of the Golf Course Review Corpus for Top 30 Ranked U.S. Courses*. The dataset includes the variables 'review_id,' 'course_name,' 'location,' 'architect,' 'year_built,' 'review_title,' 'review_author,' 'file_name,' 'review_text.'*

Exploratory Data Analysis of the Golf Course Review Data

A brief exploratory data analysis (EDA) was conducted on the numerical and categorical aspects of the dataset. Figure 1 illustrates the distribution of years that the golf courses were originally built. Notable findings from this visualization include:

- The distribution of years built is positively skewed, with the majority of courses being built between 1895 and 1937.
- There are no current top 30 courses built between 1937 and 1995.
- Three of the top 30 courses were built between 1995 and 2003.

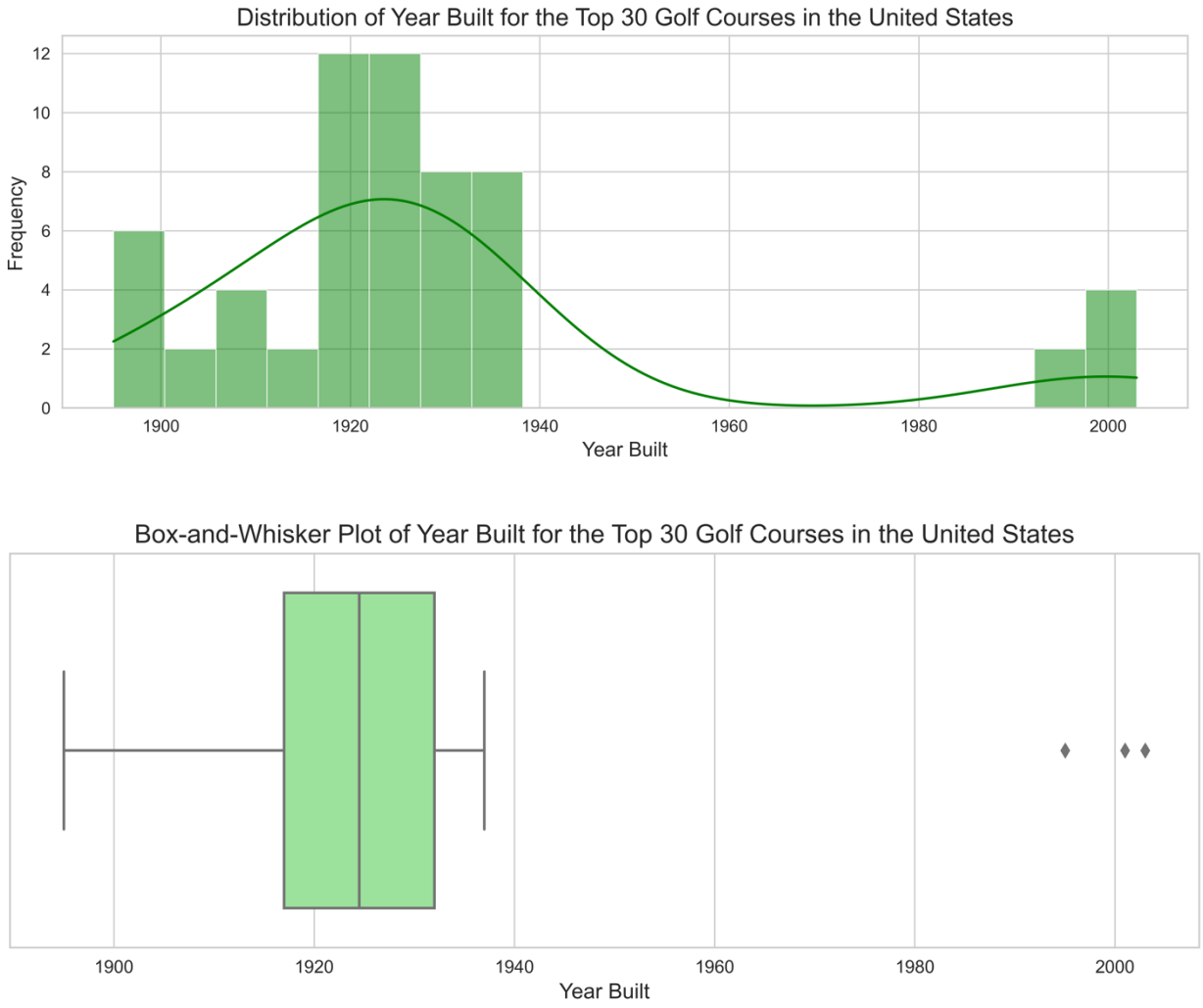


Figure 1. (a) Distribution of the year of construction for the top 30 golf courses in the United States. (b) Boxplot of the construction year for the top 30 courses in the United States. *(Note: frequency count should be halved to account for each course having two reviews)*

Figure 2 illustrates the frequency distribution of the golf course architects who built the courses. The distribution shows that the majority of architects appearing in the top 30 rankings only designed 1 course; however, notably, Seth Raynor designed two courses, Ben Coore and Ben Crenshaw designed two as a team, Ronald Ross designed three, and A.W. Tillinghast designed four of the top 30 golf courses.

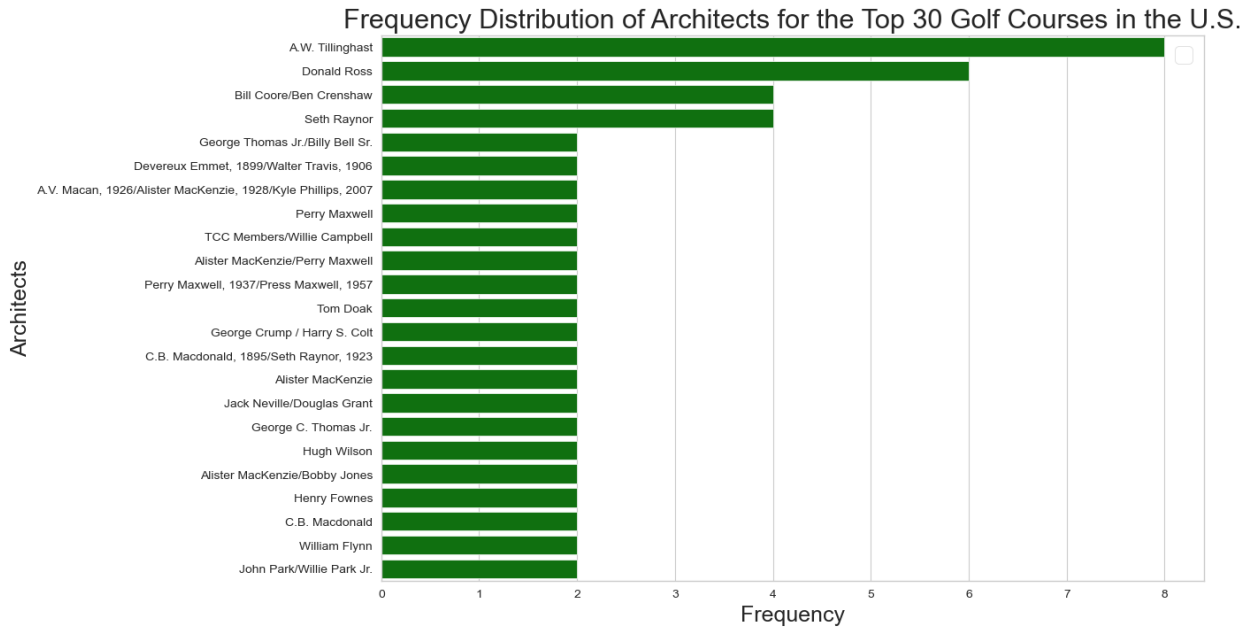


Figure 2. Frequency distributions of the golf course architects who designed the courses included in the top 30 ranking. *(Note: frequency count should be halved to account for each course having two reviews)*

Next, a frequency distribution was performed for the location of the golf courses by state to examine whether there were certain states that hosted a large number of the top courses. **Figure 3** shows this frequency distribution, and interestingly, New York has the highest frequency of top courses, with 8 out of the 30 courses located in the states. The state with the next highest frequency is California, with 6.

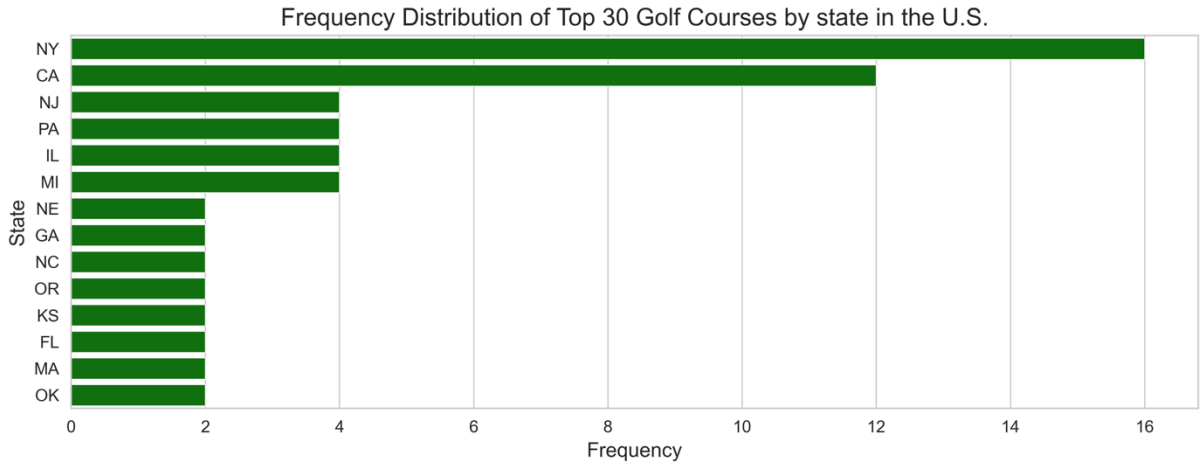


Figure 3. Frequency distribution of Top 30 Golf Courses by State in the U.S. *(Note: frequency count should be halved to account for each course having two reviews)*

Lastly, **Figure 4** shows the frequency distribution of review authors. Notably, David Jones, Bill Satterfield, and the website top100golfcourses wrote a high frequency of reviews, and this demonstrates a potential weakness in the dataset that should be considered for future experimentation.

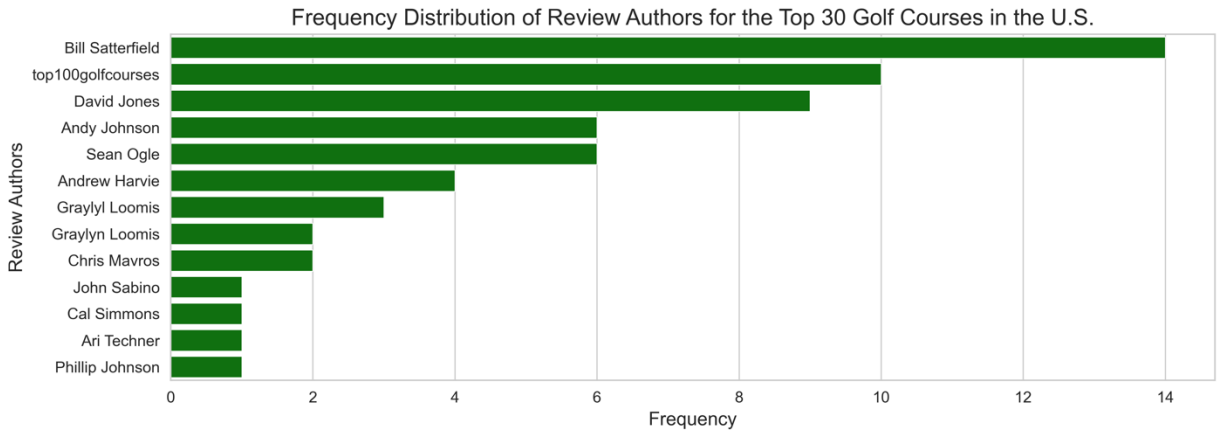


Figure 4. Frequency Distribution of Review Authors for the Top 30 Courses in the U.S. *(Note: frequency count should be halved to account for each course having two reviews)*

Research Design and Model Methods

The research design used for this study involved a thorough preprocessing of the text data followed by two feature extraction methods and two topic modeling techniques.

Text Data Preprocessing

1. **Preprocessing Libraries and Models:**
   - The preprocessing phase of this research involved using several NLP libraries, including spaCy for lemmatization and NLTK for its standard stopwords and regular expression tokenizer. The steps taken in this phase included:
2. **Stopwords:**
   - ***Standard Stopwords:*** These were loaded from NLTK's English stopwords list.
   - ***Domain-Specific Stopwords:*** In addition to the standard NLTK stopwords, a number of golf-domain-specific stopwords were defined, including "course," "play," "hole," "green," "par," "tee," "yard," "golf," "bunker," "fairway," "leave," "shot," "right," "good," and "club."

- ***Hole Numbers:*** Golf hole numbers, such as "one," "two," "three," …. "eighteen," appear frequently in these reviews and provide little semantic meaning, so these were also removed.
- ***Course Names:*** To ensure that topic clusters were formed based on the themes and characteristics of the different courses, the specific course names were also included as stopwords (e.g., "pine," "valley," "augusta," etc.).

3. **Preprocessing Function:**
   - The preprocessing function written performed tokenization, lowercasing, removal of non-alphabetic and short tokens, and removal of both standard and corpus-specific stopword somatization on the text data.

Feature Extraction

Feature Extraction Using TF-IDF and Parameter Tuning Using GridSearchCV

The first feature extraction method used was Term Frequency-Inverse Document Frequency (TF-IDF), which involved evaluating the importance of a word in a document relative to a collection of documents. In the context of this research, the aim was for this method to identify words that are unique and significant to specific reviews.

In this phase, a pipeline integrating TF-IDF and K-Means clustering was created to allow for extensive parameter testing and fine-tuning that would otherwise be too burdensome.

1. **GridSearchCV Implementation and Parameter Grid:**
   - GridSearchCV with 5-fold cross-validation was used to find the optimal combination of parameters based on the parameter grid included below. This allowed us to evaluate the model's performance across various TF-IDF and K-Means parameters and identify both the optimal set of parameters for TF-IDF and the number of clusters/topics to use for the research.

Parameters Used	Parameter purpose	Parameter Values
`tfidf__max_df`	Maximum document frequency threshold to ignore terms appearing too frequently	0.75, 0.85, 0.95
`tfidf__min_df`	Minimum document frequency threshold to ignore terms appearing too infrequently	0.01, 0.05, 0.10
`tfidf__ngram_range`	Range of n-grams to consider	unigram, bigram, trigram
`kmeans __n-clusters`	Number of clusters to form with K-Means	3 – 10

2. **Parameter Evaluation and TF-IDF Model Fitting**:
    - The GridSearchCV was used to fit the preprocessed review text and identify the best TF-IDF parameters and clustering scores.
    - The best parameters identified were: 'kmeans__n_clusters': **3**, 'tfidf__max_df': **0.95**, 'tfidf__min_df': **0.1**, 'tfidf__ngram_range': **(1, 1)**.
    - The best estimator tool from GridSearchCV was then utilized to transform the preprocessed review text into TF-IDF vectors, which were subsequently converted into a DataFrame format for further analysis.

Feature Extraction Using Doc2Vec (including Parameter Tuning)

The second method of feature extraction used was Doc2Vec, an extension of the Word2Vec model. This method represents sentences, paragraphs, or entire documents using dense vectors that capture semantic information about the document.

For this research, we applied the Doc2Vec model to the preprocessed review data and, as with TF-IDF, subsequently used K-Means clustering to identify groups within the reviews. We systematically trained several Doc2Vec models and evaluated cluster metrics to optimize their parameters.

1. **Parameter Tuning for Doc2Vec Model**
    - To identify the optimal parameters, a function was created to train a series of Doc2Vec models using 3 clusters (as determined during TF-IDF tuning) and tuning of the following parameters:
        - **`vector_size`**: Dimensionality of the feature vectors.
        - **`window`**: Maximum distance between the current and predicted word within a sentence.
        - **`min_count`**: Minimum frequency threshold to ignore infrequent words.
        - **`epochs`**: Number of iterations over the corpus.
    - The optimal parameters identified were: 'vector_size': **200**, 'window': **10**, 'min_count': **2**, 'epochs': **10**
        - A table of the different Doc2Vec parameters used and their accompanying Davies-Bouldin and Calinski-Harabasz scores is included in ***Appendix B***.

Evaluation of TF-IDF and Doc2Vec Vector Clusters (Topic Modeling)

Visualization of TF-IDF and Doc2Vec Vectors Using Multidimensional Reduction Techniques

To visualize the clustering of the TF-IDF and Doc2Vec vectors identified using K-Means clustering, we employed two different dimensionality reduction techniques, Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE). Multiple multidimensional reduction techniques were used to visualize the clusters in multiple ways and gain better insight into the cluster patterns.

Evaluation of the Top Features (words) of TF-IDF and Doc2Vec Vector Clusters

1. **Identifying Top Features in TF-IDF Clusters**
    - To identify the most important words in each of the TF-IDF vector clusters, we analyzed the cluster centers (centroids) generated by the K-Means algorithm. Using these centroids, we determined the top features by examining their TF-IDF scores within the centroid, with higher TF-IDF values indicating greater significance in the cluster.
2. **Identifying Top Features in Doc2Vec Clusters**
    - To analyze the top features for each cluster in the Doc2Vec model, we calculated the cosine similarity between the cluster centroid and the word vectors within the Doc2Vec model. This method helped us pinpoint the words that closely resembled the centroid of each cluster, indicating the primary features of each cluster.

Topic Modeling Using Latent Dirichlet Allocation (LDA)

In our research, we used LDA to analyze the preprocessed golf course reviews and identify underlying themes in the data. LDA is a widely used topic modeling technique for analyzing corpora of text, making it particularly suitable for this research.

1. **Preparation for LDA Topic Modeling**
    - To prepare the preprocessed text data for LDA, we created the necessary Gensim-compatible data structures, including a dictionary and corpus.
2. **Applying LDA Topic Modeling and Extracting Topics**
    - The LDA model was trained on the prepared dictionary and corpus, with the number of topics being set to 3 for the model training.

- After the LDA model was trained, we extracted the top words for each topic to analyze their coherence and compare them to the top features extracted using K-Means clustering of the TF-IDF and Doc2Vec vectors.

Results

Visualization of Feature Extraction Clusters (TF-IDF-K-Means vs. Doc2Vec-K-Means)

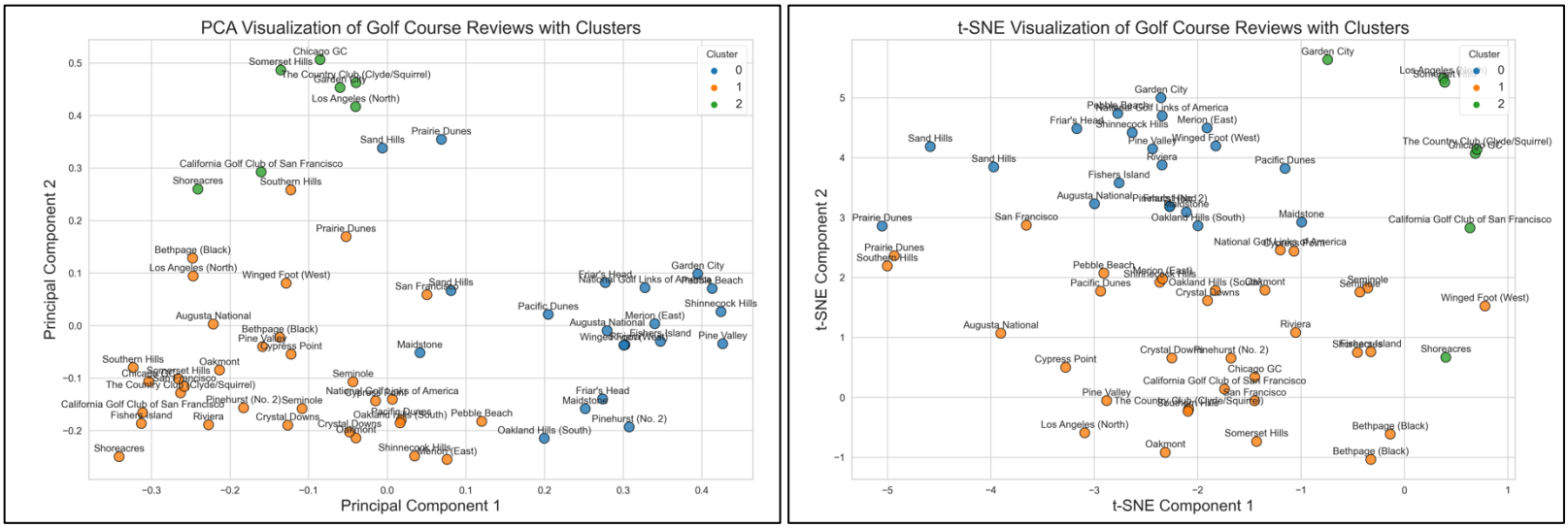


Figure 1. Visualization of TF-IDF vector K-Means clusters in 2-dimensional space through multidimensional reduction various techniques. (a) Principal Component Analysis (PCA) multidimensionality reduction. **(b)** t-distributed Stochastic Neighbor Embedding (t-SNE) multidimensionality reduction. *[note: fill page images are included in **Appendix A**]*

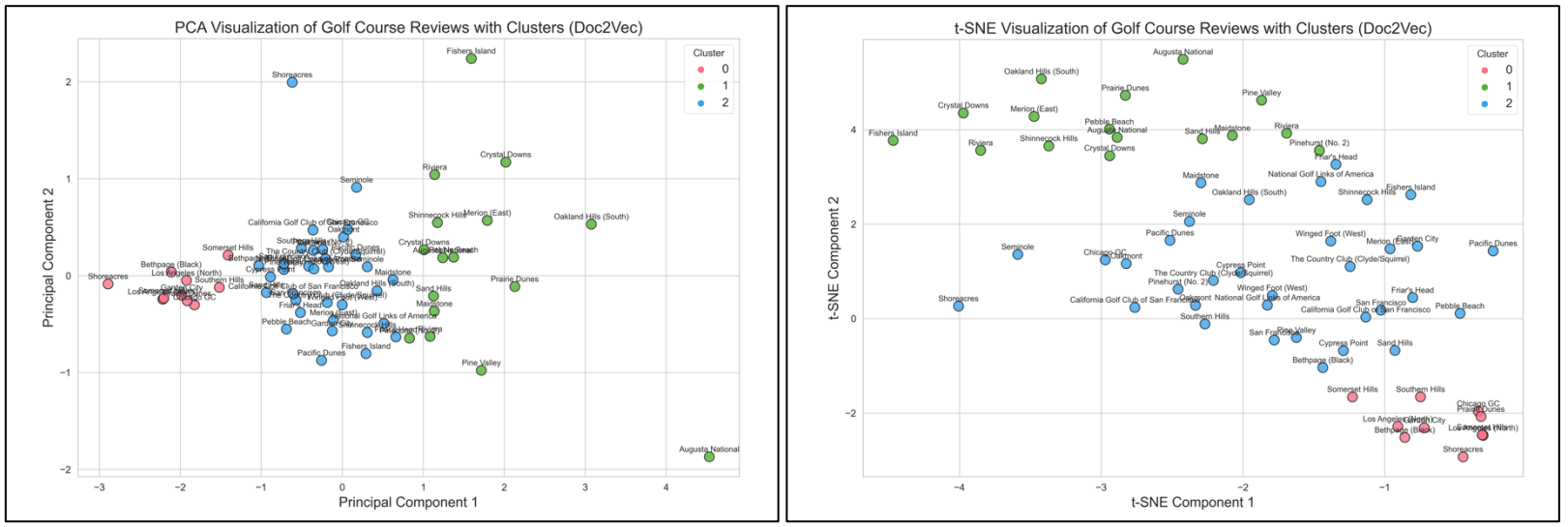


Figure 2. Visualization of Doc2Vec vector K-Means clusters in 2-dimensional space through multidimensional reduction various techniques. (a) Principal Component Analysis (PCA) multidimensionality

reduction. **(b)** t-distributed Stochastic Neighbor Embedding (t-SNE) multidimensionality reduction. *[note: fill page images are included in the **Appendix B**]*

Evaluation of Clustering Metrics (TF-IDF-K-Means vs. Doc2Vec-K-Means)

NLP Method	Davies-Bouldin Score	Calinski-Harabasz Score
TF-IDF | K-Means	3.4322	3.0339
Doc2Vec | K-Means	0.7717	53.2060

Table 2. Clustering evaluation metrics, Davies-Bouldin Score and Calinski-Harabasz Score, for the two NLP methods used in the study: TF-IDF with K-Means and Doc2Vec with K-Means. A lower Davies-Bouldin score indicates better clustering, and a higher Calinski-Harabasz score indicates better cluster.

Evaluation of TF-IDF-K-Means & Doc2Vec-K-Means Top Features and LDA Topic Modeling

TF-IDF Feature Extraction Paired with K-Mean Clustering	
Cluster 0	get, well, would, time, really, back, first, much, day, make, pretty, like, look, long, see
Cluster 1	back, side, short, front, approach, player, difficult, long, surface, make, well, ball, slope, feature, put
Cluster 2	layout, usa, design, restoration, original, perhaps, raynor, cup, today, north, former, new, american, locate, first

Table 3. The top 15 features (words) for each of the three clusters that were identified using TF-IDF feature extraction paired with K-Means clustering on the preprocessed text data.

Doc2Vec Feature Extraction Paired with K-Mean Clustering	
Cluster 0	design, open, championship, layout, american, original, build, maxwell, architecture, famous, man, macdonald, mackenzie, early, fact
Cluster 1	create, bring, property, low, turn, sit, fall, must, area, view, set, high, yet, use, land
Cluster 2	follow, shape, call, become, fine, want, layout, also, route, take, let, land, although, try, photo

Table 4. The top 15 features (words) for each of the three clusters that were identified using Doc2Vec feature extraction paired with K-Means clustering on the preprocessed text data.

Latent Dirichlet Allocation (LDA) Topic Modeling	
Topic 0	get, well, back, time, like, would, make, first, short, see, look, much, long, day, hit
Topic 1	short, back, side, put, approach, front, slope, feature, ball, make, long, round, surface, tree, find
Topic 2	back, side, front, approach, short, long, well, make, player, slope, ball, difficult, great, get, come

Table 5. The top features (words) for each of the three topics generated by the Latent Dirichlet Allocation (LDA) model on the preprocessed text data. A visualization PCA visualization of the LDA clusters is provided in Appendix C.

Discussion

Comparative Analysis of Feature Extraction Methods (TF-IDF vs. Doc2Vec) and the Application of K-Means Clustering to this Vectorized Data

Analysis of Performance Metrics

To evaluate the effectiveness of the features extraction methods, TF-IDF and Doc2Vec, paired with K-Means Clustering, we used the clustering evaluation metrics Davies-Bouldin (DB) and Calinski-Harabasz (CH) scores. These two performance metrics were chosen for this portion of the research to provide a comprehensive validation of the clustering results. The evaluation metrics are presented in **Table 2**.

- **Davies-Boudlin Score:** For evaluating Davie-Bouldin scores, a lower value indicates better, more compact clusters that are well separated (Wang and Xu 2019). The results show that the DB score for Doc2Vec (0.7717) is significantly lower than that for TF-IDF (3.4233), indicating that the clusters formed with Doc2Vec vectors are more cohesive and distinct.
- **Calinski-Harabasz Score:** Conversely, a higher Calinski-Harabasz score indicates better-defined clusters with greater between-cluster dispersion relative to within-cluster dispersion (Singh et al. 2020). The CH score for Doc2Vec (53.2060) is substantially higher than that for TF-IDF (3.0339), further supporting Doc2Vec's superior performance in producing well-defined clusters.

These results demonstrate that, from at least a quantitative perspective, Doc2Vec, when paired with K-Means clustering, outperforms TF-IDF in creating compact, well-separated clusters when using the golf course review text data. Given that Doc2Vec has the advantage of better capturing the semantic relationships within reviews, this is a likely cause of the more meaningful clustering outcomes.

Observational Analysis of Cluster Visualization via Multidimensional Reduction

We employed two different multidimensional reduction techniques, Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) to better understand the K-Means clustering results visually. Both of these multidimensional reduction methods employ different approaches to aid in the visualization of high-dimensional data, and for this reason, both were chosen to be shown in this paper.

Principal Component Analysis (PCA) Visualization

The first multidimensional reduction technique used to visualize the clusters was PCA, which transforms the data into a new coordinate system by reducing the dimensionality of the data by linear means. The first principal component, or first coordinate, accounts for the greatest variance by any projection of the data, followed by the second principal component, which accounts for the second greatest variance (Franc 2022). The PCA plots for both TF-IDF and Doc2Vec feature extraction vectors (Figure 1 & 3) reveal the following:

- **TF-IDF PCA Visualization:** The cluster appears to be relatively distinct within the 2-dimensional space; however, there is a noticeable overlap between the clusters, especially in two larger clusters (clusters 0 and 1).
- **Doc2Vec PCA Visualization:** The Doc2Vec clusters appear more distinct, better separated, and more compact compared to the TF-IDF visualization. This visualization supports the quantitative performance metrics findings that suggest Doc2Vec better captures semantic similarities between course reviews and, thus, produces better clustering results.

t-distributed Stochastic Neighbor Embedding (t-SNE) Visualization

The next multidimensional reduction technique used to visualize the clustering was t-SNE, a nonlinear dimensionality reduction method that also works very well for displaying high-dimensional datasets. Different from PCA, t-SNE works by minimizing the divergence between the distributions that measure pairwise similarities of the input object in high-dimensional space and the corresponding low-dimensional points (Nanga et al. 2021). The t-SBE plots for both TF-IDF and Doc2Vec feature vectors (Figures 2 and 4) provide the following insights:

- **TF-IDF t-SNE Visualization:** The clusters exhibit more overlap and are less distinct compared to the PCA visualization. Interestingly, this multidimensional reduction technique causes certain reviews of the same course to appear very close to each other, which wasn't the case in the PCA visualization. Some of these courses include Bethpage (Black), Seminole, Prairie Dunes, and Sand Hills. Since the course names were removed from the reviews during data preprocessing, it is likely that these courses have other very notable and unique words in their reviews, leading to their close proximity in the plot.

- **Doc2Vec t-SNE Visualization:** While the clusters are less compact compared to the PCA visualization, they appear to be well-separated with clear boundaries between them. The advantage of t-SNE here is that you can more clearly see which course reviews appear in which clusters, while still maintaining distinct clusters.

Overall Interpretation of Multidimensional Reduction Visualization Findings

Similar to examining a physical object in 3-dimensional space from multiple perspectives, combining insights from both PCA and t-SNE allows for a more nuanced understanding of the clustering results. The PCA visualization more clearly illustrated the compactness of clusters in comparison to each other. However, in the context of this study, t-SNE has the advantage of allowing for a clearer inspection of the reviewed courses in each cluster, while maintaining a similar level of distinctness and separateness of the clusters.

Comparative Analysis of Topic Modeling Results (K-Means Clustering with Feature Extraction Methods vs. LDA)

Overview of Topic Modeling Techniques

As previously discussed, two topic modeling approaches were employed in this study: LDA and K-Means clustering (paired with TF-IDF and Doc2Vec feature extraction techniques). The use of these methods aims to identify the key themes and topics within the corpus of golf course reviews and hopefully uncover insights into the attributes and characteristics that contributed to the golf courses' high reputations. In this section, we will provide an interpretation and comparative analysis of the topic modeling outcomes derived from the various feature extraction and clustering techniques.

K-Means Clustering with TF-IDF and Doc2Vec Feature Extraction

TF-IDF with K-Means Clustering (Table 3)

- **Cluster 0 (Overall Experience and Visual Sentiments)**: This cluster captures the general sentiments about the course's visual appeal and the golfers' (a.k.a. the reviewers) overall experience, with words such as “get,” “well,” “time,” “day,” “make,” “pretty,” “look,” “see,” “first,” “really,” and “much.”
- **Cluster 1 (Course Design and Technical Gameplay)**: This cluster is characterized by words that likely represent the technical aspects of the golf course design and gameplay on the course. Words in this cluster related to course design include “difficult,” “short,” “long,” “slope,” “feature,” and “surface.” Words related to technical gameplay include “back,” “front,” “side,” “approach,” and “player.” The terms related to course design indicate the discussion of the challenges and features of the golf course, while the terms related to technical gameplay likely refer to the strategic placement of golf shots as players navigate the intricacies of the course.

- **Cluster 2 (Course Design and Historical Aspects)**: This cluster primarily highlights words related to the courses' designs, such as "layout," "design," and "locate." However, the cluster also significantly focuses on the courses' historical aspects, including terms such as "usa," "restoration," "original," "cup," "Raynor," "former," and "first."

Doc2Vec with K-Means Clustering (Table 4)

- **Cluster 0 (Historical Aspects and Famous Architects)**: This cluster includes words such as "design," "open," "championship," "layout," and "American," indicating that these courses have a history of hosting major championship golf tournaments with championship layouts. Other words, including "Maxwell," "architecture," "famous," "MacDonald," and "Mackenzie," also indicate that the review for the courses in this cluster likely contains significant discussion about the famous course architects who designed them.
- **Cluster 1 (Course Layout and Landscapes)**: This cluster predominantly features words related to the physical characteristics of the golf course, including "create," "bring," "property," "low," "high," "turn," "sit," "fall," and "area." Additionally, this cluster includes words such as "view," "set," "use," and "land," which likely refers to how the course designers designed the course around the physical features and visual elements of the landscape.
- **Cluster 2 (Technical Gameplay and Course Strategy)**: Featuring words such as "follow," "shape," "want," "layout," "route," "take," "let," and "try," this cluster likely refers to the strategical aspects of the course and how players must navigate the routing and layout of the course strategically.

Latent Dirichlet Allocation (LDA) (Table 5)

- **Topic 0 (General Experiences on the Golf Course)**: This topic includes words such as "get," "well," "back," "make," "short," "hit," "look," and "see," which likely refer to general sentiments and overall experience of the golfers from both a playing and visual perspective.
- **Topic 1 (Technical Aspects of Gameplay on the Course)**: This cluster contains words such as "short," "back," "make," "long," "front," and "approach," which likely represent the technical aspects of playing the golf course. Additionally, terms such as "feature," "round," "surface," and "slope" indicate the discussion of the course's physical features that must be navigated.
- **Topic 2 (Difficulties and Shot Execution related to Gameplay)**: This topic includes terms like "back," "side," "front," "approach," "short," "long," and "slope." Additionally, it contains words such as "player," "make," "difficult," "great," and "get," which suggests that the topic focuses on course difficulty and shot execution associated with these courses.

Comparative Analysis

- **TF-IDF with K-Means:** The clusters formed using TF-IDF are relatively coherent. However, there is some notable overlap between clusters, particularly clusters 1 and 2, which both contain a mixture of course-designed-related terms.

- **Doc2Vec with K-Means:** Doc2Vec produced the most distinct and semantically rich clusters. One cluster focused on the historical aspects of the course and its architecture, another focused on the course layout and landscapes, and the third focused on the technical gameplay and course strategy required. It is likely Doc2Vec's ability to capture semantic relationships that contributed to its production of more meaningful topic clusters.
- **LDA:** There is a significant overlap in the terms included in each cluster resulting from the LDA technique. For instance, words such as "back," "make," "short," and "long" appear in all three clusters. Additionally, Topics 1 and 2 both contain the terms "side," "approach," "front," "slope," and "ball." This significant overlap is likely a result of LDA's probabilistic approach, which, while good at capturing broad themes, may struggle with the finer nuances of identifying distinct themes in a corpus of only golf reviews.

Conclusions and Actionable Insights

Conclusions

Conclusions Regarding Feature Extraction Methods

The comparative analysis of both the quantitative clustering performance metrics and the multidimensional reduction cluster visualizations demonstrate that in the domain of golf review text data, Doc2Vec paired with K-means clustering is the most effective when compared to TF-IDF. Doc2Vec significantly outperformed TF-IDF regarding Davies-Bouldin and Calinski-Harabasz scores, indicating more compact and well-defined clusters. Furthermore, both the PCA and t-SNE visualization of the Doc2Vec clusters displayed less inter-cluster overlap and greater cluster concentrations than TF-IDF.

Conclusions Regarding Topic Modeling Techniques

The comparative analysis of the topic modeling results for the three methodologies used demonstrates that Doc2Vec paired with K-Means clustering was the best at capturing and differentiating the nuanced and semantic content of golf course reviews when compared to TF-IDF and LDA. While TF-IDF also created very coherent clusters, Doc2Vec outperformed TF-IDF in terms of differentiation and richer semantic representations, likely due to Doc2Vec's use of dense vectors (versus TF-IDF's sparse vectors).

Surprisingly, and despite being a more sophisticated topic modeling methodology, LDA performed significantly worse than both Doc2Vec and TF-IDF paired with K-Means clustering. In this research, LDA was found to produce topic clusters with significant overlap, indicating that LDA is less effective at performing thematic distinctions in the realm of golf, possibly due to the

nuance of golf course reviews. Generally, golf course reviews share many similarities, and two courses that are drastically different might also be thematically differentiated by a select few—but important—set of distinguishing features. Given the significant overlap in many of these golf reviews and LDA's usefulness at specifically identifying broad topic themes, this might be a contributing factor to why LDA's topic modeling performance is lacking in comparison to the other topic modeling techniques used.

In conclusion, the results of this study found that, out of the NLP methodologies tested, Doc2Vec paired with K-Means clustering was the most effective in analyzing golf course reviews and produced the most valuable topic modeling results. Also, while less effective, TF-IDF paired with K-Means clustering also provides valuable topic modeling results and is worth implementing in addition to Doc2Vec to possibly identify additional topics. The insights extracted from these topic modeling results have the potential to provide valuable guidance in the enhancement of golf course design and player experience.

Actionable Insights from Topic Modeling Results

The results from the topic modeling methods have the potential to provide actionable insights with practical application for the design and enhancement of courses. Some of the insights uncovered in this study include:

Designing Courses with Strategic Gameplay: At least one of the cluster topics identified in all three of the topic modeling techniques was interpreted to be pertaining to the technical gameplay of the golfer on the course. Given this, golf course designers and managers should focus on designing their golf courses in a way that incorporates aspects of the course that force golfers to strategically think about how they should best navigate the golf course.

At some point on any given hole, the hole should ideally present to the golfer a hypothetical question of how they should approach the hole. Subsequently, the golfer will need to answer this question by deciding whether to hit the ball short and lay up, assess the risk of going long, evaluate which side is best to miss on, and settle on an ideal shot shape they should make (if they are that skilled). The top golf courses in the world require that a golfer not only play well but also think smartly, and this is something that all golf course designers and managers should strive to bring to their courses.

Golf Course Location, Routing, and Layout: Some other prominent themes identified from the topic modeling results included the importance of course layout and landscape. While not always related to how a golf course plays, the scenery associated with a golf course's

location was shown to be a notable feature in the reviews of the top courses. However, while this is a notable insight, changing the location of an already-existing golf course is obviously infeasible.

Although the location of the golf course is a very influential factor in how highly the course is regarded, an equally as important and more actionable aspect of a golf course is its layout and how the progression of the holes move throughout the natural landscape (a.k.a., "the routing"). Great golf courses should strive to use the land of a golf course in a way that utilizes the natural landscape and optimizes the flow of holes through the property.

Embrace, Remember, and Encourage History: Another aspect of great golf courses that was identified but is hard to quickly adopt or change is the history associated with the course. While golf courses cannot change their history, they can do things to proactivity foster a richer history and stronger reputation. One such proactive action might include bringing in notable golf course architects to help in the redesign of courses. Moreover, the course can push to be the host of either professional or amateur events, which will hopefully bolster the course's reputation and nurture a greater sense of course history as time goes on.

Directions for Future Work

Future research can build on this study's findings in several ways, including expanding the dataset of course reviews, applying additional NLP techniques using the same experimental methodology, or incorporating sentiment analysis or classification.

Expanding the Dataset of Course Reviews: One of this study's shortcomings is the relatively limited dataset that is used to explore the various NLP techniques used. The current dataset only included two reviews for each of the top 30 courses on the Golf.com top 100 list. Future research could expand upon this dataset by either (1) increasing the number of reviews per course up to three or 4 reviews per course, (2) expanding the number of courses beyond the top 30 to the top 50 or 100, or (3) a combination of the two.

Implementation of Additional Feature Extraction and Topic Modeling Techniques: Another future direction for this work can include using the same experimental methodology but expanding the number of NLP feature extraction and topic modeling techniques used for comparative analysis. Additional NLP models can be explored, such as GPT-based models or BERT.

Incorporation of Sentiment Analysis and Classification Techniques: Another direction to take the methodology and design of the research is to integrate sentiment analysis techniques to evaluate the emotional tone of the reviewer, which might help provide a more comprehensive understanding of the reviewer's experience. However, given that these are some of the nicest, pristine, and well-designed courses, it is likely that the sentiment for all of these course reviews might be similar. Alternatively, the dataset could be expanded with reviews of courses that are not included in the top 100 list, and future research could attempt to create a classification model with the ability to identify whether a review is likely to be of a top 100 golf course or a course outside the top 100.

Acknowledgments

Grammarly was used to check for grammatical errors, help correct spelling errors, and to fix run-un sentences.

ChatGPT 4o was used to help brainstorm the organizational flow of the paper's subsections.

Data Availability

The data for this research is available via the following link to the research GitHub Repository: https://github.com/stefanjenss/Natural_Language_Processing/blob/main/MSDS_453_Term_Project/golf_course_review_corpus.csv

Code Availability

Web Scrapping File in GitHub Repository

https://github.com/stefanjenss/Natural_Language_Processing/blob/main/MSDS_453_Term_Project/golf_course_reviews_scrapper.ipynb

NLP Research Code

https://github.com/stefanjenss/Natural_Language_Processing/blob/main/MSDS_453_Term_Project/V3.5_top_courses_NLP_code_(course_name_removed).ipynb

References

Baca, L., N. Ardiles, J. Cruz, W. Mamani, and J. Capcha. 2023. "Deep Learning Model Based on a Transformers Network for Sentiment Analysis Using NLP in Sports Worldwide." In

Communications in Computer and Information Science, 328–39. https://doi.org/10.1007/978-3-031-37940-6_27.

Bhukya, B., and M. Sheshikala. 2023. "NLP Based Topic Modeling for Healthcare: Analyzing Patient Reviews to Improve Quality of Care and Access to Services." *2023 International Conference on Emerging Techniques in Computational Intelligence (ICETCI)*, September. https://doi.org/10.1109/icetci58599.2023.10330957.

Franc, A. 2022. "Linear Dimensionality Reduction." *arXiv (Cornell University)*, January. https://doi.org/10.48550/arxiv.2209.13597.

Hapke, H., C. Howard, and H. Lane. 2019. *Natural Language Processing in Action*. 2nd ed. Simon and Schuster.

-Jin, H., Y. Zhang, D. Meng, J. Wang, and J. Tan. 2024a. "A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods." *arXiv.Org*. March 5, 2024. https://arxiv.org/abs/2403.02901.

Le, T., T. Ho, V.-H. Nguyen, and H.-S. Le. 2024. "How to Deeply Understand the Voice of the Customer? A Proposal for a Synthesis of Techniques for Analyzing Online Reviews in the Hospitality Industry." *Journal of Hospitality and Tourism Insights*, February. https://doi.org/10.1108/jhti-07-2023-0460.

Lomas, N. 2024. "How Many Golf Courses are in the US?" Golf Span - Golf Tips and Equipment Reviews. January 5, 2024. https://www.golfspan.com/how-many-golf-courses-are-in-the-us.

Nanga, A., A. T. Bawah, B. Acquaye, M. Billa, F. Baeta, N. A. Odai, S. K. Obeng, and A. D. Nsiah. 2021. "Review of Dimension Reduction Methods." *Journal of Data Analysis and Information Processing* 09 (03): 189–231. https://doi.org/10.4236/jdaip.2021.93013.

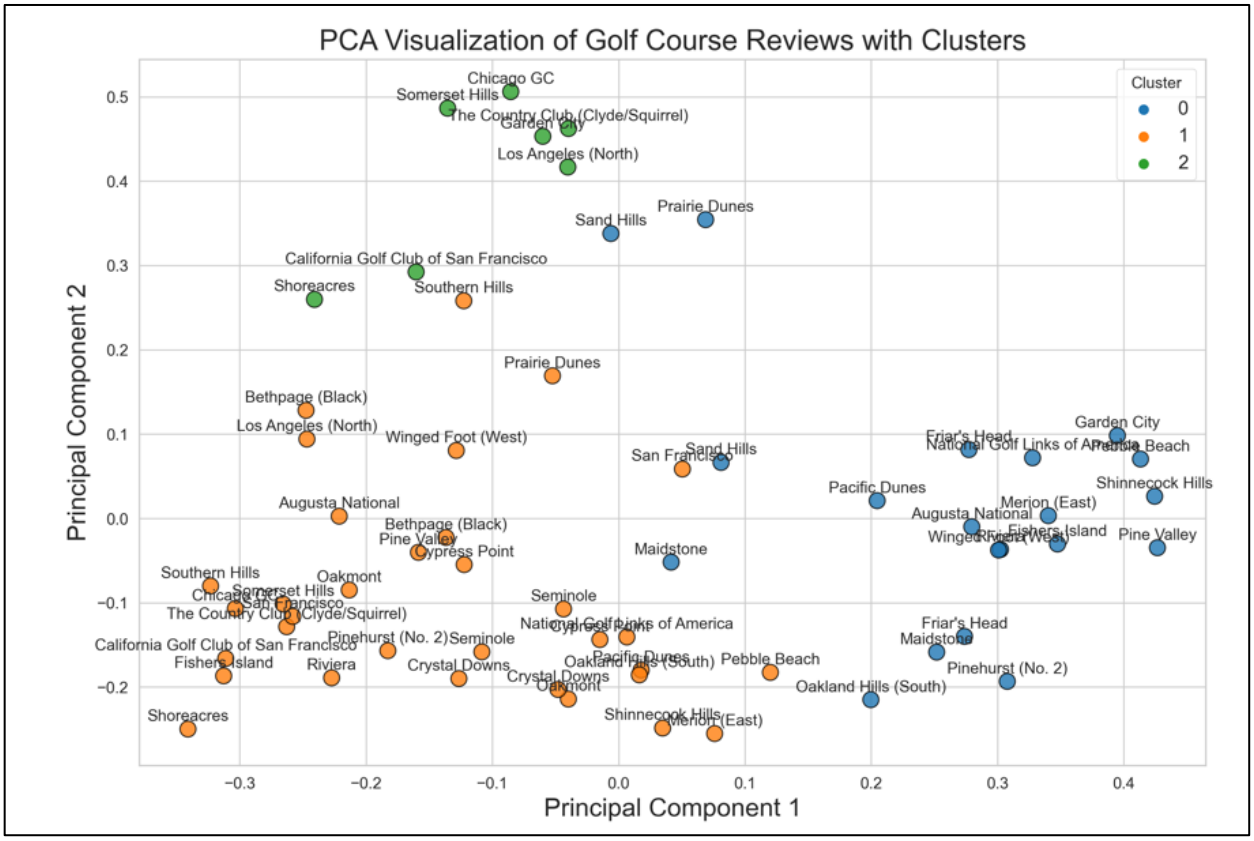
Singh, A. K., S. Mittal, P. Malhotra, and Y. V. Srivastava. 2020. "Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal Data Using K-Means," March. https://doi.org/10.1109/iccmc48092.2020.iccmc-00057.

Wahid, F., J. Hasan, and S. Alom. 2019. “Cricket Sentiment Analysis from Bangla Text Using Recurrent Neural Network with Long Short Term Memory Model,” September. https://doi.org/10.1109/icbslp47725.2019.201500.

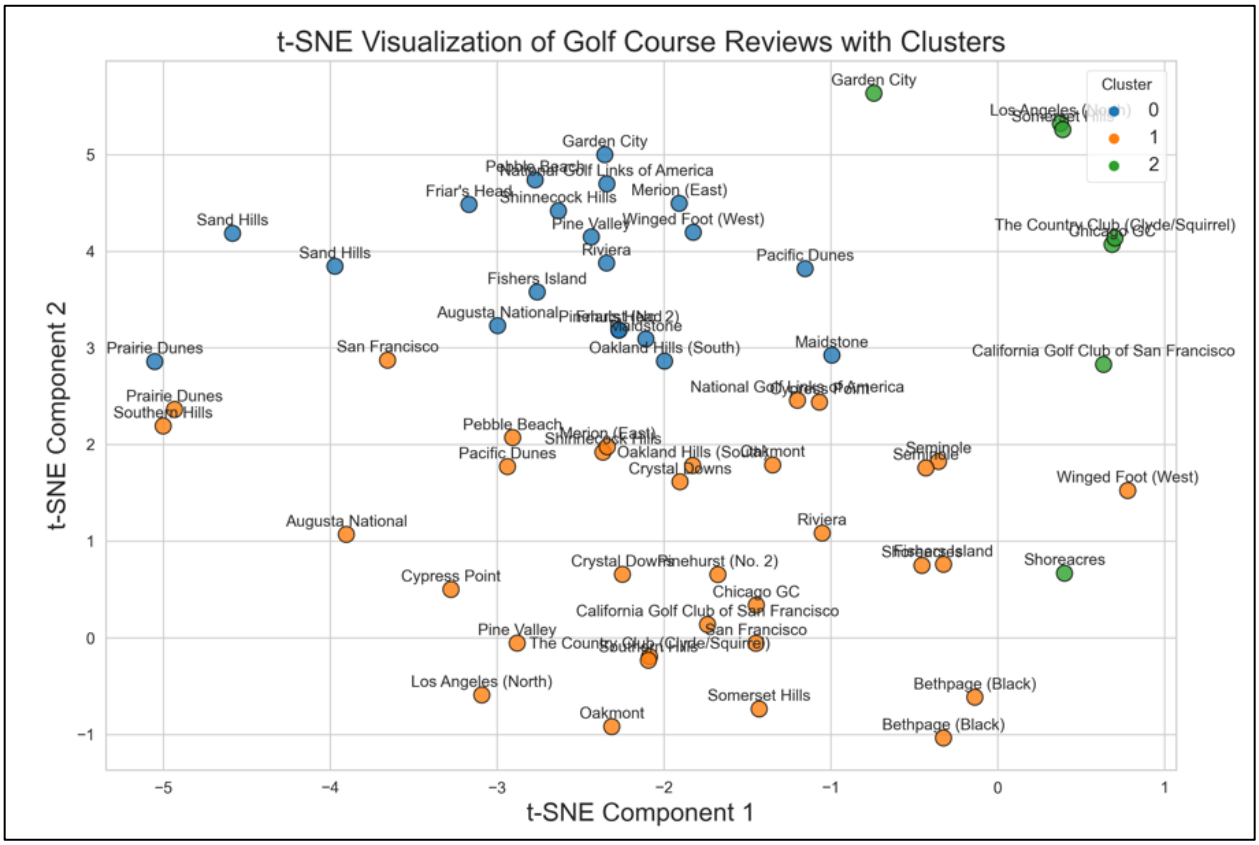
Wang, X., and Y. Xu. 2019. “An Improved Index for Clustering Validation Based on Silhouette Index and Calinski-Harabasz Index.” *IOP Conference Series. Materials Science and Engineering* 569 (5): 052024. https://doi.org/10.1088/1757-899x/569/5/052024.

Appendix A – TF-IDF/K-Means Clustering Visualizations

Appendix A.i: Full-Page Visualization of PCA TF-IDF/K-Means Clustering

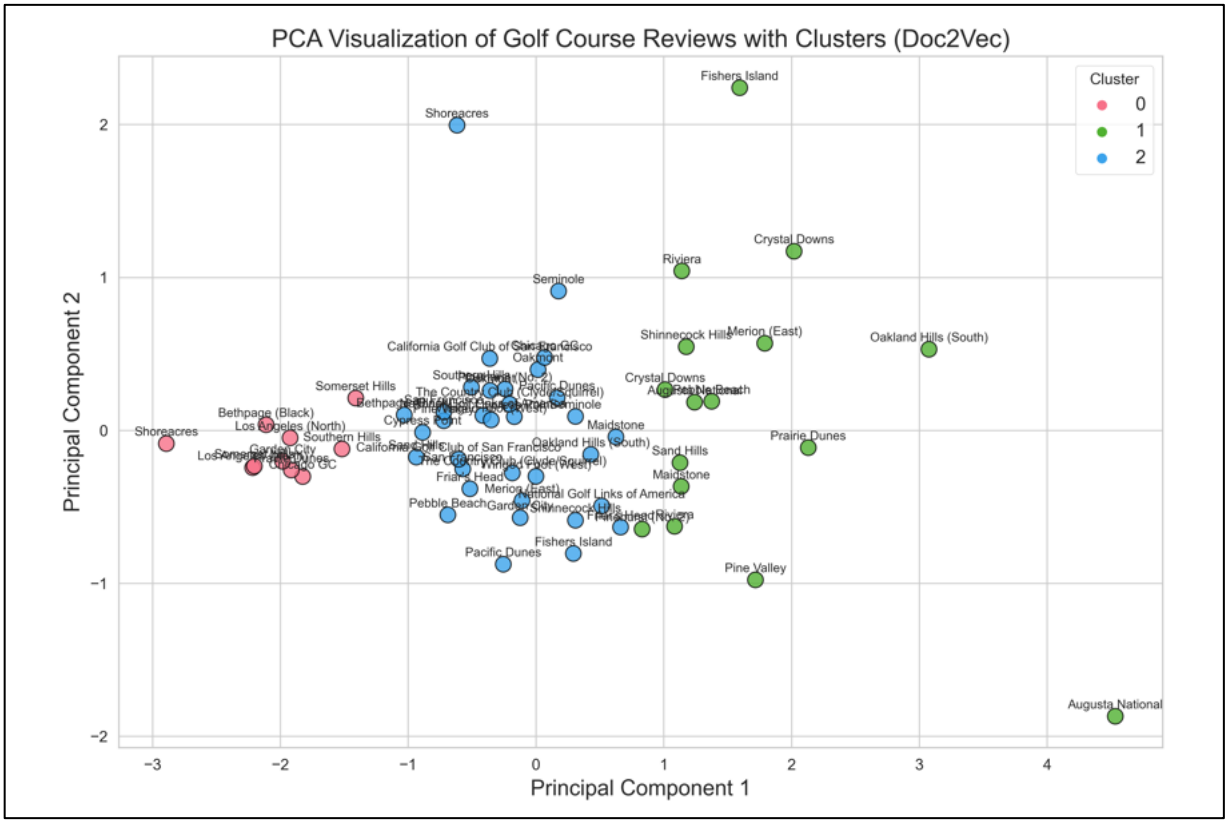


Appendix A.ii: Full-Page Visualization of t-SNE TF-IDF/K-Means Clustering

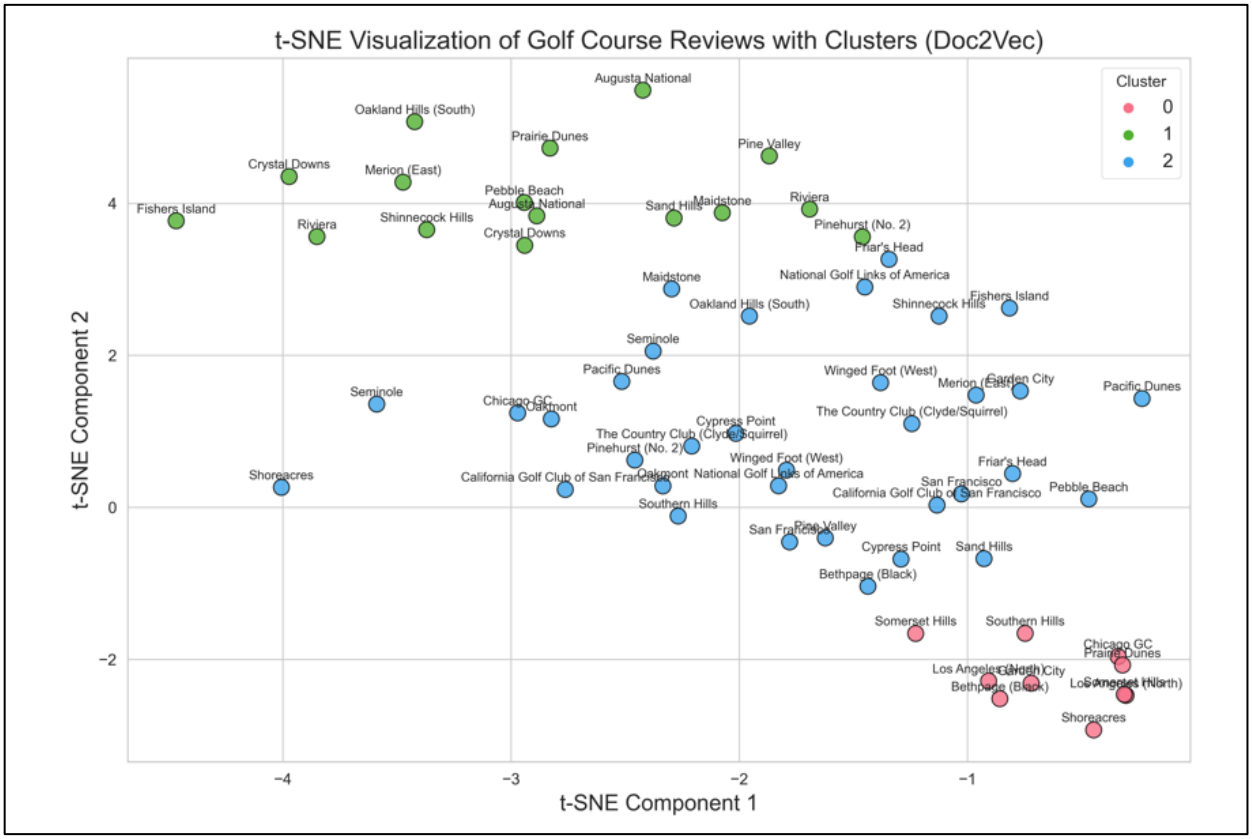


Appendix B – Doc2Vec/K-Means Clustering Visualizations

Appendix B.i: Full-Page Visualization of PCA Doc2Vec/K-Means Clustering



Appendix B.ii: Full-Page Visualization of t-SNE Doc2Vec/K-Means Clustering



Appendix C – LDA Visualization