

AAVE:
A COLLECTION
OF AAVE
BY BLACK
SPEAKERS
FOR NLP DATA

Welcome to the AAVE Corpus.

This dataset was created by AAVE speakers for AAVE speakers and the engineers, academics, researchers, and builders that endeavor to create NLP models that represent the beauty and complexity of the AAVE sociolect.

Dataset Formation

The AAVE corpora collection is an open source data collection that uses json and txt files. There are four collections: Lyric, Leadership, Book, and Social Media and each collection can be accessed through this Git repo. This Git repo is protected by MIT License for research purposes only. This repo is only the beginning of the AAVE corpora and is used as a basis for further research for Natural Language Data supplementation for downstream NLP tasks. Download the repo in a Zip file for access. Be aware that this repo is about 400 MB of unprocessed files with 231 items. Further processing each file will increase storage size.

The specific compilation of each collection is different. Below I will describe the characteristics of each.

LYRIC COLLECTION

The Lyric Collection is a combination of json files by artist that contains the lyrics of up to 350 songs from their discography. I focused on compiling the names of the most popular and influential Black artists of the 20th and 21st century. The Lyric collection is skewed male with a heavier emphasis on Hip Hop. This is largely due to the prevalence of male Hip Hop artists in the music industry in the modern day and the relative ease of access to lyrics after the 1950s compared to before. The analytical breakdown of the corpora is below:

There are 21 women and 39 men in this collection.

There is an average of 250 songs per artist bringing the overall song lyric collection to ~15,000 songs.

The oldest song present was recorded in the 1930s.


The newest song in the collection was recorded this year.

Source: Genius




LEADERSHIP COLLECTION

The Leadership collection is a collection of speeches by prolific Black leaders from Fredrick Douglass to Sojourner Truth to Martin Luther King to Ketanji Brown Jackson and many Black political leaders in between. This collection is heavily skewed male and has a heavy education bias as the majority of speakers in this collection have at least a college degree. Geographically the dataset heavily favors speakers from the South and many states, such as Washington state, are not represented as a place of origin for any of the collected speakers. Read below for the analytical breakdown:



34 speeches from Black leaders. Most speeches in the leadership collection are speeches by former President Barack Obama.





BOOK COLLECTION

The Book collection was the most difficult collection to augment. African Americans have been grossly underrepresented from the literary cannon and as a result selecting only the few to have been admitted into the Ivory tower would lend itself to self selection bias. Due to this, I worked to include works from Historically Black Book archive collections from universities. The University of Kansas has such an archive that I heavily powered from. The additions in this sample, though, are not as representative of the community as the other collections and is in need of open source contributions to truly capture the diversity of Black thought and AAVE use.

This was the hardest collection to source due to copyright standards.

There are 54 books in this collection and growing.

Source: University of Kansas



SOCIAL MEDIA COLLECTION

The Social Media Collection is my most robust and diverse collection of AAVE instances for modern day analysis. Included in this collection are json files of video transcripts, tweets, and blog posts from Black thought leaders that use AAVE prominently.

I am a Twitter Developer Research API subscriber and have retrieved tweets using the third party Python packages Tweepy and SNScrape. The Tweet collection information is as followed:

Search 30 Day Tweets*

#blm - June 2022, Count: 2000
#BlackGirlMagic - June 2022, Count: 100
#BlackTwitter - June 2022, Count: 100
#SayHerName - June 2022, Count: 100
LilNasX bet awards - June 2022, Count: 100

Search Full Archive Tweets

#thanksgivingclapback
Date: 11/16/2015 - 11/30/2015
Count: 500
Date: 11/21/2016 - 11/30/2016
Count: 1000
Date: 11/20/2017 - 11/30/2017
Count: 1000

#OscarsSoWhite:
Date: 01/14/2015 - 01/30/2015
Count: 500

#SayHerName
Date: 02/01/2015 - 07/30/2015
Count: 500

#blackmoms
Date: 06/28/2019 - 07/01/2019
Count: 500

#BlackLivesMatter
Date: 05/25/2020 - 08/01/2020
Count: 5000

Tweet Total: 11,400

Source: Twitter

CORPORA

The corpora folder includes the processed .txt files of the AAVE corpora. There are 141341 words in this corpora as of June 9, 2022.

social_media_corpora.txt is the processed .txt file of all of the tweets in this corpora. Here are the details:

- Emojis and hashtags are still in tweets for researchers to either remove or keep to their discretion
- All users and location information has been scrubbed to protect identity
- Tweets collected are not scrubbed of profanity.

leadership_corpora.txt is the processed .txt file of all speeches in this corpora.

literature_corpora.txt is the processed .txt file of all books in this corpora.

lyrics_corpora.txt is the processed .txt file of all lyrics in this corpora.

REFERENCES

- [1] Rickford, John R. 1999. African American Vernacular English: Features, Evolution, and Educational Implications. Malden, MA: Blackwell.
- [2] Alim, H. Samy, and Geneva Smitherman. 2012. Articulate White Black: Barack Obama, Language, and Race in the U.S. Oxford, UK: Oxford University Press.
- [3] Bell, Alan. 1984. Language Style as Audience Design. *Language in Society*, 13 (2): 145–204.
- [4] Britt, Erica, and Tracey L. Weldon. 2015. African American English in the Middle Class. In Sonja Lanehart (ed). *The Oxford Handbook of African American Language*. New York: Oxford University Press. 800-816.
- [5] Chambers, Jack. 2009. *Sociolinguistic Theory, Revised Edition*. Malden, MA: Blackwell.
- [6] Craig, Holly K., and Julie Washington. 2006. Malik goes to school: Examining the language skills of African American students from preschool-5th grade. Mahwah, NJ: Erlbaum.
- [7] Debose, Charles. 1992. Codeswitching: Black English and Standard English in the African-American Linguistic Repertoire. *Journal of Multilingual and Multicultural Development*, 13 (2): 157-167.
- [8] Debose, Charles. 2015. The Systematic Marking of Tense, Modality, and Aspect in African American Language. In Sonja Lanehart (ed). *The Oxford Handbook of African American Language*. New York: Oxford University Press.
- [9] <https://www.languagejones.com/blog-1/2014/6/8/what-is-aave#:~:text=AAVE%20is%20a%20dialect%20of,than%20other%20dialects%20of%20English>.
- [10] <https://time.com/6092078/artificial-intelligence-play/>
- [11] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. <https://doi.org/10.1145/3442188.3445922>

ABOUT **JAZMIA HENRY**

Jazmia Henry is an expert tinkerer that works to combine her love of social justice and ethics with her love of programming and Machine Learning. She graduated from Tulane University in 2016, Columbia University in 2018, and is currently a Technology & Racial Equity Practitioner Fellow with the Stanford Center for Comparative Studies in Race & Ethnicity (CCSRE) and an Affiliate Fellow with the Stanford Institute for Human-Centered Artificial Intelligence (HAI) at Stanford University. Her work has also been supported by the Stanford Digital Civil Society Lab. Currently, she is the Lead Instructor of Machine learning and Data Science with the Data Science for All program at Correlation One and a 2022 Reviewer for the Datasets and Benchmarks track at NeurIPS. Formerly, she has served as the Head of Machine Learning at Motley Fool, advised startups with historically marginalized founders, worked for think tanks, in electoral politics, and financial services as a Data practitioner. In her spare time, she takes care of her puppy, Moji, and renovates historic homes.

ABOUT **CCSRE**

Since its establishment in 1996, the Stanford University Center for Comparative Studies in Race & Ethnicity (CCSRE) has been a leader in producing innovating research, training, policy engagement and public education on diverse topics surrounding race and ethnic studies. In particular, the CCSRE Technology & Racial Equity Initiative works to advance racial justice in the analysis, production, and deployment of new technologies.