[redacted]
08 May 22

Country and Soul: Analyzing Trends in American Music Journalism from 1960 to Present

1. Introduction

Since its founding, the United States has distinguished itself from other nations by forging a unique cultural identity. Central to this identity is the country's musical traditions, and no genre represents America better than country music. Country music has its roots in African American spirituals, Appalachian folk, and Mexican ballads. Despite this mixed heritage, country music's highest profile stars are predominantly white and from southern states. Their songs are, at times, homogeneous and thus insufficient to characterize the entirety of America's musical identity.

Soul music, developed from blues and African American gospel, is another distinctly American genre which gained popularity contemporaneously with country music. Unlike country, the most well-known soul artists are black. While one could argue that the output of soul musicians is also homogeneous, pairing soul with country provides a more complete portrait of American music history.

This study explores the significance of country and soul music in American culture by applying various text analytical techniques to a corpus of magazine and newspaper articles. Although the corpus only contains work by professional writers, it serves as a proxy for the broader public's opinion on genre and artist. The corpus also contains several interviews which capture the views of musicians in their own words. By analyzing this diverse collection of documents, I hope to discover a linguistic style in music journalism which reveals latent patterns in the American musical tradition.

2. Corpus

Articles for this analysis were scraped from rocksbackpages.com, a digital archive of top music publications. I collected all articles that were tagged with the genres *country, country rock, and bluegrass,* or *soul, funk, and R&B.* Then, I removed documents that were inconsistent with the broader corpus. This included book passages and obituaries which were likely to contain too much biographical information and insufficient musical descriptions. Lastly, I filtered out any articles that were not tagged to an artist. The remaining corpus contains 3,651 documents published between 1960 and 2022. The corpus primarily contains interviews, reviews, profiles, and sleevenotes.

Each article was tokenized using NLTK's sentence and whitespace tokenizers. Then, tokens were tagged with part of speech (POS) labels using NLTK's POS tagger. It was assumed that the structure of each article could be organized in an ordered hierarchy of content objects (OHCO). The OHCO structure contains levels for article, paragraph, sentence, and token. Each token in the corpus was indexed based on the OHCO. In total, the corpus contained 3,876,953 tokens.

After tokenization, I created a vocabulary of unique terms (62,946 terms). Then, I converted the corpus into two bag-of-words representations (BOW), using whole articles as bags. The first BOW, hereinafter the *term BOW*, contains all terms in the vocabulary. The second BOW, hereinafter the *n-gram BOW*, contains the top 20,000 most frequent unigrams, bigrams, and trigrams. Each BOW records the number of occurrences and the term frequency-inverse document frequency (TFIDF) for each term or n-gram. I used Scikit-learn's TFIDF transformer to compute values. TFIDF is preferred to term occurrence since it discounts low-information words that appear in a high number of documents.

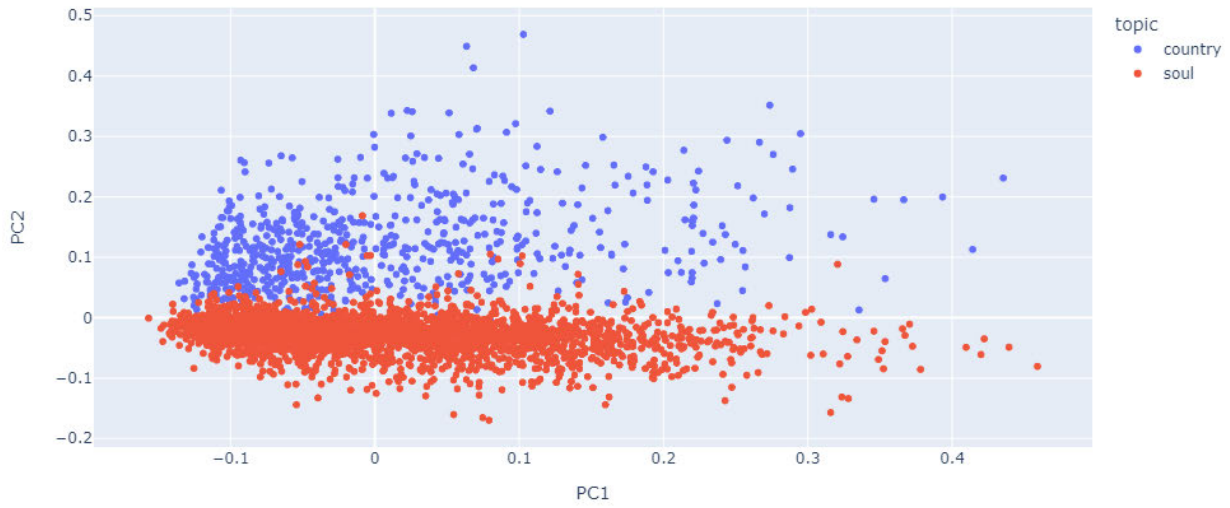

Figure 1. Scatter plot of articles along the first two principal components, colored by genre.

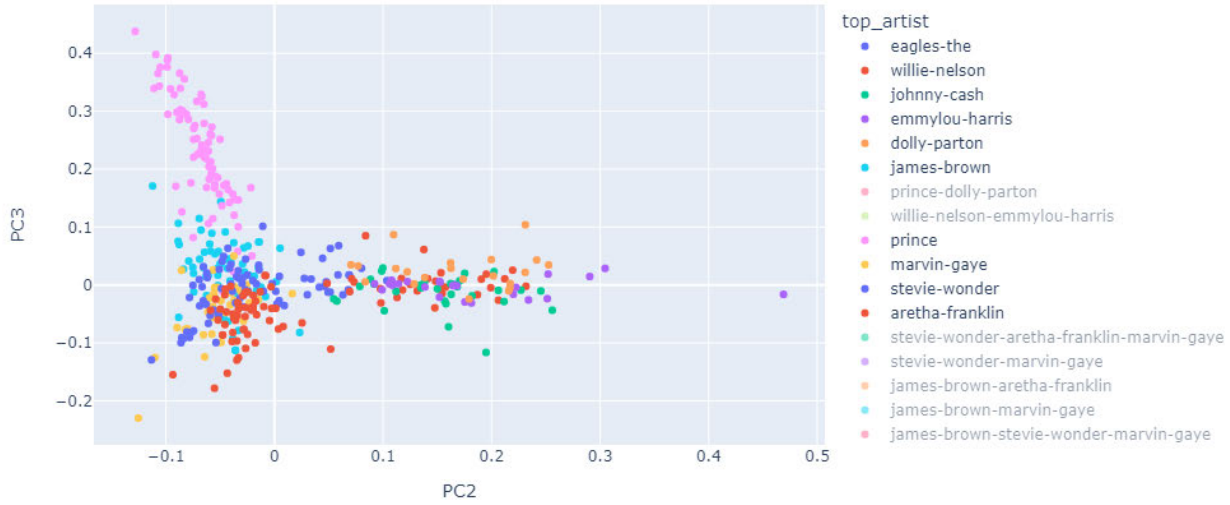

Figure 2. Scatter plot of articles along the second and third principal components

3. Vector Space Modeling

I created a document-term matrix representation of the corpus using TFIDF values from the term BOW. I applied principal component analysis (PCA) to the matrix and plotted articles along the first four principal components.

First, I searched for genre-level clustering. Figure 1 shows that articles separate by genre along the second principal component. It is incorrect to attribute meaning to any principal component; however, the second principal component seems to distinguish genre information in each article.

Next, I investigated clustering by artist. Figure 2 demonstrates that articles about the same artist tend to cluster together. Unlike the genre clusters, the artist clusters do not separate themselves along a single principal component.

I also looked for clusters based on article type and publication date. I was unable to identify any meaningfully groupings.

From this process, I conclude that the language used in music journalism is genre specific. The words used to describe country music are different from those used to describe soul music. In contrast, clustering of articles by artist is likely caused by the

occurrence of artist names and not due to specific language. The language used in different article types, such as reviews and interviews, is not distinct. There were no shifts in word-choice or style that would cause articles to cluster based on publication date.

While the document-term matrix is useful for examining properties of each article, it is not the only useful vector-space representation. I created a time-term matrix that encodes terms occurrence by publication date. The time-term matrix features a narrower vocabulary of only the top 20,000 terms by mean TFIDF. With a time-term matrix representation, I can use kernel density estimation (KDE) to visualize how terms fluctuate in popularity over time.

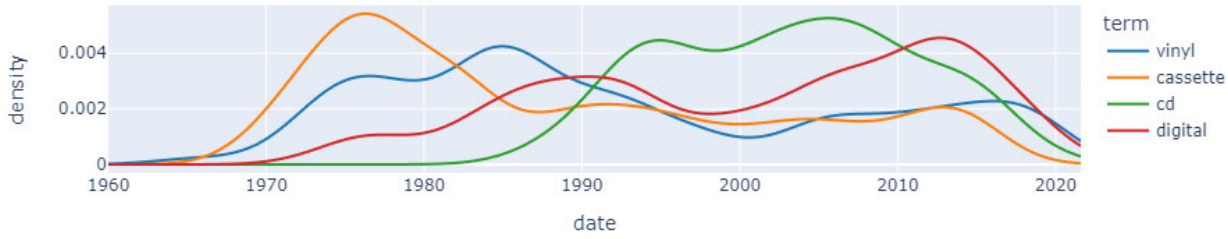

Figure 3. KDE plot of recording media over time.

Figure 4. KDE plot of supremes and diana ross over time.

Figure 3 shows the shift in recording media from 1960 to present. The term *cassette* peaks in the mid-1970s whereas the term *vinyl* doesn't peak until the 1980s. This runs counter to intuition since vinyl records predate cassette tapes by several years. Figure 4 compares the occurrence of *diana ross* to *supremes*. While Diana Ross never peaked as high as The Supremes, one could argue that Ross extended her cultural significance by going solo.

Using publication dates as the time-step for the time-term matrix potentially skews the KDE since articles are not uniformly distributed over time. For example, the corpus may contain only a handful of articles from a given month, and then hundreds of articles from the following month. Time periods that are dense with articles are necessarily dense with terms. To investigate the consequences of uneven document spacing, I created an additional time-term matrix where the timestep simply corresponds to the article's order in the corpus when sorted by publication date.

Figure 5 compares the calendar time scale to the corpus time scale. While there are certainly some quantitative differences, the graphs are qualitatively similar. This suggests that the distribution of articles over time is uniform enough to use publication date as a time scale.

Figure 5. KDE plots of various artist by calendar time and by corpus time

4. Principal Component Analysis

As previously discussed, principal component analysis was used to investigate clusters of articles in the document-term matrix. I formalize the principal component analysis by computing 12 principal components and associated loadings from the document-term matrix.

Figure 6 shows terms plotted against the third and fourth loadings. Term markers are colored according to their part of speech. Most of the terms are clustered around the origin, indicating low influence on the components. Proper nouns scatter away from the origin which means they have the strongest effect on the third and fourth components.

Why are proper nouns so influential on the principal components? This is likely due to the use of TFIDF values in the document-term matrix. Proper nouns, like artist names, have high TFIDF values since they occur frequently in the corpus, but are restricted to small subset of documents. PCA attempts to retain as much information as possible while reducing the dimensionality of the provided data. It seems logical that PCA relies on high information words to compute components.

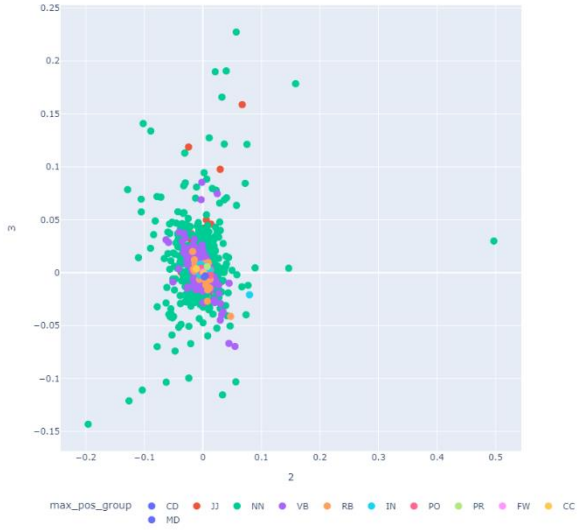

Figure 6. Scatter plot of terms against the third and fourth principal components.

5. Topic Modeling

I performed topic modeling to identify subtopics that may reveal any cultural insights that are not be detectable from vector-space modeling alone. I used Scikit-learn's implementation of latent Dirichlet allocation (LDA) with 20 topics. Topics were characterized by their most probable terms and by their most relevant terms. Relevance measures

how likely a term will appear under a topic and not appear under any other topic.

The top terms were not particularly insightful. The most probable terms in one topic were also probable in other topics. This phenomenon was limited to general terms like *music*, *group*, *song*, and *album*. In contrast, the relevant terms were more distinct, but harder to interpret as topics. This was because relevant terms were mostly names. Some topics were identifiable. For instance, *Toussaint* (Allen Toussaint), *Rebennak* (Dr. John), and *Neville* (Aaron Neville/Neville Brothers) are all notable musicians within the subgenre of New Orleans soul. Other topics were murkier. *Winehouse* (Amy Winehouse), *Debarge*, and *Braxton* (The Braxtons) were all the most relevant terms for a topic. Yet, it is unclear how any of these musical acts are related.

To demystify the topics further, I repeated LDA on each genre separately. I generated ten topics for each genre. The problems encountered in the full corpus persisted at the genre level. Surprisingly, some of the topics from the full corpus were preserved. For example, both the full corpus and the soul corpus have a topic for New Orleans musicians.

Once topics were generated, I applied hierarchical agglomerative clustering with ward linkage to determine which topics were the related. Figure 7 shows which topics in the soul corpus are most similar. Table 1 displays the labels for each of these topics.

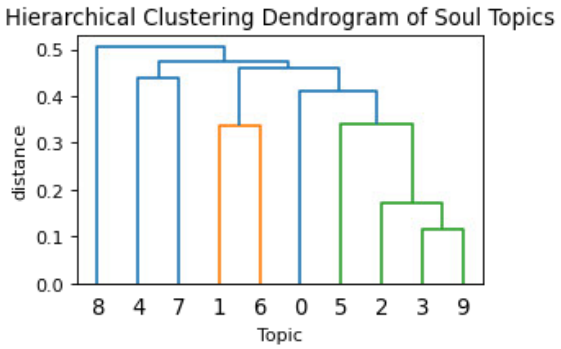

Figure 7. Dendrogram of soul music topics

topic_id	relevant_terms
0	heron hancock nile herons worrell defunkt
1	drifters kerr berns sigma chimes dionnes
2	hp linx tinas graves castor verdine
3	syreeta mt hardcastle jonzun colonel jasper
4	dangelo winehouse aaliyah macy neneh alicia
5	mavis cherrelle thelma hodges obrien pervis
6	jazzie gorrie bradley sar dowd ruffin
7	ulmer lidell encores schmaltz vandrosss choruses
8	toussaint meters neville dorsey rebennack nevi...
9	carey mariah monáe natalie lopez chaka

Table1. Relevant terms for soul topics.

Unfortunately, few conclusions can be drawn from the dendrogram. Topics 9 and 3 are the most closely related, but the relevant terms do not suggest a cause for their similarity. The jumble of proper nouns makes it challenging to see any patterns. This suggests that topic modeling is ill-suited for this corpus.

6. Word Embedding

I examined term similarity in the corpus through word embeddings. Specifically, I used the Gensim implementation of the Word2Vec algorithm. I chose a window size of four based on the results of several informal trials. After creating a word embedding, I clustered similar terms with t-distributed stochastic neighborhood embedding (t-SNE).

Figure 8 shows the top 2000 terms by DFIDF. As one might expect, terms tend to cluster based on parts of speech. This is intuitive—adjectives will occur in similar contexts to other adjectives. More interesting are the subgroupings within the parts of speech clusters. Proper nouns subdivide into names of people and locations. Nouns like *musician*, *vocalist*, and *performer* are more similar to names than to other common nouns. Adjectives like *instrumental* and *rhythmic* are located near nouns like *harmony* and *beats*.

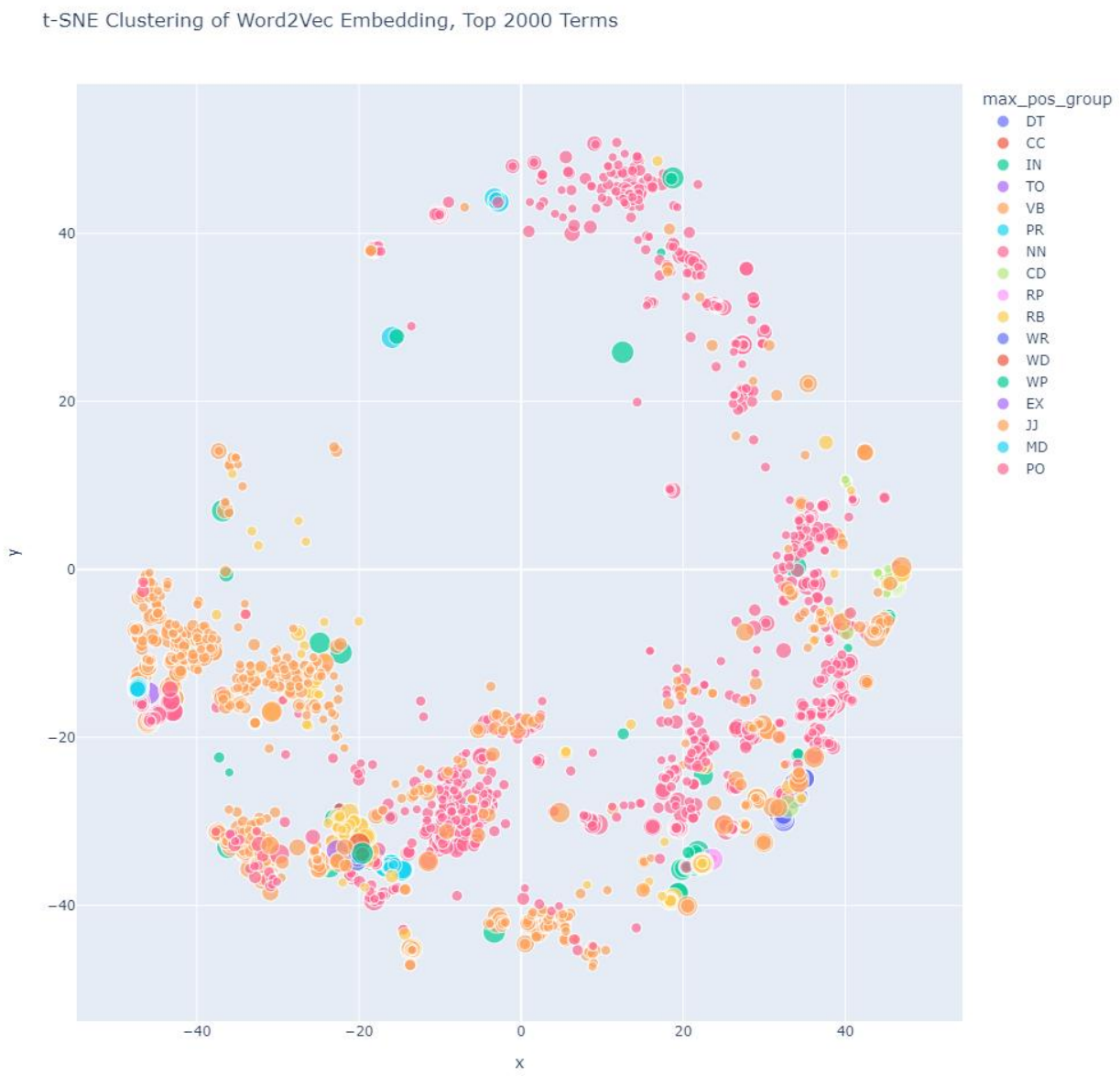

Figure 8. t-SNE plot of Word2Vec embedding of top 2000 terms by DFIDF

.

What does this word embedding suggest? Broadly, it shows what language is used in music journalism and how ideas are organized. It reveals less about the role of country and soul music in American culture, and more about the stylistic conventions of music publications.

7. Sentiment Analysis

I analyzed sentiment within the corpus using the NRC emotion lexicon, the NRC valence, arousal, and dominance lexicon, and the VADER sentiment analyzer. The NRC lexicons show that on average music articles are neutral to slightly positive in valence. The strongest emotional sentiments are joy, anticipation, and trust. Tagging with the NRC lexicon also allows us to visualize changes in sentiment over time. Interestingly, figure 9 shows that the valence of the corpus decreases over time. This could be due to the country and soul music falling out of favor.

Critics and journalists may subconsciously choose language that reflects shifts in popular opinion.

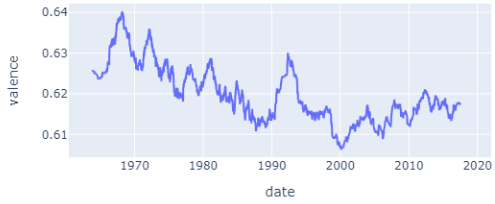

Figure 9. Rolling average of NRC valence

Unlike the NRC lexicons which were used to tag individual tokens in the corpus, the VADER sentiment analyzer is applied to whole sentences. VADER does an adequate job of judging the sentiment of sentences with extreme valence. Occasionally, it mislabels more neutral sentences. When the sentiments of each sentence are averaged over an article, the results are consistent with the NRC Lexicons.

Figure 10 depicts the average sentiment of articles based on type. Reviews tend to be more positive than interviews which are more positive than overviews and retrospectives. This is not wholly surprising. Music reviews often function as recommendations to the reader. Quality albums and performances are more likely to be written about than neutral or poor musical output. Interviews often feature conversation, which can vary in valence.

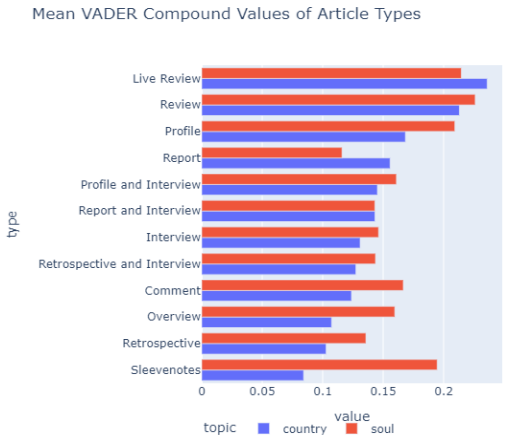

Figure 10. Average valence of article types

8. Conclusion

Initially, this analysis sought to characterize the American music tradition by surfacing patterns from magazine articles. At the outset, it did not occur to me to ask if articles were even suitable for this task. With each analytical method, I had trouble relating my results to American culture at large. Despite this struggle, the analysis was not unfruitful. The methods explored in this paper reveal a distinct style and language used in music journalism. Through this research, I have come to realize that music journalism is a component of culture just as worthy of study as music, art, and social behavior.

Given more time, I would focus my analysis specifically on music criticism. I would like to include additional genres of American music. I would attempt to resolve the issues caused by high frequency proper nouns by anonymizing artists' names. Lastly, I would like to balance the articles by publication date so that they are uniformly distributed over time.

9. Resources

Alvarado, R. (2022). DS5001-2022-01, Github Repository. https://github.com/ontoligent/DS5001-2022-01

Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Mohammad, S. (2018). Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL).

Mohammad, S., & Turney, P. (2013). Crowdsourcing a Word-Emotion Association Lexicon. Computational Intelligence, 29(3), 436–465.

Music of the United States. (2022, April 29). In Wikipedia. https://en.wikipedia.org/wiki/Music_of_the_United_States

Music reviews, articles & interviews from the online library of pop writing. Rock's Backpages. Retrieved March 26, 2022, from https://www.rocksbackpages.com/

Pedregosa *et al*. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Reback *et al.* (2022). pandas-dev/pandas: Pandas 1.4.1 (v1.4.1). Zenodo. https://doi.org/10.5281/zenodo.6053272

Rehurek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (pp. 45–50). ELRA.