**Hip-Hop Lyrics vs. Geography: Linguistic Analysis**

**Introduction**

Hip-hop, as a cultural and musical phenomenon, is deeply rooted in geography. From the East Coast’s lyrical intensity to the West Coast’s laid-back storytelling, regional differences have been recognized anecdotally for decades. This project aims to quantitatively analyze these differences by investigating whether linguistic patterns in hip-hop lyrics correlate with the geographical locations of the artists. More specifically, we explore whether large language models (LLMs) like BERT encode information predictive of geography.

The research questions guiding this project are as follows:

1. Can we predict the geographical location of a rapper based on the lyrics of their songs?
2. What linguistic patterns are predictive of location?
3. To what extent do pre-trained LLMs encode information relevant to geography?

While prior studies have mapped the geography of hip-hop, few have leveraged computational methods like LLM embeddings to examine regional linguistic styles in lyrics. By addressing this gap, this project combines insights from computational linguistics, cultural studies, and machine learning to contribute to our understanding of hip-hop’s regional diversity.

**Literature Review**

Several studies and resources provide a foundation for this project. B. Piri’s article *The Geography of Hip-Hop* highlights the regional variations in hip-hop’s sound and lyrics, emphasizing environmental influences like car culture on the West Coast and trap music’s rise in the South. Similarly, *The Pudding*’s interactive website visualizes hip-hop artists in a semantic space based on word usage but lacks a geographical component.

Other works, like *MCFlow: A Digital Corpus of Rap Transcriptions and African American Vernacular English (AAVE) Features in Rap Lyrics: A Case Study*, focus on rap’s linguistic and musical dimensions. However, these studies do not explore the relationship between lyrics and geography using modern machine learning models.

Additionally, fields like stylometric analysis (linguistic style attribution) and dialectometry (computational dialect analysis) offer methods that could complement this project. While these techniques provide valuable insights, this project uniquely combines them with pre-trained LLM embeddings and machine learning to analyze the geography of hip-hop lyrics.

**Data and Preprocessing**

The dataset used in this project comes from the *Kaggle Hip-Hop Encounters Data Science* dataset, containing lyrics from 36 famous rappers. Each file compiles lyrics from an artist’s released songs, scraped from the Genius API. To assign geographical labels, each rapper was associated with their birthplace or the city where they began their career, stored in cities.json. Latitude and longitude coordinates for these cities were obtained using the geopy library and saved in city\_coords.json.

The data were preprocessed to resolve imperfections in the original dataset and prepare for further analysis.

1. Cleaning lyrics: Removing non-lyrical content (e.g., track listings, screenplay) and standardizing text using the unidecode library.
2. Embedding generation: Lyrics were embedded into a 768-dimensional feature space using BERT (bert-base-uncased). Each rapper’s lyrics were averaged across tokens to create a fixed-size embedding vector from BERT’s last hidden state vector.
3. Dimensionality reduction: Principal Component Analysis (PCA) was applied to reduce embeddings, retaining the most informative dimensions for downstream tasks.

**Methods**

***Attempt 1: Regression with PCA and Random Forest***

In the first approach:

1. PCA reduced the embeddings to three dimensions for visualization and to 10 and 29 dimensions for regression.
2. A Random Forest regression model was trained to predict latitude and longitude based on the reduced embeddings. Hyperparameters were optimized using grid search, and performance was evaluated using R^2 scores and visual inspection of predicted vs. actual locations.

***Attempt 2: Lasso Regression, Hierarchical Classification, and Data Cleaning***

The second approach focused on lasso regression and hierarchical classification, as well as resolving data cleaning issues with the first approach:

1. Lasso reduced the embeddings to 111 dimensions for two additional regression models: ElasticNet and another Random Forest regression model. Hyperparameters for each were optimized using grid search, and performance was once again evaluated using R^2 scores and visual inspection of predicted vs. actual locations.
2. A custom PyTorch model predicted geographic hierarchy (region → state → city) using shared and task-specific layers. The model trained on embeddings, optimizing separate cross-entropy losses for each hierarchical level.
3. Extra cleaning steps resolved data shortcomings of the first approach.

***Evaluation Metrics***

Models were evaluated using:

1. R^2 scores: For regression models, though negative scores indicated poor performance.
2. Visual analysis: Maps of predicted vs. actual locations provided qualitative insights.
3. Classification accuracy: Accuracy was reported for hierarchical models at the region, state, and city levels.

**Results**

***Attempt 1 Results***

*3D PCA Visualization*

A 3D scatterplot of embeddings revealed limited clustering by geography. For example, Kendrick Lamar and Kanye West (from California and Illinois, respectively) formed a small cluster. Artists from different regions often overlapped in the reduced embedding space.

*PCA + Random Forest Regression:*

* 10 components: The model achieved a cross-validated R^2 score of -5.18. Predicted locations were biased towards the center of the map, with poor generalization to unseen artists like Diddy.
* 29 components (95% variance): The R^2 improved to -2.82, and predicted locations were closer to actual locations. Visual analysis showed some alignment, but many predictions remained inaccurate.

***Attempt 2 Results***

*ElasticNet Regression*

ElasticNet, a regression method that combines the L1 and L2 penalties of the Lasso and Ridge methods to improve regularization and feature selection, produced sparse models with a cross-validated R^2 of -0.97. While better than Attempt 1, the model still underperformed in explaining geographic variance, especially in areas less represented in the training data.

*Lasso Feature Selection + Random Forest*

Combining Lasso for feature selection with Random Forest regression yielded an R^2 of -0.70. This hybrid approach slightly improved predictions.

*PCA + Random Forest Regression*

Applying Random Forest on reduced embeddings from PCA with cleaned data and 29 components yielded small further improvement, with an R^2 of -0.66.

*Hierarchical Classification*

The hierarchical model struggled, achieving:

* Region Accuracy: 12.5%
* State Accuracy: 12.5%
* City Accuracy: 0.0%

Notably, the model misclassified Snoop Dogg (California) as being from Harlem, New York.

**Discussion**

All regression models exhibited predictions biased toward the Center-East of the U.S., a manifestation of *regression to the mean*. This behavior occurs when the models, trained on geographically diverse data, minimize their loss function (e.g., Mean Squared Error) by predicting locations closer to the geographic center of the training data distribution. The effect is particularly pronounced when data points are sparse or when the models lack sufficient complexity to capture extreme variations. As a result, predictions for rappers from regions with fewer representatives in the dataset (e.g., the West Coast or Southern U.S.) were often inaccurately biased toward more centrally located areas, such as the Midwest or the Eastern Seaboard.

I had hoped to fix this by using a classification model, but this approach introduced its own set of challenges. The hierarchical classification model I implemented struggled to achieve meaningful accuracy, particularly at the state and city levels, where the data was sparsest. The region-level predictions were slightly better, but still biased toward the East Coast and Midwest regions, reflecting the imbalanced distribution of rappers in the dataset. Furthermore, the model’s reliance on embeddings that were not fine-tuned for geographic or cultural distinctions likely limited its ability to learn the nuanced linguistic patterns needed to differentiate between locations. The hierarchical structure added complexity to the task, but the lack of strong signal in the data and the small number of rappers per location resulted in predictions that often defaulted to the most frequent or centrally located categories. These results suggest that without significant improvements in data quality, model architecture, or task-specific fine-tuning, a purely classification-based approach may not resolve the geographic biases inherent in the regression models.

**Future Work**

This project revealed challenges in predicting geography from hip-hop lyrics but also highlighted several promising areas for improvement.

1. Pre-trained BERT models could be fine-tuned on a dataset of rap lyrics to better capture regional slang, AAVE features, and stylistic nuances specific to hip-hop.
2. Collecting a larger and more geographically diverse dataset would address regional imbalances and improve model training, especially for underrepresented regions like the West Coast and South.
3. Future work could experiment with embeddings from models already fine-tuned on song lyrics (e.g., RoBERTa or GPT) or use clustering methods to reveal latent geographic patterns.
4. A hybrid model could first classify rappers into broad regions and then refine predictions with regression, potentially improving accuracy and reducing bias.
5. Probing embeddings to identify linguistic features that drive predictions could reveal how models encode geographic information and guide future feature design. This could become an interesting research project in interpretability.
6. Clustering techniques like k-means or t-SNE could uncover natural groupings in the lyrics that align with geography.
7. The current process could be reversed by training models to generate lyrics based on a specified location or region.
8. The methods used here could be applied to study other dimensions of hip-hop, such as temporal trends or subgenre distinctions.

**Conclusion**

This project explored the relationship between hip-hop lyrics and geography, leveraging pre-trained language models and machine learning to predict the location of artists based on their lyrics. While current models underperformed, exhibiting a bias toward central locations and low prediction accuracy, the findings suggest that geographic linguistic patterns exist and can potentially be captured with improved methods. The visualizations and analyses indicate that some geographic distinctions may be encoded in lyric embeddings, but more robust datasets and fine-tuned models are needed to fully capture these nuances. By bridging machine learning, cultural studies, and linguistics, this work sets the foundation for future explorations of hip-hop’s regional diversity.

More details and documentation can be found in the [GitHub repository](https://github.com/jpsank/engl353).

**Informal Reflection**

I chose this project centering on hip hop because I wanted to address a subject that is important to me and, in my opinion, understudied in educated circles. At home in Arkansas, my friends and I connect over creating and listening to rap songs. Coming to Yale, I found hip hop, and especially Southern trap music, to be widely taboo and generally scorned as unintellectual. However, I believe that there is much to be learned from the lowbrow. Life is not all about literature, philosophy, and being more learned than your peers; at least for me, the hard part is getting out of my head, not bringing more into it. When I’m back home, my down-to-earth friends help bring my focus back to the simple, important externalities in life, like staying fit, having fun, and being a good son, friend, and brother. People who strongly identify as intellectual sometimes forget that the simplest things are what make life meaningful. They may also seek to set themselves apart from others, but it’s usually healthier to lose the ego and join the crowd.

Via this project I sought to pursue an intellectual curiosity in a balanced fashion. I hoped that some of what I learned from this project could help me better enjoy the music I listen to. I also hoped to maybe share my findings with my friends. This way, even though I would spend hours on a computer to complete the project, I could then see at least a little benefit in my real life from what I’ve learned. I’m a senior Computer Science major, and I’m pretty burnt out from software engineering work, but this project was fun, because it combined some of my hobbies and helped me dive deeper into them. I’m extremely grateful that I could do this project for class instead of a more rigidly structured assignment like a problem set or essay. Creative projects are the best; thank you Professor Glaser!

I want to keep exploring this, since I think there’s potential in the hierarchical classification and hybrid regression-classification models; I just need to tune the hyperparameters and model architecture to make sure my method is correct. But for now, I have other projects to finish, so I’m going to do those, since I’ve spent a lot of time on this. Thanks for reading!