Beyond the Melody: A Comparative Text Analysis on Taylor's Greatest Hits

Group members: Annika Salpukas, Manvi Kottakota, Ruby Mason

Inspiration

Taylor Swift is one of the world's greatest pop stars with a massive fan base made up of millions of followers. "Swifties" as they are called spend hours listening to Swift's hundreds of songs, showing their true dedication to her music. Swift stands not only as a great singer and songwriter, but also as a role model and symbol of feminism in today's society.

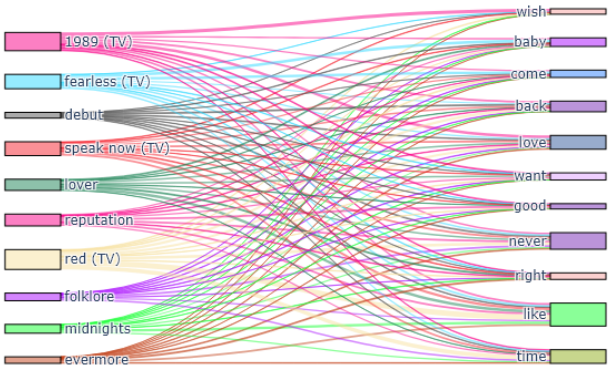
This project aims to leverage data science skills such as natural language processing in order to take a deeper dive behind the lyrics in some of her most popular songs. These text files used in the project contain the full lyrics of the songs off of 10 of her albums, and are cleaned to produce the desired results. The objective is to uncover potential connections and trends between Swift's work that have contributed to her massive success, shaping the idol that she is known as today.

Process

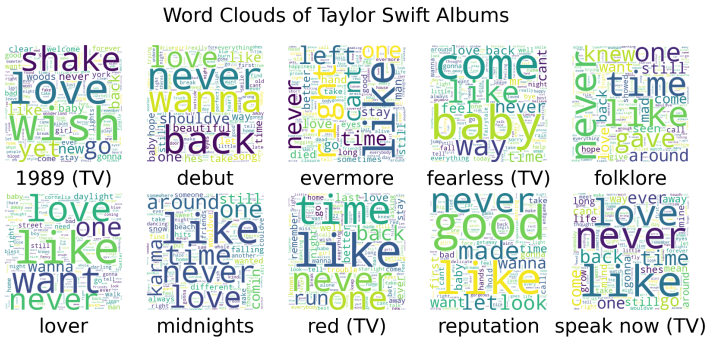
Data in the form of text files is loaded into the Textastic class where it proceeds to be cleaned and grouped based on different statistics. Commonly used words known as "stop words" have been taken out of the data to prevent statistics such as word frequency from being skewed. Additional words that commonly appear in songs such as "ooh," "la," and "woah" have been removed from the data as well. The files used are already in a .txt format, so the default parser is

appropriate to use in this case. The parser returns a dictionary of different calculated statistics that are then stored in another dictionary containing all of the compared text files. Statistics that were relevant for the produced visualizations were word count and varying measures of sentiment scores, but additional stats such as mean word length, and how many unique words occur were also collected. These data points were then used to create relevant visualizations for comparing the different albums.

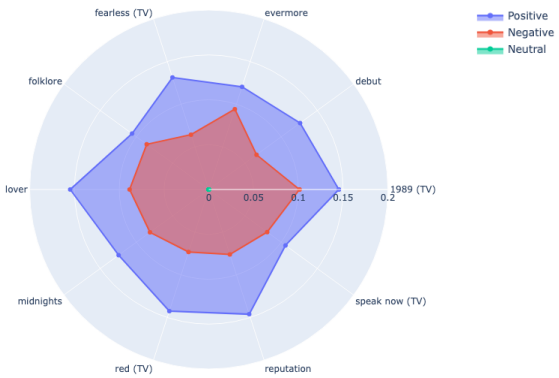
Insights



The first visualization created uses a Sankey diagram to track which albums use which words the most. Since data from ten albums were loaded for this visualization, the top two most commonly used words from each album were linked to them to maintain the readability of the plot. Here, the thickness of a line connecting an album to a word represents the frequency of that word in the album. Some of the strongest connections observed were between the “Red” album and words such as “time” and “never”, which makes sense in the context of the album’s focus on the complex feelings resulting from past heartbreaks. Other stronger connections included those between “1989 (Taylor’s Version)” and words like “wish” and “love”, which may be explained by the themes surrounding coming of age and finding young love. The re-recorded albums seem to have the most connections to common words, which is possibly due to their length but also their dependence on chorus as opposed to her albums in the country and folk genres.



The second visualization creates word clouds for each of the texts and displays them in an array of subplots. This visualization displays the most commonly used words across each album, with the size of the word representative of how often it appears. From this it is clear that the words "love", "never", and "time" are common across all ten of the albums. This is validated by the fact that she writes many love songs which sing about love, time, "never getting back together" and so on. This visualization is both informative and displayed in a way that is appealing to look at, making it a good representation of the data.



Our third visualization, a radar chart, portrays a sentiment analysis of Taylor Swift's album collection. Swift is known for each of her albums portraying a unique set of experiences and emotion that pertain to her various life experiences. Through this visualization, we were able to dissect the emotional and sentimental nuances across her albums, where the length of these areas from the center serves as a gauge for the intensity of sentiment, allowing for ease of comparison. It becomes apparent that certain albums, like Lover, were especially positive as reflected in the word use of the word "love" in our previous word clouds.

We can also observe that her earliest albums Taylor Swift (Debut) and Fearless, likely reflected lighter themes regarding her positive experiences as a teenager in love considering the significantly higher positive sentiment over negative. The notably higher negative sentiment in Evermore and 1989 can be attributed to a harder period of Swift's life with the album covering loss of family members, struggles in relationships, and closure in withering friendships.

Conclusions

Looking carefully at the language choice that Swift makes can give insight to what makes her such a successful artist. Many of the words/themes that arose from the data had to do with love, which is a very popular topic to sing about and aligns with the success that Swift sees. Text analysis also verified Swift's themes within her albums such as love, friendships, and life experiences as the words associated with them followed trends within the specific albums. Analyzing her work through a data science lens gives a deeper understanding of Swift's impressive career.

Author contributions

Annika Salpukas: Implemented functionality for building sankey diagrams from loaded data, helped compile lyrical data by album and stopwords, wrote code for loading stopwords.
Ruby Mason: Implemented word cloud visualization by adding on to default parser and creating word_clouds function. Wrote and configured parsing exceptions. Drafted intro and process sections of report.

Manvi Kottakota: Implemented the radar chart function and visualization, adding required data for semantic analysis onto default parser. Wrote the code for the default parser besides the word cloud functionality code within it. Wrote out the description for the third visualization.

Repository link: https://github.khoury.northeastern.edu/kkmanvi123/nlp-framework.git