<table>
<thead>
<tr>
<th>No</th>
<th>Authors</th>
<th>Reference</th>
<th>Year</th>
<th>Topic</th>
<th>Genre of Songs</th>
<th>Data</th>
<th>Size</th>
<th>Analysis</th>
<th>Special Visualizations</th>
<th>Key Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Daniels, M.</td>
<td><a href="http://polygraph.co/shorty.ht">http://polygraph.co/shorty.ht</a> ml</td>
<td>2015</td>
<td>Analysis of the use of the word Shorty in rap lyrics as well as other associated words over time</td>
<td>Rap/HH</td>
<td>Lyrics from RapGenius.com and album release data from Musiablox</td>
<td>NA</td>
<td>Counts of the usage of the word Shorty by year and by meaning</td>
<td>Stacked area charts for various meanings of Shorty by year as well as bar charts of other word usage</td>
<td>Single Words shorty</td>
</tr>
<tr>
<td>2</td>
<td>Daniels, M.</td>
<td><a href="http://polygraph.co/timeless/">http://polygraph.co/timeless/</a></td>
<td>2015</td>
<td>Analysis of play counts of various tracks and the associated artist by year of release</td>
<td>Variety</td>
<td>Spotify data cover full range of years from 1950 until 2015</td>
<td>NA</td>
<td>Frequency analysis of Spotify play counts by various artists and tracks by year</td>
<td>Interactive rank charts, bar charts and time lines</td>
<td>Popularity of tracks via Spotify</td>
</tr>
<tr>
<td>3</td>
<td>Daniels, M. and Beacham, K.</td>
<td><a href="http://polygraph.co/labels/">http://polygraph.co/labels/</a></td>
<td>2015</td>
<td>Ranking and Analysis of Success of Record Labels</td>
<td>Rap/HH</td>
<td>No of weeks that their artists tracks were on Billboard's Hot Rap Artists 1989 to 2014</td>
<td>16 years by 52 weeks of top 25 on bill board</td>
<td>Time and geographically based statistical analysis of the performance of the top rap record tables</td>
<td>Interactive Table and Sprite Charts by Label, Artist and Year.</td>
<td>No weeks on Bill Board Hot Rap</td>
</tr>
<tr>
<td>4</td>
<td>Malmi, E.</td>
<td><a href="http://mining4rea">http://mining4rea</a> ning.com/2015/01/ 731/raplyzer</td>
<td>2015</td>
<td>Presents Raplyzer, a computer program which automatically detects rhymes from rap lyrics and which is used to rank popular rappers based on their average Rhyme factor. Presents another program called BattleBot, which is a search engine for rhyming rap lines based on the algorithm used in Raplyzer</td>
<td>Rap/HH</td>
<td>Scraped the lyrics of 94 rap artists from a lyrics website.</td>
<td>10,082 songs.</td>
<td>For each artist, I computed the Rhyme factor averaged over all the songs of the artist (based on assonance rhyming). Rank ordered and then compared rhyme factor score by number of unique words (ala Daniels)</td>
<td>Scattergram of rhyme factor by unique words for each artist.</td>
<td>Rhyme factor for assonance</td>
</tr>
<tr>
<td>5</td>
<td>Malmi, E.</td>
<td><a href="http://arxiv.org/abs/1505.04771v1">http://arxiv.org/abs/1505.04771v1</a></td>
<td>2015</td>
<td>The highly structured nature of rap makes it particularly amenable to computer analysis. And that raises an interesting question: if computers can analyze rap lyrics, can they also generate them?</td>
<td>Rap/H</td>
<td>A list of 104 popular English-speaking rap artists and scraped all their songs available on a popular lyrics website.</td>
<td>In total, we have 583,669 lines from 10,980 songs</td>
<td>Our approach is based on two machine learning techniques: the RankSVM algorithm, and a deep neural network model with a novel structure. For the problem of distinguishing the real next line from a randomly selected one, we achieve an 82% accuracy. We employ the resulting prediction method for creating new rap lyrics by combining lines from existing songs. In terms of quantitative rhyme density, the produced lyrics outperform best human rappers by 25%.</td>
<td>None</td>
<td>Rhyme factor for assonance</td>
</tr>
<tr>
<td>6</td>
<td>Mauch, M. et al.</td>
<td><a href="http://pirsa.royaloppublishing.org/content/797/1980-">http://pirsa.royaloppublishing.org/content/797/1980-</a> 83</td>
<td>2015</td>
<td>Using music information retrieval and text mining tools, analyze the musical properties of approximately 17000 recordings that appeared in the charts and demonstrate quantitative trends in their harmonic and timbral properties. Use these properties to produce an audio-based classification of musical styles and study the evolution of musical diversity and disparity, testing, and rejecting, several classical theories of cultural change. Investigate whether pop musical evolution has been gradual or punctuated. Show that, although pop music has evolved continuously, it did so with particular rapidity during three stylistic ‘revolutions’ around 1964, 1983 and 1991. Conclude by discussing how our study points the way to a quantitative science of cultural change.</td>
<td>Variety including Rap/HH</td>
<td>Focused on songs that appeared in the US Billboard Hot 100 between 1960 and 2010. We obtained 30-s-long segments of 17094 songs covering 86% of the Hot 100. The earlier years had more missing songs than the later years.</td>
<td>30-s audio of 17094 songs</td>
<td>LDA was used to determine the underlying harmonic and timbral topics - 8 in all. Each track was represented as a distribution over eight harmonic topics that capture classes of chord changes (e.g. ‘dominant-seventh chord changes’) and eight timbral topics (T-topics) that capture particular timbres (e.g. ‘drums, aggressive, percussive’, ‘female voice, melodic, vocal’, derived from the expert annotations), with topic proportions q. These topic frequencies were the basis of the analyses. Topics then used to constructed a taxonomy of 13 styles by k-means clustering on principal components derived from our topic frequencies. This was the base of the analysis of the rise and fall of the styles as well as the changing audio features.</td>
<td>Single Chart showing time series graphs for each of the 8 harmonic and 8 timbral audio feature topics. Specialized comba chart showing dendrogram of style/genres and a spindle graph of those styles by years. Combined Vertical Time Series/Boxplot of 4 Diversity Measure by Time. Quarterly pairwise distance matrix of all the songs in the Hot 100 based on 16 audio topics. Graph of Footestimate novelty for each year.</td>
<td>Audio features of songs - harmonic and timbre</td>
</tr>
<tr>
<td>7</td>
<td>Powell-Morse, A.</td>
<td>Lyric Intelligence in</td>
<td>2015</td>
<td>Using song lyrics, tries to answer questions about which genres</td>
<td>Variety including</td>
<td>Analyzed 225 tracks from</td>
<td>255 tracks</td>
<td>Analysis included overall scores by year, artist sex and year, genre and year, and avg. word count by Time line charts and bar charts</td>
<td>Lyric Intelligence</td>
<td>Hip-Hop Page 1</td>
</tr>
</tbody>
</table>
Billboard charts for Big Data, MC Examines social relationships
Hemphill, T. Examines 150 of the top songs
Hemphill, T. F1
Inspired by Lyrics taken from Sexual relations
Raschka, D. Billboard ranking
NA Machine Building a classification model
Variety Downloaded the
Hand-labeled a
Focused on optimizing precision and recall via
Stacked bar charts,
SAD vs. happy

8 Ruiz, E. Why Cam'ron's Hey Ma is Peak Neotraditional
2015 Examines 150 of the top songs from 2001-2005 and called them down to sixteen tracks looking at their current
RapHH Billboard ranking and Spotify ranking 16 tracks Relationship btw. Billboard percentile at year of release and Spotify ranking in 2014 Parallel coordinate chart displayed horizontal Popularity of tracks via Spotify

9 Thompson, J. The Evolution of Rap lyrics and a Tale of Two LDA's
2015 Inspired by The Evolution of Popular Music: USA 1960-2010 (R above) which utilizes audio signaling processes to understand the evolution, this paper explores investigates the possibility of tracing the same evolution through lyrics.
Variety including RapHH Used artists and songs from Evolution of Music study (songs that appeared in the US Billboard Hot 100 between 1960 and 2010.) Lyrics obtained by Chartlyrics API ~80% of 17094 songs (~13672 songs)
Used Latent Dirichlet Allocation (LDA) to determine ordering topics. After exploration settled on 25. Unfortunately these did not do well in predicting the 13 genres assigned in the original evolution study (only 16% predicted but rap once one of the better ones-60%).
Word Clouds, Topic Charts, Heat Map (all created with D3.js).
Analytically derived Topics of popular music

10 Big Data, MC Riding Dirty: The Science of Cars and Rap lyrics
2014 Which car makes appeared most frequently in top lyrics from 1996-2014
RapHH RapGenius.com lyrics (no mention of actual sample) Measured the frequency of various makes across the years.
Frequency of mention in lyrics by car make and across time Time series line charts Car Brands

2014 Textual analysis of hip-hop songs—both lyrics and music videos. Using standpoint theory, the study examines hip-hop songs and music videos' gender dynamics, the power structure between men and women, and popular themes.
RapHH A list of 25 hip-hop music videos was compiled, based on the top downloaded hip-hop music videos on iTunes as of September 2014. 25 of top downloaded hip hop music videos from iTunes in 2014 A codebook was created whose purpose was to identify coding categories for analyzing the lyrics and music. For lyrics, these were the subjects analysed: the occurrence of derogatory words, references alluding to money and extravgant living, references to having or getting any number of women and descriptions of their appearances, mention of drug and alcohol use on the performers themselves, the use of drugs and alcohol on women, the mention of partying, descriptions of sexual activities, shows of masculinity through money, sexual prowess or acts of violence. For music videos, these were the subjects analysed: the gender of the hip-hop artists', hip-hop artists' style of dress, the presence of dancers, dancers' style of dress (if applicable), dancers' gender(s), the style of dance choreography, the use of a 'video vision', the overall theme of the music video. Frequency analysis was done on list for the various subjects.

12 Daniels, M. Outcasts in Charts: Look at Outkast's History Through the Lens of Data
2014 Analysis of Outkasts' use of various words and themes RapHH Data based on lyrics and themes in various tracks over the years Specific Word and theme usage from 1985 to 2012 Time line and bar charts used to analyse use of specific words and themes Stacked area charts and bar charts Specific words - skat, crank, play, guns, drugs, the south, pimpin, cars/taxis

13 Daniels, M. The Largest Vocabulary in Hip Hop: Rappers, Ranked by the Number of Unique Words Used in their Lyric
2014 Analysis of lyrics by number of unique words in track RapHH Lyrics taken from Genius.com 35K words from first albums of 85 prominent Rappers Rank ordering of Rappers (vocabulary); special, Interactive Ranking Ordering of Rappers by unique word counts by region
Lexical diversity

14 Hemphill, T. Naic: Maximum Distance. Minimum Displacement.
2014 Used the Hip Hop Word Count's new semantic analysis results to extract all geographic mentions from the complete bodies of work of 32 rappers. These locations were translated into geo coordinates which were then made into points that plotted the location's movements. The location moved each week while holding a light pen.
RapHH Tahir Hemphill's Hip Hop Word Count database Body of works of 12 artists from 50,000 rap songs from 3,000 artists, from 1979 to the present
Semantic analysis and extraction of geographical mentions. Locations translated into geo coordinates. Light pen display of moving from point to point in the geographical mentions. Geographic mentions

15 Hemphill, T. Picasso Baby!
2014 Examines social relationships between rap artists and fine artists using mentions of the fine artists (e.g. Picasso) in the lyrics of the rap artist (say Jay-Z)
RapHH Hip Hop Database including lyrics from 500 tracks and 3000 artists 50,000 rap songs from 3,000 artists, from 1979 to the present
Creates bi-modal graph: artist creates track; artist name appears in track lyrics Interactive SNA Named (fine art) artists

16 Musuman, S., Moore, J. and Coventry, B. Using Machine Learning to Understand Song Popularity: Can Lyric Content Predict Song Trajectories on the Billboard 200?
2014 Use song lyrics to predict the ranking of a song on next week's Billboard's Hot 100. Does this by predicting song popularity trajectory labels and then utilizing the trajectory labels to predict song rankings.
Variety MusicXtrace (MMX) bag-of-words version of the Million Songs Dataset (MSD) from 1999-2008. The Hot 100 songs by week were scraped from BB's Hit 100 site -- this is the set to predicted.
Given the year and the songs for which lyrics are available, the dataset was reduced to 940 songs.
The analysis was fairly complicated. Essentially, they gathered a series of NLP measures for each track - top 3 songs, average word length, overall memorability of songs based on novelty determined by tf-idf, song sentiment (polarity, pos-neg, subjectivity - along with a series of attributes from the MSD database including artists familiarity scores, artist hotness, year released, energy and tempo of song, and loudness of the song. These features were used to predict various trajectory labels (e.g. songs in the top 10,20,50) for a week or average change in rankings > [5,10]. These trajectory labels were then used to predict the next weeks ranking.
Standard bar charts, line charts, and scatterplots Word frequencies, entropy, and sentiment.

17 Raschka, D. Building a classification model
2014 Variety Downloaded the
Hand-labeled a
Focused on optimizing precision and recall via
Stacked bar charts

Hip-Hop Page 2
Learning Approach to Classify Music by Mood Based on Song Lyrics

That can automatically predict the mood of music based on song lyrics, i.e., classifying them as happy or sad.

Lyrics from LyricWiki, filtered out songs for which lyrics were not available, and automatically removed non-English songs using Python's Natural Language Toolkit.

Subset of 1200 songs: 1000 songs for the training dataset and 200 songs for the validation dataset.

The F1-score performance metric rather than optimizing the overall accuracy – i was primarily interested in filtering out sad songs, the differences between the different pre-processing steps and parameter choices were rather minor except for the choice of the n-gram sequence lengths.

Chart of number and one for percentage vs. sad for 10 periods since 1960-2030.

emotion

18

Saiaiki, S. et al.

https://staff.aist.go.jp/m.goto/PAPER/2013/48/Saiakif.pdf

LyricRadar: A Lyrics Retrieval System Based on Latent Topics of Lyrics

2014

Presents a novel system - LyricRadar - that enables users to interactively browse song lyrics by visualizing their "topics" rather than relying on simple word search.

Variety

Utilizes authors’ database of popular Japanese [J-Pop] songs and popular English songs taken from the Music lyrics Database.

6902 J-Pop (1847 artists, 2285 song writers, and 26k words) and 3351 English songs (487 artists and 36k words).

Employed Latent Dirichlet Allocation (LDA) to derive 5 topics and their weights for songs and artists. The weights were derived using key noun, verb, and adjective components from the lyrics (derived using “MeCab”) after the elimination of stopwords and the weighting of the words via “tf/idf.”

Lyrics Radar Chart to both display and interactively search for songs and artists; 20 cluster display of song and artists. Both are based on the topic weight and not single topic placement.

Topic profile for 5 derived topics

19

Ying, T., Doraism, S. and Abdullan, N.

http://scientiar.net/fulltext/7-1003-2891-RUG01-002

Lyrics-based Genre Classification Using Variant H-ID Weighting Schemes

2014

Study an approach to lyrics-based musical genre classification was presented which utilizing mood information. From the analysis of the lyrics text in the data collection, correlation of terms between genre and mood was observed. Based on this correlation, a new weighted equation with combine weights from genre and mood was introduced and implemented in two different ways.

Variety but not R&HH

Musical genre classification experiments were performed using a test collection consists of 3000 English songs. A total number of 10 genres were used, including pop, blue, country, folk, R&B, reggae, grunge, punk rock, soul and metal. In test collection were selected.

1000 English songs classified by genre. Mood of text set coded by 5 undergrads into happy, sad, angry, calm, gloomy, romantic, confused, disgusted and aggressive.

Combined all documents of single mood (10) class and then computed specialized weighted tf-idf for 10 docs and looked at correlations

None

Gene based on mood

20

Cundiff, G.


The Influence of Rap and Hip-Hop Music on Audience Perceptions of Misogynistic Lyrics

2013

Analyzed the misogynistic lyrical content of 10 popular rap and hip-hop songs found BB Hit 100 from 2000-2010, as well as their impact on audience views of violence against women.

Rap/HH

Selected Billboard Hit 100 tracks on which to perform qualitative content analysis

20 tracks

Content analysis determining the frequency of mentions of various types of misogyny (demeaning language, rape/sexual assault, sexual conquest and/or physical violence)

Pie charts of % of types of misogynistic lyrics (out all mentions) and % by “level of mentions” - low (1-2), med (3-4), high (5+)

Misogynistic lyrics by type

21

Higa, S.


Analyzing Rap Lyrics with Python

2013

The most beloved car brand in hip-hop?

Rap/HH

Created list of car brands by manufacturer and country using manufacturer’s list. Originally 10 countries and 2599 brands reduced to 4 countries – Germany, Japan, UK and US – and 153 brands.

Radio 1000

Black Youth Group Rap Lyrics Database contains lyrics for all of Billboard Music’s rap songs from 1989 through 2009.

Frequency counts by brand and year.

Note: provides Python code.

None

Google Charts

Car brands

22

Abraham, T., Koul, N. and Morales, J.

http://people.ischool.berkeley.edu/~michaela/rapsongs/index.html

R.A.P. - Rap Analysis Project

2013

Applied machine learning techniques and data science principles to a database of rap lyrics from 1980 to 2015. After an active exploration of the data, focused on “hit prediction”, particularly on what it takes to make it onto the weekly Billboard Top 100 charts.

Rap/HH

Genius.com database of rap lyrics

24,175 entries. This dataset contained 1,491 rap songs that had successfully made it onto the top 100

General analysis of % of Billboard 100 rap songs by month and year. Investigated lyrics for brand names, geographical locations, crime & geography, and theme. Trained an support vector machine (SVM) on the features of song topic, vocabulary, and release data. Classified dataset of 600 songs (300 successful and 300 unsuccessful) achieving over 71% accuracy for classifying the success of songs. Utilized Acc.kNeighAPI for entity identification and search for songs.

CartoDB maps, brand/topic by country, bar chart, pie chart, and “Tableau” for charting.

Cabo/DB maps, brand/topic by country, bar chart, pie chart, and “Tableau” for charting.

23

Cheney, C.


Hip-Hop Word Count is a Living, Breathing Database of Every Word in Every Rap Song Ever

2013

Discusses the analysis mentions of brand products in Tahir Hemphill’s Hip-Hop Word Count Database which includes more than 50,000 rap songs from 3,000 artists, from 1979 to the present. T

Rap/HH

Tahir Hemphill’s Hip Hop Word Count database

50,000 rap songs from 3,000 artists, from 1979 to the present

Focus on brand mentions of Champagne by brand

775 Mentions

Ace Of Spades: 9
Asti: 8
Chameno: 5
Cold Duck: 2
Cristal: 297
Dom Perignon: 209
Meet: 249

Streamgraphs champagne brands over time; also US Map of mentions by artist/track location

Champagne brands

24

Stercke, L.


Topic Detection in a Million Songs

2013

Music Information Retrieval (MIR) is the international research is the so-called “Million Song Datasets” (MSD) with metadata for 1,000,000 songs. This metadata is matched with 237,662 lyrical tags from commercial lyrics catalogues.

Variety including Rap/HH

The main dataset used for this research is the so-called “Million Song Datasets” (MSD) with metadata for 1,000,000 songs. This metadata is matched with 237,662 lyrical tags from commercial lyrics catalogues.

Labeled Latent Dirichlet Allocation, a super-vised topic model, was trained using a labeled subset of stopwords and the weighting of the words via “tf/idf.”

Point graph and bar graphs displaying kurtosis of various bars.

Top 20 topics

Hit-Hop Page 3
<table>
<thead>
<tr>
<th>Reference</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trykowski, T.</td>
<td>Rap Game Riff Raff Textual Analysis. 2013 Undertook a textual analysis of Riff Raff's lyrical cannon. We cleaned up every Riff lyric. We then threw the gigantic word document into a few fun experiments. RapHH Ragenius.com lyrics of Riff Raff songs. 116 songs, 20000 words. Diagrammatic analysis of selected words - ice, candy, versace, ball, played for, rap game. Word cloud and word trees for selected words. Selected words - ice, candy, versace, ball, played for, rap game.</td>
</tr>
<tr>
<td>Zichermann, S.</td>
<td>The Effects of Hip-Hop and Rap on Young Women in Academia. 2013 Investigates the rise of the cultures and music of hip-hop and rap in the world and its effects on its female listeners and fans, especially those in academia. Conducted a content analysis of 95 lyrics from the book, Hip-Hop &amp; Rap: Complete Lyrics for 175 Songs by Spence. The songs had lyrics that repeated misogynist and sexist messages. Also conducted focus group on female students to see impact of lyrics. RapHH lyrics from the book, Hip-Hop &amp; Rap: Complete Lyrics for 175 Songs by Spence. Complete Lyrics for 175 Songs. Content analysis looking for 5 potential themes: Sexual exploitation and disrespect towards women, Overtreatment display of wealth, Glamorization of tobacco, alcohol, illegal substances, and weapons, Territory establishment, inclusion of derogatory and racially charged terminology. None Misogynist and sexist messages.</td>
</tr>
<tr>
<td>Davis, K.</td>
<td>Get Rich or Die Trying: A Semiotic Approach to the Construct of Wealth in Rap Music. 2011 Signification and Representation of Wealth Themes - sex, living the good life, and respect - and signifiers (money, cars, attire, liquor and bling) as displayed in text of rap lyrics. RapHH Top Billboard songs of last decade. 11 songs. &quot;Homemeutic&quot; content analysis of songs to identify significations and representations with focus on major genres (e.g. Gangsta Rap) None Wealth - Materialism themes</td>
</tr>
<tr>
<td>Mody, M.</td>
<td>A Rhetorical Analysis of the Meaning of the &quot;Independent Woman&quot; in the Lyrics and Videos of Male and Female Rappers. 2011 Combines feminist and critical cultural theories to explore the meanings of the &quot;independent woman&quot; in the lyrics and respective videos of male and female rappers. RapHH Analysis of the rap songs/videos yielded from search of songs in 2010 containing the keywords &quot;women&quot; and &quot;independence&quot; Songs/videos of 6 rappers. A rhetorical analysis of collected songs, videos and video comments to compare and contrast perceptions of independence by male and female rappers and audiences. None Misogynistic language vs images of independence with material wealth for women</td>
</tr>
<tr>
<td>Oware, M.</td>
<td>Decent Daddy, Imperfect Daddy: Black Male Rap Artists’ Views of Fatherhood and the Family. 2011 Exploratory research examines how black male rappers talk about motherhood, fatherhood, and parenthood. RapHH Sample includes platinum selling “rap” albums (selling at least one million copies) from 2004 to 2009. I employed the Recording Industry Association of America (RIAA) to help identify platinum selling rap albums. I distinguish “rap” albums from “hip hop” albums. 391 of rap songs from the years 2004 to 2009. Contents analysis to understand how they talk about motherhood, fatherhood, and parenthood. None sexism, misogyny, violence, and hypermasculinity: attitude toward marriage and family; love, protection, and material support toward their mothers and children; relationships with their biological fathers; attitudes towards (biological mothers of their children—baby mamas)</td>
</tr>
<tr>
<td>Primack, B. et al.</td>
<td>Content Analysis of Tobacco, Alcohol, and Other Drugs in Popular Music. 2010 Analysed the 279 most popular songs of 2005 according to Billboard magazine. Two coders working independently used a standardized data collection instrument to code portrayals of substance use. Outcome Measures—Presence and explicit use of substances and motivations for, associations with, and consequences of substance use. Varieties including R&amp;B/HB and Rap Most popular songs from 2005 Billboard magazine. Lyrics from 279 songs. Content analysis of songs resulting in frequency analysis of references to substance abuse, motivations for use, as well as consequences. Detailed coding for these dependent variables are in the paper. The type of substance was noted along with coded descriptive information related to each song from Billboard’s records, including song title, artist, album, song length in minutes and seconds, sex of singer, and primary song genre. Genre included 5 —country, pop, R&amp;B/hip-hop, rap, and rock). Chi square analysis used to determine differences between types, motivations and consequences. None Substance abuse</td>
</tr>
<tr>
<td>van Zaanten, M. and Kanters, P.</td>
<td>Using a collection of lyrics and corresponding user-tagged Automatic Mood. 2010 Variety From 10K songs in Moodo database, lyrics for 5631 songs. All docs of particular mood are combined together leaving 10 songs to compare with None Mood</td>
</tr>
</tbody>
</table>


34 Lee, M. 
http://www.orthointelligenc...com/2009/08/13/textualizing-b...d/ta...er.com/2009/08/13/textualizing-born-toru...m/ta...o.html

35 Oware, M. 
http://www.icto.../static/4028298...he.html

36 Weltzin, R. and Kubrin, C. 
http://www.icto.org/static/4028298/ko...e.html

37 Petrie, K., Pennebaker, J., and Stetten, R. 
http://homepages.../static/4028298/us...lyrics.html

38 Mahedero, J., Martínez, A., and Cano, P. 
https://www.research...y/edu/veh...c/15373471

39 Logan, B., Kositsky, A., and Moreno, P. 
Semantic Analysis of Song Lyrics.
to characterize semantic content and discover genre clusters among artists.


Quantitative Analysis of Literary Styles 2002 Overview and brief history of the analysis of literary styles. Use canonical discriminant analysis and principal component analysis to identify structure in the data and distinguish authorship. Assume: (1) the style of an author remains the same throughout his/her life; (2) successive occurrences of function words are independent. Neither assumption tends to hold in practice.

None Raw data for this study were obtained from Internet websites such as Project Gutenberg. Multiple works for each author were downloaded in text format and processed. The titles and website URL are listed in Appendix A.

9 authors and 59 works Divide each author’s work into 1700 words. For each block, we tabulated the frequency of the 69 words chosen from the Miller-Newman-Friedman list of function words used. Canonical discriminant analysis was used to provide dimension reduction and graphical displays of the differences between authors (canonical vector plots). Also, CDA was useful for identifying key function words which were most effective at discriminating between authors. The key words were identified by examining plots of the loadings for each function word. Function words were chosen as the unit analysis because they are highly variable between authors, abundant, and easy to count and identify. Function words include:

(a been had its one the were all but has may only their what also by have more or then when an can her must our there which and do his my should things who any down if no so this will are even in not same to with as every into now such up would at for is of than upon your be from it on that was)

Scatterplots of Components within Authors and between authors Function words