An information theoretic characterization of Drake's lyrics

Reese AK Richardson[1]

[1]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208

July 23, 2023

Abstract

Drake scholars have and will continue to take machine learning approaches to Drake's songs and lyrics. However, standards for procedures like stopword handling are ill-defined in the field of natural language processing and entirely unexplored in the field of Drake studies. This work aims to quantitatively characterize the frequency and information content of words in a corpus representing Drake's discography. This analysis yields insights on stopword handling in Drake's lyrics specifically and for song lyrics in general. A dataset quantitatively describing words usage in Drake's lyrics is provided.

1 Introduction

Aubrey Drake Graham, also known as Drake, is a Canadian rapper and musician [1]. Drake's music career spans nearly two decades, beginning with his debut mixtape *Room for Improvement* in 2006. Since then, Drake has released a series of incredibly successful and influential singles, features and albums. The nascent field of Drake studies seeks to achieve an academic understanding of Drake's career and projects, musical and otherwise.

As the field of Drake studies develops, researchers will undoubtedly begin to take a quantitative lens to Drake's songs and lyrics. Some have already done so [2–4]. However, before approaches like document clustering, topic modeling and sentiment analysis can be properly applied to Drake's lyrics, researchers should understand at a basic level the artist's approach to language. For instance, it is unclear what words, if any, should be considered stopwords (i.e. uninformative words that could be removed without affecting a lyric's meaning) when processing Drake's lyrics. Removing stopwords is a common step in natural language processing [5]. However, if the stopwords selected for removal are not appropriate for the body of writing in question, one risks removing informative or important words that change the text's meaning. This work aims to quantitatively characterize the frequency and information content of words in a corpus representing Drake's discography.

2 Methods

Lyrics were obtained for all songs on all of Drake's released solo albums, compilation albums, mixtapes, playlists and extended plays (EPs) on July 17, 2023 using the Python package *LyricsGenius* v3.0.1 [6]. This package returns user-contributed lyrics from the platform Genius. Specifically, projects included in this analysis are *Honestly, Nevermind*, *Certified Lover Boy*, *Scary Hours 2*, *Dark Lane Demo Tapes*, *Care Package*, *The Best in the World Pack*, *Scorpion*, *Scary Hours*, *More Life*, *Views*, *If You're Reading This It's Too Late*, *Nothing Was the Same*, *Take Care*, *Thank Me Later*, *So Far Gone* (EP), *So Far Gone*, *Comeback Season*, *Room for Improvement*, and Drake's 2006 demo disk. This analysis does not include Drake's collaborative projects *What a Time to be Alive* (with Future) or *Her Loss* (with 21 Savage). Some songs were featured as tracks on multiple projects. The corpus was thus composed of 275 unique songs with lyrics (available as a spreadsheet in **Supplementary Data 1**). No distinction was made between lyrics performed by Drake or performed by other artists.

Lyrics were cleaned of punctuation and extraneous textual artefacts with *scikit-learn* v1.3.0 [7] and converted to a count matrix of word occurrences in each document in the corpus. A list of English stopwords was obtained from *NLTK* v3.8.1 [8]. Information content for each word was determined using the approach developed by Gerlach et al. [5]. This approach quantifies the information content of each token (in this case, English words) in a corpus by comparing the Shannon entropy of each token's distribution across documents (in this case, songs) compared to a null distribution of entropy from corpora where all tokens have been shuffled across documents while preserving the marginal counts for each document and token.

Words that are used at roughly the same frequency across songs (e.g. the words "the", "of", "be" and "to") will have a high-entropy distribution near the null model expected value and thus will have low information content. On the other hand, words with a very skewed distribution across documents (e.g. "houstatlantavegas", which occurs 21 times in only a single song) will have a low-entropy distribution far from the null model expected value and thus will have high information content. Usage statistics and information content for all words are available as a spreadsheet in **Supplementary Data 2**.

3 Results

Across 275 songs in 19 albums, mixtapes, playlists and EPs, Drake's lyrics contained 8,932 unique words used 137,585 times. The median song length was 511 words (5$^{\text{th}}$ percentile 167 words, 95$^{\text{th}}$ percentile 836 words). The longest song was 2021's *Lemon Pepper Freestyle* from *Scary Hours 2* at 1,090 words. Word frequency was highly unequal, distributed with a Gini coefficient of 0.873 and such that 5% most frequent words accounted for 77.6% of word occurrences.

Drake's most frequent word was "you", used a total of 5,167 times. **Table 1** shows the 25 most frequent words. "You" and "to" were the most frequent words across songs, present in 273 out of 275 (99.3%). **Table 2** shows the 25 words found in the most songs.

The words with the highest information content included "preach", "sexy" and "fancy" (**Figure 1**). **Table 3** shows the 25 words with the highest information content. Many of these words occur multiple times in a single song's chorus and nowhere else, such as the word "fireworks", featured only in 2010's *Fireworks* from *Thank Me Later*. 2,096 words (23.4%) had negative information content (i.e. higher entropy than the null model value). This includes words that only occur once in every song in which they are present, like "seem" (19 times in 19 songs), "twice" (18 times in 18 songs) and "weed" (16 times in 16 songs). **Table 4** shows the 25 words with the lowest information content.

Some putative English stopwords had relatively high information content, such as "own" and "yours", which both fell in the highest 1% informative words. The 25 English stopwords in Python package *NLTK* with the highest information content are shown in **Table 5**. Several of these stopwords are contractions of "you", reflecting Drake's tendency to directly address an individual in his lyrics. Some stopwords that would be removed by a *NLTK* user would actually mask highly informative words with a different meaning. For instance, the word "won", included in *NLTK* to capture instances where a space is inadvertently inserted into the contraction "won't", has higher information content than 99.8% of words in Drake's corpus.

A visual summary of the information content of certain words is shown in **Figure 1**. A comparison of the information content of certain words to term frequency inverse document frequency, another measure of the importance of tokens in a corpus [5], is shown in **Figure 2**.

4 Discussion

This work makes several novel contributions to the field of Drake studies.

First, this analysis reveals interesting idiosyncrasies in Drake's lyrics. For instance, the most common word in the constructed corpus of Drake's lyrics is "you", capturing the artist's tendency to directly address an individual in his songs. Several contractions of "you" and the genitive case "yours" also appear frequently and are among the most informative words in the corpus.

Second, this analysis provides a dataset that can inform future quantitative approaches to Drake's lyrics. Machine learning approaches are prone to so-called "shortcut learning", by which models make decisions based on uninformative or irrelevant features, leading to failures in model generalization [9]. This dataset could be used to mitigate shortcut learning in the machine interpretation of

Drake's lyrics by describing which words in these lyrics actually encode information relevant to the task of characterizing and distinguishing songs.

Third, this analysis raises questions about stopword handling in the machine interpretation of Drake's lyrics specifically and in song lyrics in general. Many English language stopwords were actually highly informative Drake's lyrics. This suggests that stopword lists cannot easily be generalized and may need to be tuned to individual artists for certain tasks. If one used Python package *NLTK* to process Drake's lyrics with stopword removal, they would remove "you"-related stopwords, erasing crucial information on Drake's relationship to the listener. On the other hand, song lyrics, Drake's included, tend to feature repetitive sections, especially in hooks and choruses. This could inflate the information content and perceived importance of particular words: consider the word "own", which occurs 73 times in 2013's *Own It* from *Nothing Was the Same*, mostly in the chorus. Whether words like this should be considered for certain tasks is subject to the interpretation of a chorus' importance over the rest of the lyrics. Altogether, these questions suggest that stopword handling may require different approaches for musical and non-musical texts.

This work is also subject to several limitations. First, it does not include Drake's collaborative projects *What a Time to be Alive* or *Her Loss*. Second, it includes all lyrics in a song, including those not sung or authored by Drake himself. Third, the lyrics were obtained from the platform Genius, where lyrics are user-contributed. Some words having the same meaning may have been transcribed differently by users, such as "tryin" versus "trying". Future works may benefit from seeking out a more authoritative source for lyrics.

5 Figures and tables

Table 1: 25 most frequent words in corpus.

Token	I (bits)	I (percentile)	TFIDF	Count	Songs	Stopword
you	0.193	75.358	0.138	5167	273	TRUE
the	0.196	75.392	0.587	4694	266	TRUE
to	0.100	74.048	0.079	2946	273	TRUE
and	0.181	74.922	0.393	2812	265	TRUE
me	0.261	76.556	0.577	2494	259	TRUE
it	0.323	77.967	0.344	2460	265	TRUE
im	0.367	78.627	0.662	2118	254	FALSE
my	0.227	75.918	0.468	2023	259	TRUE
that	0.205	75.459	0.466	1886	258	TRUE
in	0.185	74.978	0.566	1633	252	TRUE
on	0.237	76.030	0.411	1562	257	TRUE
know	0.433	80.307	0.624	1378	246	FALSE
like	0.371	78.639	0.990	1311	231	FALSE
yeah	0.625	84.136	1.311	1293	220	FALSE
for	0.397	79.825	0.624	1225	243	TRUE
they	0.563	83.251	1.655	1155	205	TRUE
with	0.257	76.243	0.635	1081	239	TRUE
dont	0.383	79.109	0.550	1041	242	TRUE
just	0.247	76.142	0.499	1018	244	TRUE
all	0.335	78.168	0.829	1008	228	TRUE
up	0.411	80.038	0.838	992	227	TRUE
of	0.380	78.829	0.806	981	228	TRUE
its	0.529	82.859	0.827	953	226	TRUE
we	0.457	80.631	1.044	912	215	TRUE
but	0.172	74.832	0.605	905	235	TRUE

Table 2: 25 words present in the most songs (total 275) in corpus.

Token	I (bits)	I (percentile)	TFIDF	Count	Songs	Stopword
to	0.100	74.048	0.079	2946	273	TRUE
you	0.193	75.358	0.138	5167	273	TRUE
the	0.196	75.392	0.587	4694	266	TRUE
and	0.181	74.922	0.393	2812	265	TRUE
it	0.323	77.967	0.344	2460	265	TRUE
me	0.261	76.556	0.577	2494	259	TRUE
my	0.227	75.918	0.468	2023	259	TRUE
that	0.205	75.459	0.466	1886	258	TRUE
on	0.237	76.030	0.411	1562	257	TRUE
im	0.367	78.627	0.662	2118	254	FALSE
in	0.185	74.978	0.566	1633	252	TRUE
know	0.433	80.307	0.624	1378	246	FALSE
just	0.247	76.142	0.499	1018	244	TRUE
for	0.397	79.825	0.624	1225	243	TRUE
dont	0.383	79.109	0.55	1041	242	TRUE
with	0.257	76.243	0.635	1081	239	TRUE
but	0.172	74.832	0.605	905	235	TRUE
like	0.371	78.639	0.99	1311	231	FALSE
all	0.335	78.168	0.829	1008	228	TRUE
of	0.380	78.829	0.806	981	228	TRUE
up	0.411	80.038	0.838	992	227	TRUE
its	0.529	82.859	0.827	953	226	TRUE
got	0.321	77.944	0.837	890	223	FALSE
yeah	0.625	84.136	1.311	1293	220	FALSE
when	0.270	76.903	0.709	699	220	TRUE

Table 3: 25 words with the highest information content.

Token	**I (bits)**	**I (percentile)**	**TFIDF**	**Count**	**Songs**	**Stopword**
preach	5.822	100	182.174	74	2	FALSE
sexy	5.146	99.989	57.111	54	4	FALSE
fancy	4.792	99.978	58.736	39	3	FALSE
ay	4.779	99.966	45.478	43	4	FALSE
aye	4.673	99.955	58.169	55	4	FALSE
dedicate	4.618	99.944	52.712	35	3	FALSE
fireworks	4.492	99.933	134.803	24	1	FALSE
woop	4.489	99.922	134.803	24	1	FALSE
gangstas	4.486	99.910	134.803	24	1	FALSE
tri	4.485	99.899	134.803	24	1	FALSE
hannenin	4.476	99.888	68.931	28	2	FALSE
tryin	4.455	99.877	11.595	101	22	FALSE
houstatlantavegas	4.311	99.866	117.952	21	1	FALSE
won	4.310	99.854	117.952	21	1	TRUE
faded	4.249	99.843	27.856	63	8	FALSE
falling	4.175	99.832	56.622	23	2	FALSE
wishin	4.098	99.821	30.671	29	4	FALSE
brea	4.022	99.810	95.485	17	1	FALSE
controlla	4.022	99.798	95.485	17	1	FALSE
brand	3.777	99.787	14.631	50	11	FALSE
nails	3.752	99.776	25.383	24	4	FALSE
account	3.751	99.765	25.383	24	4	FALSE
leaving	3.697	99.754	41.851	17	2	FALSE
dollar	3.677	99.742	18.488	29	6	FALSE
grammy	3.666	99.731	24.325	23	4	FALSE

Table 4: 25 words with the lowest information content.

Token	**I (bits)**	**I (percentile)**	**TFIDF**	**Count**	**Songs**	**Stopword**
seem	-0.073	0.011	2.672	19	19	FALSE
twice	-0.068	0.022	2.726	18	18	FALSE
rapper	-0.064	0.034	2.784	17	17	FALSE
weed	-0.060	0.045	2.844	16	16	FALSE
future	-0.059	0.056	2.909	15	15	FALSE
each	-0.052	0.067	3.052	13	13	TRUE
bread	-0.048	0.078	3.132	12	12	FALSE
memphis	-0.048	0.090	3.052	13	13	FALSE
uncle	-0.047	0.101	3.132	12	12	FALSE
sitting	-0.047	0.112	3.132	12	12	FALSE
dinner	-0.046	0.123	3.132	12	12	FALSE
finish	-0.046	0.134	3.219	11	11	FALSE
album	-0.045	0.146	3.132	12	12	FALSE
pretty	-0.044	0.157	3.219	11	11	FALSE
thatll	-0.044	0.168	3.132	12	12	TRUE
plane	-0.043	0.179	3.219	11	11	FALSE
message	-0.042	0.190	3.314	10	10	FALSE
mr	-0.042	0.202	3.219	11	11	FALSE
hours	-0.042	0.213	3.314	10	10	FALSE
throwin	-0.041	0.224	3.219	11	11	FALSE
purple	-0.041	0.235	3.219	11	11	FALSE
fashion	-0.041	0.246	3.314	10	10	FALSE
prove	-0.040	0.258	3.314	10	10	FALSE
ignore	-0.040	0.269	3.314	10	10	FALSE
thoughts	-0.040	0.280	3.314	10	10	FALSE

Table 5: The 25 English stopwords in Python package *NLTK* with the highest information content.

Token	**I (bits)**	**I (percentile)**	**TFIDF**	**Count**	**Songs**	**Stopword**
won	4.310	99.854	117.952	21	1	TRUE
own	3.282	99.541	5.480	127	43	TRUE
yours	3.046	99.295	11.91	67	16	TRUE
youd	2.191	98.231	8.874	34	12	TRUE
youll	1.759	96.977	6.815	52	20	TRUE
youve	1.703	96.865	6.043	63	25	TRUE
doing	1.384	94.996	4.324	53	28	TRUE
too	1.310	94.660	2.227	415	134	TRUE
were	1.290	94.492	3.222	127	60	TRUE
again	1.215	94.223	3.853	113	50	TRUE
down	1.185	94.122	1.802	454	151	TRUE
over	1.177	93.831	2.631	212	90	TRUE
after	1.143	93.316	3.419	85	45	TRUE
did	1.136	93.271	2.460	182	86	TRUE
shes	1.060	92.801	4.700	49	25	TRUE
she	1.048	92.734	2.729	482	131	TRUE
about	1.029	92.600	1.774	319	132	TRUE
no	0.974	89.834	1.725	693	177	TRUE
where	0.963	89.722	1.725	294	129	TRUE
am	0.95	89.633	3.151	142	65	TRUE
through	0.948	89.610	2.355	202	93	TRUE
more	0.943	89.588	1.868	251	116	TRUE
under	0.936	89.420	4.042	33	21	TRUE
why	0.922	89.084	2.123	250	109	TRUE
he	0.882	88.849	2.457	198	90	TRUE

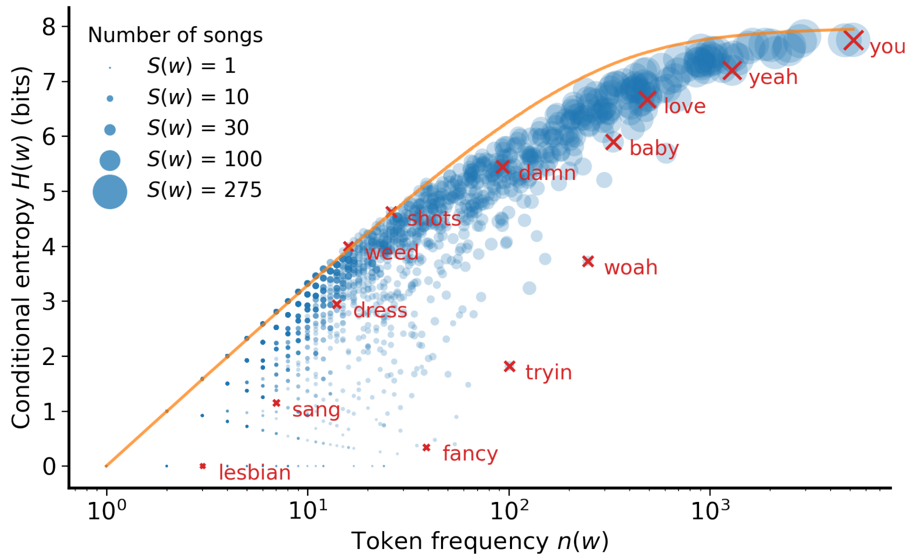

Figure 1: The conditional entropy $H(w)$ for each word w (blue dots) as a function of word frequency $n(w)$, compared to the null model expected value (orange line). Dots are scaled by the number of songs that contain a word. Information content is the difference between the null model expected value and $H(w)$.

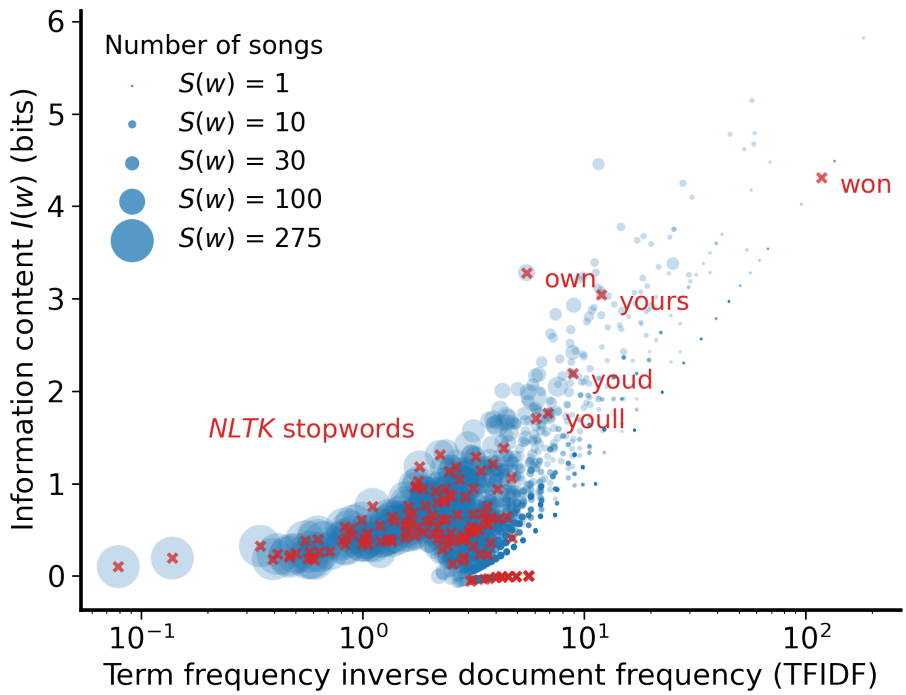

Figure 2: The information content $I(w)$ of each word w as a function of TFIDF of each word w. All words are shown as blue dots, scaled by the number of songs that contain that word. Stopwords from Python package *NLTK* are highlighted with red crosses.

6 Data and code availability

All data and code is available at `github.com/reeserich/drake_information`.

References

[1] *Drake. Encyclopaedia Britannica*. Accessed July 18, 2023 [Online]. URL: `https://www.britannica.com/biography/Drake`.

[2] Brandon Punturo. *Drake — Using Natural Language Processing to understand his lyrics. Towards Data Science*. Accessed July 18, 2023 [Online]. Aug. 2018. URL: `https://towardsdatascience.com/drake-using-natural-language-processing-to-understand-his-lyrics-49e54ace3662`.

[3] Peter Li. *The Many Clusters of Drake. Towards Data Science*. Accessed July 18, 2023 [Online]. May 2019. URL: `https://towardsdatascience.com/the-many-clusters-of-drake-8718607401ad`.

[4] Hamza Kazmi. *Old vs New Drake Lyrics With AI. Towards Data Science.* Accessed July 18, 2023 [Online]. Aug. 2020. URL: https://towardsdatascience.com/understanding-differences-between-drakes-old-vs-new-songs-using-text-classification-and-lstm-6381fb568b31.

[5] Martin Gerlach, Hanyu Shi, and Luıs A Nunes Amaral. "A universal information theoretic approach to the identification of stopwords". In: *Nature Machine Intelligence* 1.12 (2019), pp. 606–612.

[6] John W Miller. *LyricsGenius*. Version 3.0.1. Apr. 18, 2021. URL: https://pypi.org/project/lyricsgenius/.

[7] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[8] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[9] Robert Geirhos et al. "Shortcut learning in deep neural networks". In: *Nature Machine Intelligence* 2.11 (2020), pp. 665–673.