Gender analysis in American rap using word embedding

Owein Dourneau
Université de technologie de Belfort Montbéliard (UTBM)
owein.dourneau@gmail.com

Abstract

Word embeddings, a machine-learning technique, assign a numerical vector to each word in a given vocabulary. The geometric relationship between these vectors captures the semantic similarity and relationship between the words they represent. Inspired by the method detailed by Nikhil Garg et al. , the aim of this project is to quantify the gender stereotypes and bias presents in the American rap and their potential changes over 3 decades : 1900's, 2000's and 2010's. Word embeddings trained on the lyrics of songs from rappers of theses decades suggest that some concepts are, independently of the decades, gendered but also that while gender may play a role in the language used in American rap music, it does not appear to have changed significantly and differently than for a corpus of headlines of the New York Times over the past three decades.

1 Introduction

In last decades, the study of genders have become a particularly highly and salient debated area of research. This is due in part to the current cultural and political conversations surrounding gender identity and equality, as well as the recognition that gender plays a significant role in shaping individual experiences and societal structures. As a result, understanding the ways in which gender is represented and perpetuated in various forms of media and communication has become increasingly important.

As demonstrated by Nikhil Garg et al. (Garg et al., 2018) word embeddings were among the usual methods such as primarily leverage human surveys , dictionary and qualitative analysis, or in-depth knowledge of different languages performing well to capture stereotypes.

The purpose of this project is to use word embedding to investigate the representation of gender in American rap music over 3 lasts decades. The study uses this method in order to understand if trends in the representation of gender in American rap music align with those found in a corpus of headings of the New York Times. By comparing the two, it is expected to uncover any similarities or differences in the representation of gender in these two sources, and track how they may have changed over time. This can provide a nuanced understanding of the representation of gender in American rap music and how it relates to societal norms and expectations.

2 Theory

Word embedding is a computational method that maps words or phrases to a high-dimensional vector space, where the distance between vectors reflects semantic similarity. The theoretical foundation of this approach is based on the idea that words that are semantically similar should be represented by vectors that are close to each other in the vector space.

One of the main theoretical concepts behind word embeddings is distributional semantics, which posits that words that occur in similar contexts tend to be semantically similar, thus, the context-based co-occurrence information of words is captured through the word embedding, this is done through training algorithms on large corpora of text.

This paper use the GloVe word embeddings method (Pennington et al., 2014) to create the vectors. In the GloVe method, the model learns a vector representation for each word, such that the dot product of the vectors for two words is proportional to the logarithm of the number of times the words co-occur in the corpus. The main intuition underlying the model is the simple observation that ratios of word-word co-occurrence probabilities have the potential for encoding some form of meaning.

3 Data

3.1 Rap lyrics data

The first part of this project's data-set is constituted of exactly 276 rappers' lyrics over 3 decades : 1990's, 2000's and 2010's. For each rappers the 10 most famous songs have been fetched making a total of 2760 lyrics text. The distribution over decades is the following :

Decade	Number of rappers
1990	101
2000	100
2010	75

Table 1: Number of rappers by decades

The chosen rappers have been selected using the weekly Hot Rap Songs Billboard Ranking (Billboard, 2023). Taking rappers appearing the most in the ranking as a single artist (collaboration have been ignored) for each decades.

3.2 New York Times Data

The second part of the data-set is constituted of the 5000 first headings of the New York Times for each year since 1990 aggregated by decades.

3.3 Preprocessing

Each documents have been preprocessed following this schema :

- Remove punctuation and non-alphabetic characters
- Lowercase
- Remove stop words

And then two different versions have been tested, one with lemmatization and one without.

The corpus of documents have been processed this way because punctuation, non-alphabetic characters and stop words were not adding any meanings. For the lemmatization part, a comparison between the results obtained with and without is discussed further in the results section (5).

4 Method

4.1 Data Gathering

Since the data-set was created for the project, multiple sources were used to gather the data.

Gather the list of rappers of the billboard ranking has been done by scrapping the website since there is no API provided. The script (BillboardScrapper/scrapper.py) use the library BeautifulSoup to do it.

Retrieve the songs for each rappers is then done by using the API of the website RapGenius (RapGeniusScrapper/scrapper.py). The lyrics are then stored by decades by artists.

For the New York Times headings, the data have been gathered using the New York Times API and then stored on Kaggle.

4.2 Training the word embedding

The first step after gathering all the data is to create the word embeddings. To do that the GloVe method is used as described by (Pennington). By first creating a file where each line corresponds to one of the document in the corpus (lyrics from a song or a corpus of headings) and then create a word-word co-occurence matrix with it. Using this co-occurence matrix to create the vectors and the vocabulary in two separate files.

In order to compare between the decades, 6 different vector spaces of 50 dimensions have been created (1 per decades for rap lyrics and New York times headings). This step can be done using the file GloVe/create_vectors.sh and have been executed for this project on a Unix system.

4.3 Loading the word embedding

The second step to then analyze the word embeddings as convenient as possible is to create a new Spacy english blank vocabulary and add the customs vectors to it. This has been done to use the different functions provided by Spacy to deal with the vectors.

4.4 Analysis angles and plot showing

All the plots are then shown executing the cells of the bias_analysis.ipynib.
The analysis angles for this project have been each time separated by decades and gender. Here are the different data computed :

- The Part of Speech tags types and counts.
- The overall positiveness of the 100 closest (Cosine similarity wise) adjectives, nouns, verbs.
- The distance to certain concepts.

The positiveness of each words have been evaluated using the SentimentIntensityAnalyzer from the library NLTK with the vader_lexicon. The scale going from -1 corresponding a highly negative word and +1 connoted to a highly positive word.

The cosine similarity have been computed each time against a centroid vector of different word related to female gender (women, girl) and male gender (men, boy).

5 Results

5.1 Gender analysis in the rap lyrics

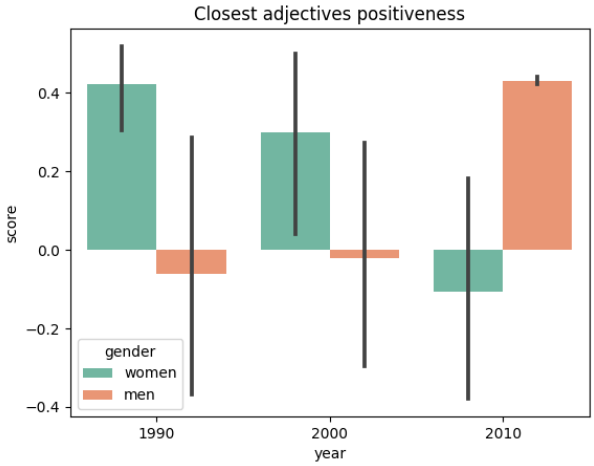

Figure 1: Closest adjectives positiveness without lemmetization

The fig. 1 shows two different trends when comparing by gender. Female adjectives positiveness decrease over the decades whereas the male one tends to be the same for the 2 first decades then increasing drastically. However the errors bars representing the confidence interval tends to show a global high variability of the data .

The fig. 2 shows an overall similarity in the positiveness of the nouns over the gender. The positiveness seems to be decreasing over the decades. We could estimate that the 1990 data having an error bar relatively smaller than the 2000 decade model better the reality. Moreover, in the 2010 decade, the men error bar is also relatively smaller than for the 2000 decade but also smaller than the one women's one. So the decreasing trends for the men looks more robust.

The fig. 3 shows two inverse trends for the genders. The female tends to be more positive over the decades whereas the male one seems to be more negative. However, the errors bar being particularly big, makes the result hardly interpretable.

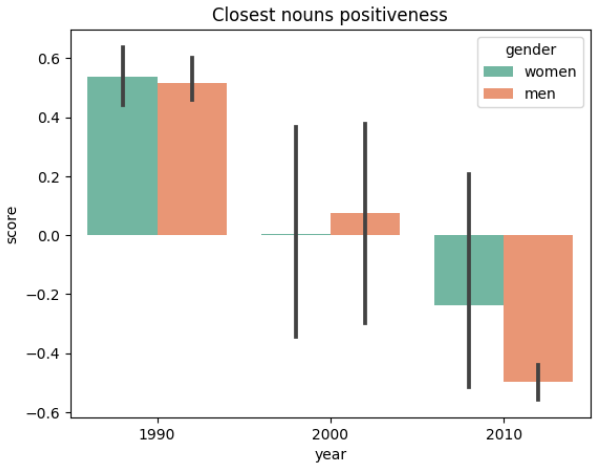

Figure 2: Closest nouns positiveness without lemmetization

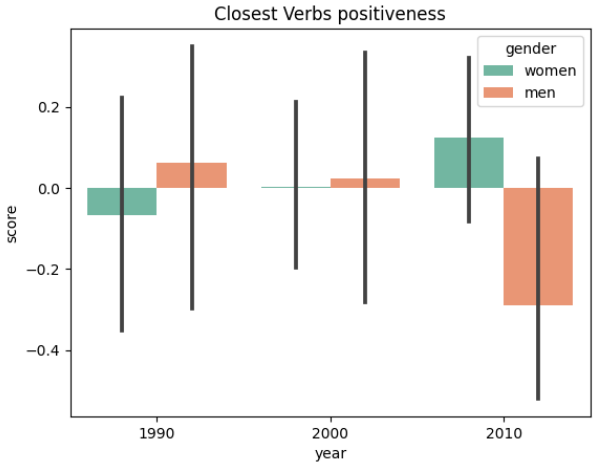

Figure 3: Closest verbs positiveness without lemmetization

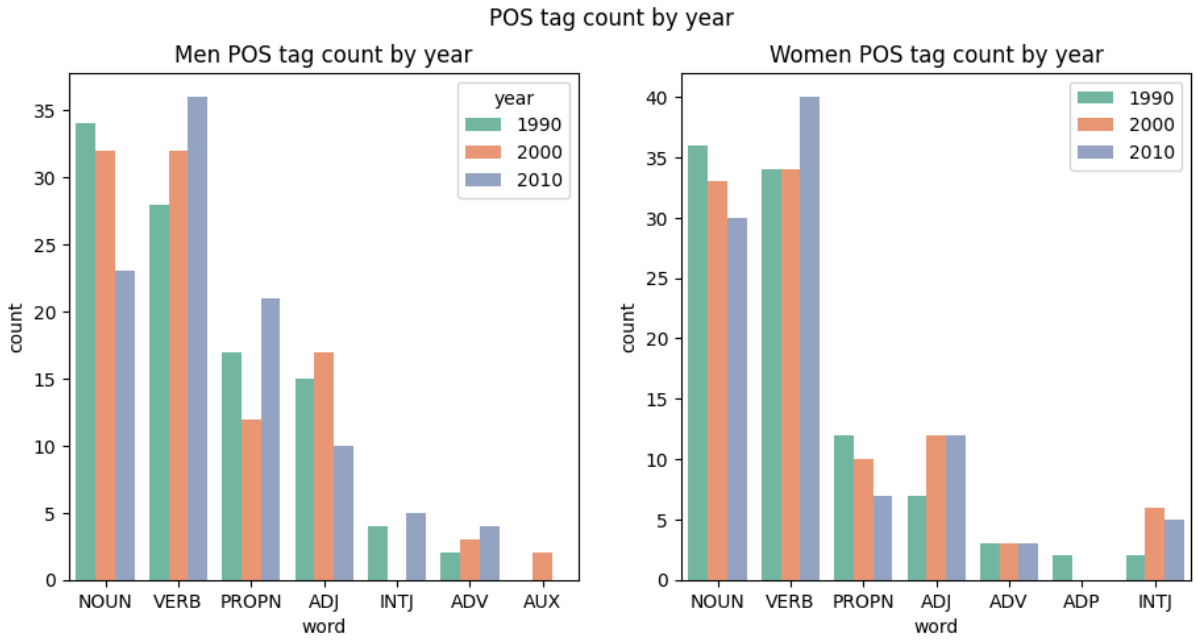

Figure 4: POS tag count by year without lemmetization

The fig. 4 shows that the nouns and verbs are the most present in the lyrics and we can observe a non-gender related decreasing presence of the nouns in the closest words over the decades. The distribution is globally the same over the gender.

The fig. 5 and fig. 7 shows the similarity difference that have been computed as the difference : female_similarity_with_the_word - male_similarity_with_the_word. A high value indicate a concept closer to the female vector whereas a negative one indicate a concept closer to the male vector.

By similarity, we refer to the parlance of the library Spacy. The similarity is computed as the Cosine similarity.

We can see for instance that the notions of "work", "love" and "child" are, independently of the decades, related to the female vector whereas the "surprise", "intelligence", "wisdom" and "strength" are more related to the male one.

We can also see a limit, when comparing the vectors "child" and "children" that are not related to the same gender vector.

Overall, the data about positiveness seems to be hard to interpret because of the high variability of the data. The POS tags analysis seems not very relevant too. However the concept similarity analysis seems to be more convincing but shown some limits when taking two concept that should be similar in meaning and thus related to the same gender vector but aren't.

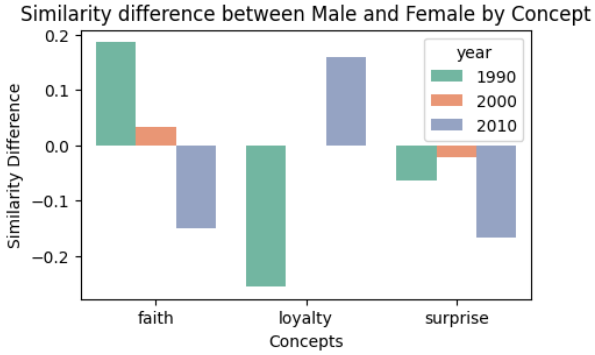

Figure 5: Similarity difference between Male and Female by concept without lemmatizing

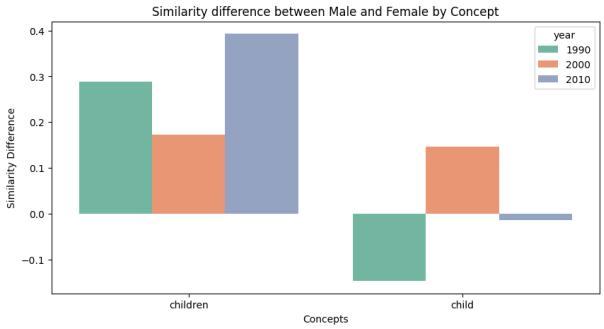

Figure 6: Similarity difference between Male and Female by concept, focus on child and children

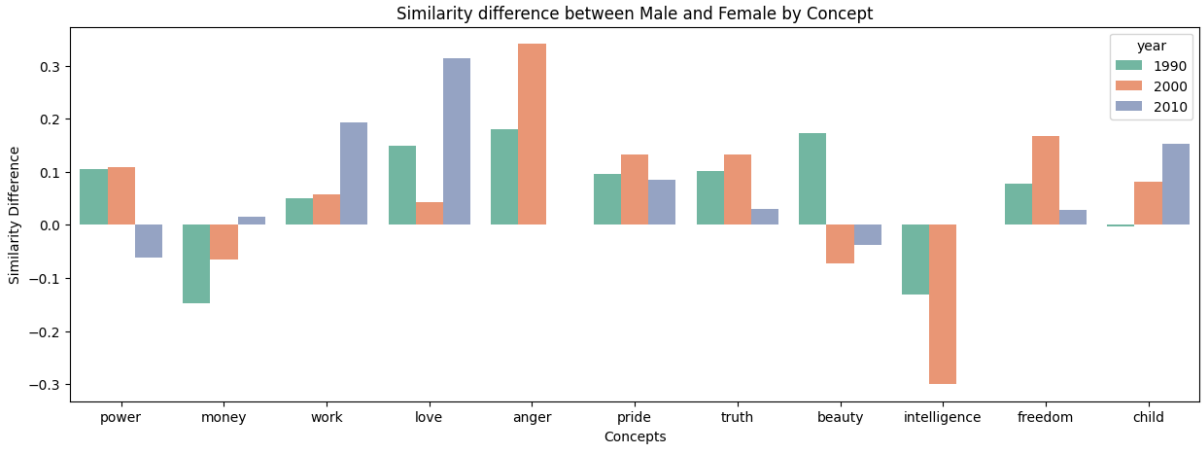

Figure 7: Similarity difference between Male and Female by concept without lemmatizing

5.2 Comparison with the lemmatized data

In order to see if the lemmatization process was affecting the results, I computed the same analysis but by adding lemmatization to the preprocessing.

For the positiveness analysis, the result for all the angles are different. For instance, we can see on the fig. 8 that the result are totally different than from the fig. 2.

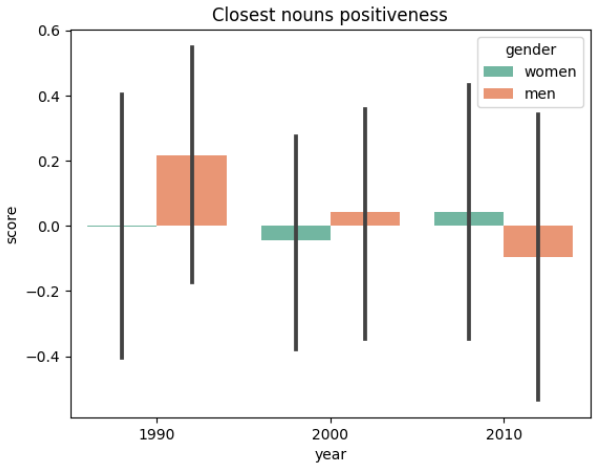

Figure 8: Closest noun positiveness with lemmatizing

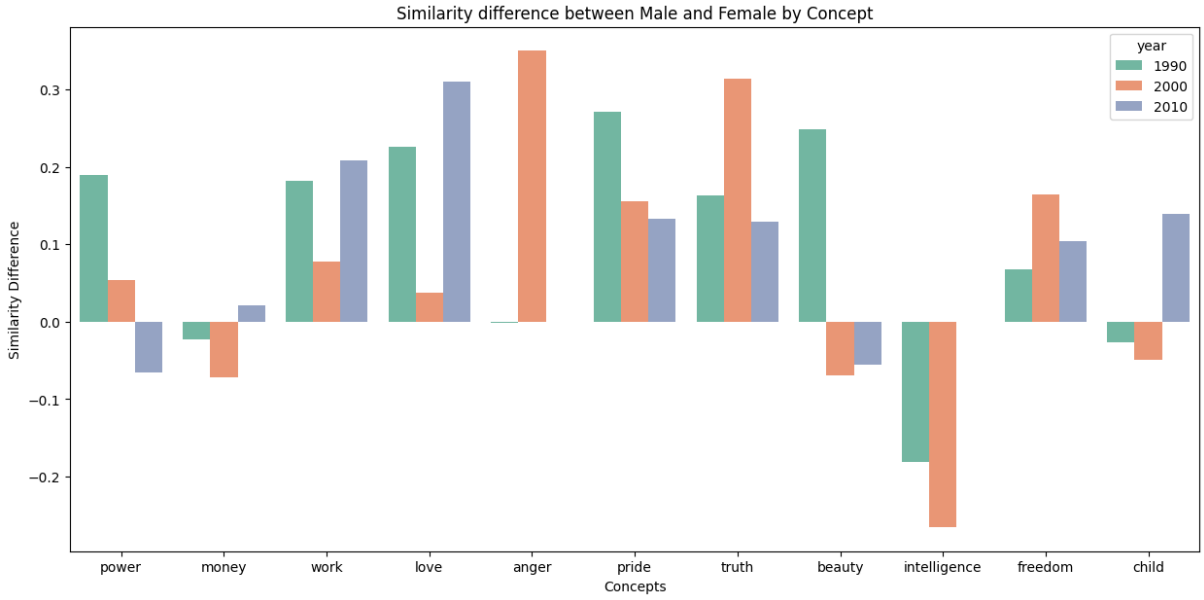

Figure 9: Similarity difference between Male and Female by concept with lemmatizing

However, the results for the concept similarity are slightly the same as we can see on the fig. 9 but child vector have been modified a bit and tend to be more related to male vector for the first 2 decades.

The comparison between lemmatize and unlemmatized shows that the positiveness analysis is not usable to interpret since the results are changing but seems to reinforce the trustness that we can have in the concept similarity analysis.

5.3 Comparison with the New York times headings data

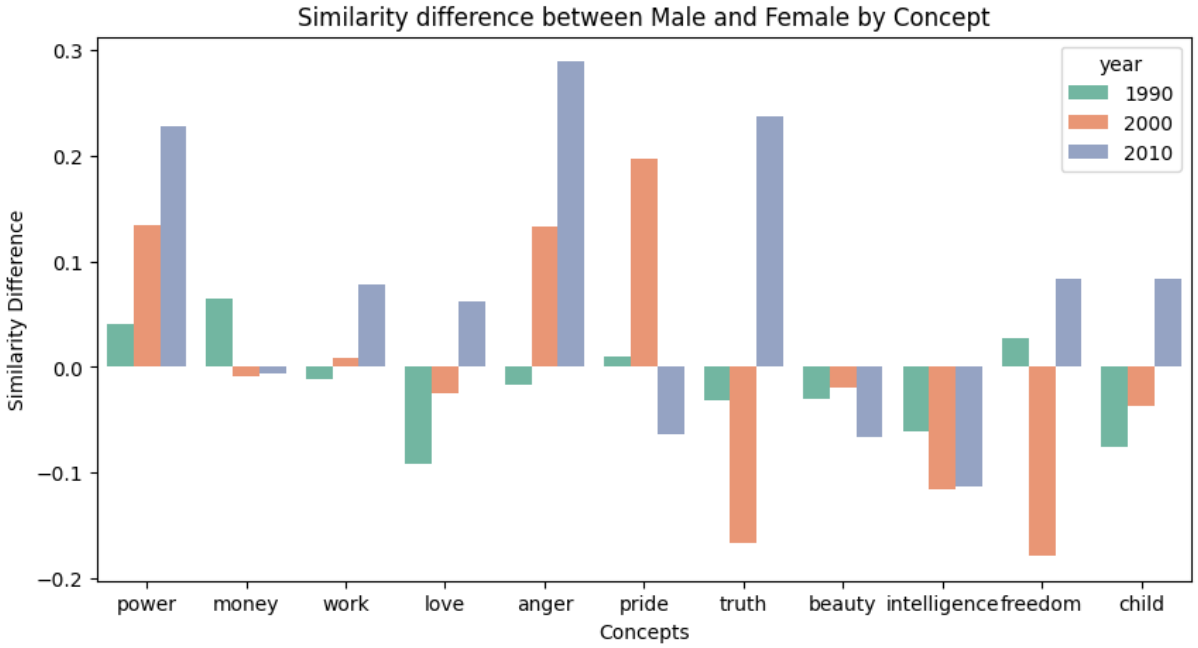

Figure 10: Similarity difference between Male and Female by concept for the NYT headings

By comparing fig. 9 and the fig. 10, we can see some similarities between the distance of gender vectors and concept vectors from the NYT headings and the rap lyrics ones. For instance we can see that "work" is closer to the female vector on the two last decades.
But also some dissimilarities in the trends, for instance for the "power" vector, where it's more and more related to the female vector in the NYT headings. The "child" vector seems to evolve in the same direction for both of the corpus.

6 Discussion

6.1 The Part of Speech tags types and counts

This angle of analysis didn't really show any difference gender wise and therefore didn't bring any really meaningful information.

6.2 The overall positiveness of the 100 closest (Cosine similarity wise) adjectives, nouns, verbs.

The results show that this angle of analysis was not meaningful for multiple reasons :
The high variance of the data shown by the large error bar on the bar plot, certainly due the size of the data-set.
The way of measuring the positiveness is also debatable. Indeed, it has been done using the vader lexicon but since rap has its very own vocabulary and word meaning, some words could have been miss scored. And also, this way of doing does not take in consideration the evolution of the potential positiveness of the words over the years.

6.3 The similarity to certain concepts to gender vectors.

The results obtained, would suggest that some concepts are gendered over all the decades. Interpreting the results, we could say that they corroborate certain of the typical stereotypes of the society such as "love" or "child" linked to female and "intelligence" and "money" to male. Overall, we can see that the results obtained are going in the same directions as the one obtained in the literature (Garg et al., 2018).

We can also see some unexpected results in the sens that we would expect them, even more in the rap, to be related to the opposite gender such as the word "beauty" that is for the 2 lasts decades more related to the male vector.

The comparisons with the NYT headings corpus would suggest that the rap does not break the trend on certain concept such as "intelligence" or "child" but also goes against it especially on the "power".

Given the result, it's hard to conclude on the fact that rap is going against the general trend or is following it since some representations are shared and some other aren't.

6.4 Overall bias

We need to take care about all the bias this project has. First, the domain of the rap is a highly unequally gendered mostly represented by male artist and the data set do not break this rule. However since the aim of this project was to quantify the stereotypes, the unequal gender distribution could be seen as a part of it.

Secondly, as seen in the literature (Williams and Best, 1990), the stereotypes varies with the culture and only the American rap have been taken in consideration trough this project.

Moreover, the size of the data-set is way smaller than the usual one used by the word embedding algorithms. Therefore the results cannot be as significant as the one seen in the literature.

The process of selection of the rappers and the songs is questionable as well since it's rely on the trust we give to the Billboard ranking system to evaluation the most famous artists.

The prepossessing method could also be discussed because it has shown for instance that lemmatizing words changed a significant part of the results.

Conclusion

In conclusion, the gender analysis using word embeddings in American rap music from 1990, 2000, and 2010 revealed that no significant trends were identified when compared to a corpus of headlines from the New York Times.
This suggests that while gender may play a role in the language used in American rap music, it does not appear to have changed significantly and differently than for the corpus of headlines over the past three decades.
However, I was able to observe the importance of data, particularly the influence of quantity. The difficulty in retrieving and aggregating them in a way that can lead to conclusions. I was able to deepen my understanding of the concept of word embeddings and more broadly text-mining at all stages of a project.

References

Billboard. 2023. Hot rap songs.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic ... - pnas.

Jeffrey Pennington. Glove method.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

John E Williams and Deborah L Best. 1990. *Measuring sex stereotypes: A multination study, Rev*. Sage Publications, Inc.