Analyzing Rap Lyrics – Part 3: Analytical Results and Implications

Posted on **June 20, 2016**

Part 3 is divided into 3 major sections. The first section focuses on the results produced by analyzing the changes and trends in lyrical "style" across time for our sample of 1206 rap songs. In this section the results were generated from a series of standard one-way ANOVA tests (with polynomial contrasts) run against the data contained in the *song-property matrix* described at the end of Part 2. The second section of this part discusses the results produced by analyzing the trends in lyrical "content" across time for the sample. More specifically, these results come from applying an analytical technique called *non-negative matrix factorization* (NMF) to the *song-stem (td-idf) matrix* also described at the end of Part 2. Finally, the third section deals with the practical implications of this and similar sorts of research, most of which revolve around a larger body of research known as **Music Information Retrieval** (MIR).

Trend Analysis: Selection of Time Periods

Even those unfamiliar with the genre can recognize that rap and hip-hop are not what they used to be. A pre-2005 hip-hop or rap hit can be easily distinguished from a track released in the past decade, and artists who have gotten into the game within the last ten years bear little similarity to what was the norm for '90s-era rappers... Perhaps the most striking difference between 1990s hip-hop and more modern tracks is the lyrics (McNulty, 2014).

One book that attempts to provide a historical understanding of some of the shifts in (American) rap lyrics is *The Anthology of Rap* (edited by Bradley and Dubois, 2010). The book is divided into four historical sections: (1) Old School (1978-84); (2) Golden Age (1985-1992); (3) Mainstream (1993-1999); and the New Millennium (2000-2010). Similar distinctions were made in a special magazine issue devoted to *Hip Hop Legends* (Jaime, 2015), although the time periods and genres were slightly different and included: Old School (1970-1982); Golden Age (1983-1987); Gangsta (1988-1998); and Millennium (1999-2010). In these, as well as other informal discussions of the historical evolution of rap, the naming and delineation of the evolutionary periods have been based on "practical means of organizing the material" rather than on a well-formulated and consensual set of terms and times.

In the trend analysis of rap lyrics provided in this discussion, time is treated differently for the two types of analysis – lyrical "style" and lyrical "content." This mimics the approaches taken by earlier researchers dealing with the same topics.

In the case of "style," one-way ANOVA (with contrasts) has often been used to examine the changes in various stylistic variables across time. Here, *time* has been typically divided into

groups or periods. For example, this is the approach used by Petrie et al. (2008) and Fell and Sporleder (2014) in their historical analyses of the lyrics of various genres. In the analysis of style that follows time is divided into six periods. The specific periods along with the number of (sample) songs in each period are shown in Table 1:

Period	No. of Songs
(P1) 1980-9	40
(P2) 1990-4	197
(P3) 1995-9	247
(P4) 2000-4	335
(P5) 2005-9	345
(P6) 2010-15	42
Total	**1206**

Table 1. Time Slices with Sample Song Counts

With the exception of the first and last periods, each of the other time periods is 5 years long. The first (1980-9) and last (2010-15) periods were extended beyond 5 years primarily to ensure that there were enough songs for comparison with the other periods.

This is the same sort of approach used by Petrie et al. (2008) and Fell and Sporleder (2014) in their historical analyses of other types of song lyrics (although the exact slices of time were different). It is not the approach used in "The Evolution of Popular Music: USA 1960-2010" (Mauch 2015), nor is it the approach used in "The Evolution of Pop Lyrics and a Tale of Two LDA's" (Thompson 2015). As noted earlier, these two "Evolution" studies examined shifts by year in the audio and lyrics, respectively, of the US Billboard Hot 100 between 1960 and 2010. In each instance the analysis was based on the individual years rather than larger time periods because the sample size was large enough to support more detailed analysis (i.e. with approximately 17K in the sample there was an average of about 350 songs per year). With the sample of 1206 rap songs analyzed here, many of the earlier and later years had less than 5 songs which makes meaningful comparisons among the individual years difficult.

Trends in Style

To reiterate what was said in Part 2, while lyrical style covers a wide range of properties, the focus of this analysis is on a subset of four stylistic dimensions including:

Dimension	Feature	Measures for Each Song in Corpus
Vocabulary	Lexical Diversity	No. of Lowercase Voc./No. of Lowercase Terms
	Uncommon Words	No. of Lowercase Terms not appearing in Wiktionary
	Slang Words	No. of Lowercase Terms appearing in Urban Dictionary
Style	Length	Lines per song, Lowercase Terms per song, Lowercase Tokens per line
	Rhyming	Rhyming Factor computed from Rapalyzer Algorithm (Malmi 2015)
	Echoisms	No. of Lowercase words with repeating letters (3 or more) and/or repeating substrings
Orientation	Temporal	Ratio of past tense verb forms to present and future tense (LIWC)
	Egocentricity	Ratio of 1st person pronouns to 2nd and 3rd person(LIWC)
Semantics	Conceptual Imagery	Inventory of Psychological and Social Processes and Personal Concerns ascertained by the Linguistic Inquiry and Word Count (LIWC)

Table 2. Dimensions, Features and Measures Incorporated in Statistical Style Analysis

With the first three dimensions – Vocabulary, Style and Orientation – only a handful of specific measures are considered. In the case of the fourth dimension – Semantics – the feature of "Conceptual Imagery" is a surrogate for a large subset of stylistic features and measures based on the research of *Pennebaker et al.* and produced by a computer application called the Linguistic Inquiry and Word Count (LIWC) that was derived from this research. The larger subset of conceptual features and measures considered here includes:

Feature	Measure	Feature	Measure
Affective	% Affective	**Perceptual**	% Perceptual
	% Positive emotions		% See
	% Negative emotions		% Hear
			% Feel
Social	% Social	**Biological**	% Biological
	% Family		% Body
	% Friends		% Health
	% Female References		% Sexual
	% Male References		% Ingestion
Cognitive	% Cognitive	**Drives**	% Drives
	% Insight		% Affiliation
	% Causation		% Achievement
	% Discrepancy		% Power
	% Tentative		% Reward
	% Certainty		% Risk
	% Differentiation		

Table 3. Features and Measures for Semantic Dimension

Each of the measures listed in Table 3 denotes the % of total words that are found in a dictionary of words associated with that particular concept. For example, in LIWC there is a list of words that denote "positive emotions." Thus, if a song has a total 500 words of which 5 are found in the *dictionary of positive emotion words*, then for this song the measure would equal 1% (= (5/500)*100). In reviewing the results for these measures it is important to note :

(1) Sub-features are not exhaustive – The top measure for each feature is an "overall" measure for that feature, while the subsequent measures associated with that feature represent key sub-features. The sub-features are not exhaustive, so the percentages for these do not sum to the overall percentage. For instance, the *% Affective* measure is the overall measure for the *Affective*

feature. The *% Positive* and *% Negative* emotions are key features for the *Affective* feature. You can't add the *% Positive* to the *% Negative* to get the *% Affective*.

(2) Features (and sub-features) are not mutually-exclusive – Words can, and often do, appear in the dictionaries for more than one feature of sub-feature. For instance, the word "love" is in the dictionaries for the *Affective* and *Drive* features (as well as the dictionaries for *Positive* emotions and *Affiliation*). So, again, you can't add the overall percentages for the various features (like Affective, Social, … Drives) to come up with some total percentage for the Semantic Dimension or to come up with the total *Word Count*.

One-way ANOVA with Linear and Quadratic Contrasts

Following the example set by Petrie et al. (2008) in their analysis of changes in the Beatles lyrics over time, simple one-way analysis of variance (ANOVA) was used to determine whether there were any statistically significant changes (p values >= .01) in the various features over the six time periods delineated in Table 1. For those measures where the overall ANOVA was significant, *polynomial contrasts* were computed to determine whether there were any significant *linear*, *quadratic*, or *cubic* effects. In general, the nature of the historical trend is best described by the highest power or order of the significant effects. So, for instance, if the only significant effect is the linear, then the measure under consideration is either trending upward or downward in a linear fashion. If the only significant effect is the quadratic, then the trend is curvilinear with a single inflection point (either convex or concave). Finally, if the only significant effect is cubic, then the trend is curvilinear with multiple inflection points (like a wave pattern). Of course, often two or three of the effects are significant. In these cases the interpretation is a little more complex but is usually (but not always) best to describe the effect with the highest power (e.g. if the linear and quadratic effects are both significant, then the curvilinear trend usually prevails).

While Python was used to perform most of the processing done in Parts 1 and 2, in this analysis I've relied primarily on R. Even though Python has a variety of statistical packages that can be used for doing one-way ANOVA, it's easier to analyze the polynomial contrasts with R. The results from the one-way ANOVA of all the stylistic features and measures are shown in the attached *spreadsheet of ANOVA results* and *PDF of trend graphs*.

The spreadsheet provides: (1) the overall means (averages) for each measure; (2) the means for each measure for the six time periods; (3) the overall significance of the F-value from the one-way ANOVA, as well as the significance of each of the effects in the polynomial contrasts (linear, quadratic and cubic); and (4) for significant relationships, a summary description of the general pattern for the trend (e.g. positive convex) . Additionally, where available, the spreadsheet provides an overall mean for each measure based on statistics compiled from a number of LIWC studies conducted by Pennebaker et al.'s (2015). These studies were based on word usage in a variety of settings including blogs, expressive writing samples, novels, natural speech, the New York Times and Twitter (in the remainder of the discussion they are collectively referred to as the *LIWC Studies*). Combined, they represent the "utterances of over 80K writers or speakers totaling over 231 million words." These combined LIWC means are provided as a point of reference to see how rap lyrics compare to general usage. So, for example, from the spreadsheet we can see that rap lyrics utilize fewer *Past Tense* verbs than the earlier LIWC studies (2.8% of all the words in the

songs are past tense verbs, while the number is 4.6% of all the words in the LIWC studies), while the use of *Power* words in rap lyrics is surprisingly very close to the general LIWC usage (2.4% to 2.5%, respectively).

The PDF file contains a series of line graphs plotting the trend in means and associated standard errors for each of the measures by time period. For each line graph a "dot" represents the mean for a given time period and the "whiskers" represent 1 standard error above and below that mean. The plots for each measure also contain a blue line representing the mean of the measure for all rap lyrics and, where available, a red line representing the mean for that measure in the earlier LIWC Studies. In addition, the overall significance of the F-value from the one-way ANOVA and the significance of each of the effects in the polynomial contrasts is noted at the top of each plot along with a summary description of the overall trend. For instance, the plot in Figure 1 below shows the trends in the mean usage for *first person pronouns* (I and we) in the 1206 rap lyrics from 1980 to 2015. From this figure we can see that overall mean percentage of first person pronouns (*PerFirst*) in the sample was just under 9% (blue line), while the mean for the LIWC Studies in general was under 6% (red line). The percent declined in the earlier years from P1 to P3 (19905-1999) but has been on the rise since then. Looking at the standard errors from one period to the next, there was obviously more variability from song to song in P1 (se ~ 5% from 1980-89) than for the other periods (se ~ 3%) which raises some issues about homoscedasticity (which I'll ignore for the moment). Based on one-way ANOVA, the overall F-value is significant at the .001 level (***) with both the linear and quadratic effects also significant at this level, while the cubic effects were not significant (note the line at the top "***,***,***,NS: convex w/ increasing slope"). Based on these significance levels, one could describe the overall trend as a "convex curve with an increasing slope" (which is also known as a *concave upward* or *convex downward* curve).

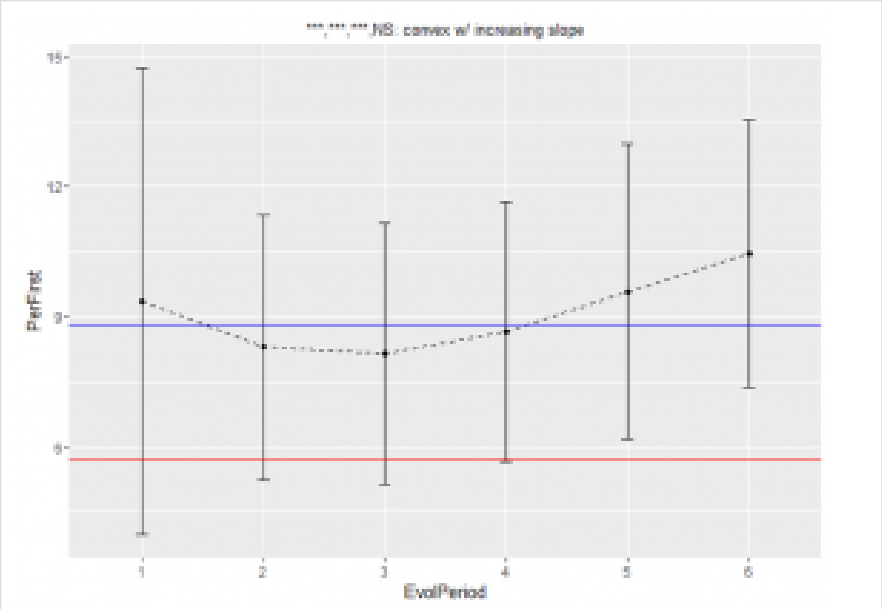

Figure 1. % of First Person Pronouns for 1206 rap lyrics by time period (listed in Table 1).

"A convex curve with an increasing slope" is only one of a number of trend patterns that can occur. Figure 2 provides a pictorial summary of the types of trends that are possible in this analysis.

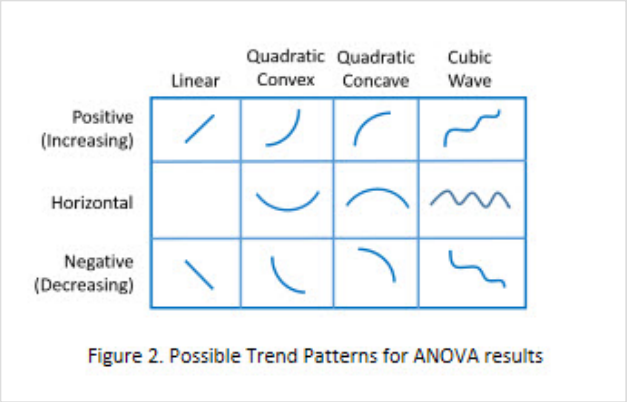

Figure 2. Possible Trend Patterns for ANOVA results

Detailed ANOVA Results

1. Vocabulary – A number of lyrical studies have indicated that for most popular genres there has been an increase in the average *Word Count* per song. This is also the pattern in this sample, although the increase was not uniform across the time periods. In P1 the average was around 450 words per song, in P2-P5 around 560, and in P6 (2010+) around 680. With this increase in the total *Word Count*, there was also a shift in the ratio of the number of unique words to the total number of words (i.e. *Lexical Diversity*). From P1 to P3 the *Lexical Diversity* went from 42% to 46% but steadily declined to an average of 34% in P6. In terms of the use of uncommon words (i.e. *% not in Wiktionary*) or slang words *(i.e. % in Urban Dictionary* but not in Wiktionary) there was no significant variation across the time.
2. Length – The changes in word count are almost a direct function of the changes in the *Number of Lines* per song and the *Number of (lower case) Words per Line*. In this case, the average *Number of Lines* rose from 63 in P1, to 72 for P2-P5, and then to 84 in P6. Not only did the number of lines increase on average but so did the *Number of Words per Line* from approximately 7 to 8. Part of the increase in the *Number of Lines* came from repetition. While the average *Number of Unique Lines* shifted from 50 in P1 to 65 in P3 and back down to about 60 in P6, the percentage of Unique Lines (not included in the spreadsheet) went from 82 to 90 to 70 in those same periods (so the percentage of non-unique or repetitive lines went from 18 to 10 to 30). So, on average one of the reasons songs are getting longer is because there is more repetition (and thus less lexical diversity).
3. Rhyming and Echoisms – There are clearly differences among rap artists with respect to their rhyming structures. While automated detection of rhyming in rap music is a very complex (as the 2010 article by Hirjee and Brown clearly demonstrates), this analysis only looks at *assonance* (i.e. “repetition of vowel sounds to create internal rhyming within phrases or sentences”) and relies on an algorithm developed by Eric Malmi (2015) which computes the length of the *Longest Matching* vowel sequence that ends with one of the 15 previous words along with the *Average Rhyme* length which averages the lengths of the longest matching vowel sequences of all words. Unfortunately, there were significant trends in either the average rhyme length nor are there any systematic changes in the longest sequences, even though they seem to vary a little bit from one time period to the next. The same is true for echoisms which measure the % of words with repetitive letters (e.g. “fuuuture”) or repetitive words (e.g. “love, love, love”). Again, there are no significant differences across time.
4. Temporal – This feature measures the degree to which a song is focused on the *past,* the *present* and the *future*. It is calculated by determining the percentages of all words that are

past, *present*, or *future tense verbs* which in this case wereapproximately 3%, 14%, and 2%, respectively. This pattern also holds for the other LIWC Studies, although the differences are not as great (with the percentages being 5%, 10% and 1%, respectively). In terms of changes across time, the *past tense* verbs declined slightly from 4% in P1 to 2.5% in P2, and to 3.0% in P6. Similarly, the *present tense verbs* declined from 15% in P1 to 13% in P3 and then to rose 15% again in P6. In contrast there were no significant changes in the use *of future tense verbs*. As a consequence, *the ratio of the past tense to present and future tense* followed the same curvilinear convex pattern as the *past* and *present tense verbs* alone.

5. Egocentric – The ratio of the percentage of words that are *first person pronouns* in comparison to the percentage that are *second* or *third person pronouns* provides an indicator of how egocentric a song is. In this sample, on average the percentage of *first person pronouns* was about 9%, while the average percentage of *second person* and *third person* combined was around 6%. For the earlier LIWC Studies, the percentages were close to 6% and 4%, respectively. Obviously, the lyrics for the sample of rap songs were generally more egocentric. From the standpoint of trends, the percentage of *first person pronouns* followed a curvilinear path going from 9% in P1 to 8% in P3 to 10% in P6, while the combined percentage of *second* and *third person pronouns* oscillated around their overall mean. The result is that the ratio of *first* to *second* and *third personal pronouns* also followed a curvilinear path from 2.4 (P1) to 1.7 (P4) to 2.6 (P6). These data suggest that the focus on oneself versus others was on the downswing until 2005 but has been on the upswing for the last 10 years.
6. Semantics – The remainder of the data deals with the "topics that are mentioned and the images they invoke." The assumption is…"that the words people use convey psychological information over and above their literal meaning and independent of their semantic context." As noted, these results were produced by applying Linguistic Inquiry and Word Count (LIWC) analysis to the lyrics. Before we consider the detailed results, consider for a moment the overall mean percentages for the various dimensions and features shown in Figures 3 and 4.

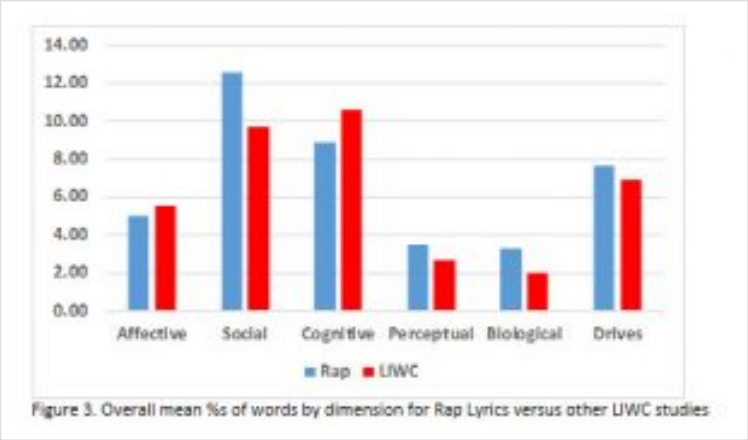

Figure 3. Overall mean %s of words by dimension for Rap Lyrics versus other LIWC studies

As Figure 3 shows, while there are differences in the percentages for a given dimension between Rap Lyrics and the other LIWC studies, the differences are usually not that large and the relative importance of the various categories is basically the same for the two. For example, while the percentage of *Social* words was higher in the Rap Lyrics versus the other LIWC Studies, the percentages for both were high relative to the other categories. In fact if you correlate the means of the dimensions for the two groups (Rap vs. Other LIWC) the correlation is .90.

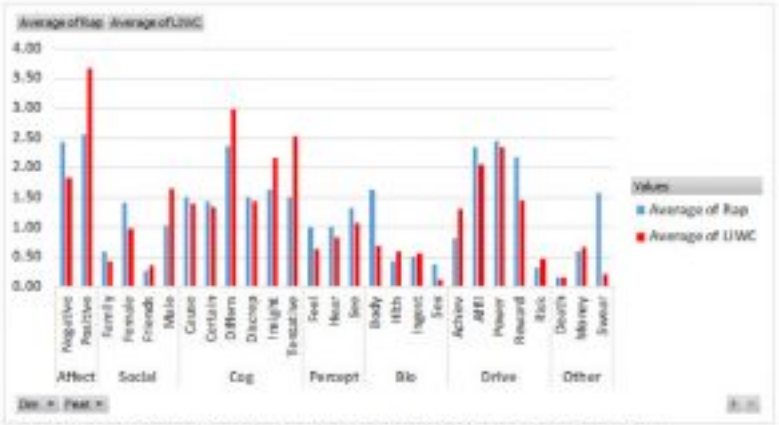

Figure 4. Overall mean %s of words by feature and dimension for Rap Lyrics versus other LIWC studies

As shown in Figure 4, the same pattern holds when you look at the various features within these dimensions. For example, while there is a major difference between the percentage of *Positive* words for Rap Lyrics and the other LIWC Studies, for both it was still the highest percentage among all the features. Again, there is a strong correlation between the percentages for the two groups (correlation is ~.8).

With these similarities as a back drop, here are the details for the various (sub)dimensions and features under the *Semantic* umbrella.

1. Affective Words – *Affective* words deal with moods, feelings and emotions. In our sample of rap lyrics about 5% of the words match the words in the LIWC Affective list or dictionary. This holds true regardless of the year. Regardless of this count, there is variation in the percentage of words that express *positive* and *negative emotions*. Overall, the percent of *positive* words (e.g. love, like) was around 2.5%. This was lower than the average percentage of positive words (5.6%) for the LIWC Studies in general. Additionally, for rap lyrics the percentage of *positive* words dropped in a curvilinear fashion from 3.6% (P1) to 2.4% (P6). In contrast, the percentage of *negative* words (e.g. hate, kill, hurt, sad) in rap lyrics increased slightly over time from 1.5% (P1) to 2.6% (P3) where it remained.
2. Social Words – *Social* words provide details about social relations and social processes (e.g. who has status and who doesn't, or who is dominating a relationship). Overall, a substantial percentage (13%) of the words in rap lyrics fall in this category, as opposed to the percentage (10%) for the LIWC Studies in general. Originally, in P1 the percentage of *social* words was 15%, but it dropped in P2 to 12%, and oscillated above and below 13% after that. In this category, fewer than 1% of the words (combined) refer to *family* and *friends* which was also true for other LIWC studies. Overall, about 1.5% of the words referred to *females*, while only 1% referred to *males*. These figures are reversed for other LIWC Studies. Additionally, the percentage of words referring to *males* steadily declined over the six time periods.
3. Cognitive Words – *Cognitive* words (e.g. think, know, because) provide clues as "to how people process information and interpret to make sense of their environment." Approximately 9% of the words in our sample of rap lyrics fell into this category which is slightly lower than it was for other LIWC Studies. For most of the subcategories in this dimension (including *insight*, *causation*, *discrepancy*, *tentative*, *certainty*, and *differentiation* words), the percentage of words was around 1.5% and, for the most part, there was little significant change in the percentage from P1 thru P6.

4. Perceptual Words – *Perceptual* words deal with the acts of seeing, hearing and feeling (i.e. touching). Overall, 3.5% of the words were in this category which was slightly higher than it was for the LIWC Studies (2.7%). The percentage in this category declined somewhat from 4.0% in P1 to 3.3% in P3 but was back to 4.3% in P6. Among the subcategories, the percentage of *See* words rose steadily from 1% in P1 to 2% in P6, the percentage of *Hear* words declined slightly from 1.5% in P1 to 1% in P6, and the percentage of *Feel* words remained pretty much the same (around 1%).
5. Biological Words – Biological words (e.g. blood, pain, intercourse) deal with various biological processes, actions, and parts of the body. Again, the percentage of words in this category was relatively small (around 3%). This was higher than it was for the other LIWC Studies (at 2%) and remained relatively unchanged from P1 to P6. The subcategories of this dimension include *Body*, *Health*, *Sexual* and *Ingestion*. The overall percentage of *Body* words is around 1.6% (the highest of the subcategories). The percent in this category increased somewhat from P1 to P4 and P5 but is now close to the overall mean. Surprisingly, the percent of *Sexual* words is very small (around .4%). However, the percent has risen from close 0% in P1 to .6% in P6.
6. Drive Words – Drive words refer to the (psychological) factors driving behavior including *Affiliation*, *Achievement*, *Power*, *Reward* and *Risk*. On average, close to 8% of the words fell into this category. This percentage was only slightly higher than it was for the other LIWC Studies (7%) and was relatively unchanged over time. Among this category, the key subcategories are *Affiliation*, *Power*, and *Reward*. These subcategories are also more prominent in the other LIWC Studies. In rap lyrics, references to *Affiliation* (e.g. friends, enemy, ally) were somewhat higher in P1 (3%) then leveled out to 2% in the remainder of the time periods, references to *Power* peaked at around 3% in P3, and references to *Reward* increased slightly from P1 (1.7%) to P5 (2.5%) but declined in P6 to the overall average of 2%. Again, it is somewhat surprising that the percentage of words dealing with *Risk* (e.g. danger) was so low (overall at .3%).
7. Other Words – Besides the dimensions and features discussed above, LIWC has been applied to a range of other words. Three classes of words that are particularly pertinent to rap lyrics are *Death*, *Money*, and *Swear* words. First, on average the percent of words pertaining *to Death* in Rap Lyrics (.2%) was the same as in other LIWC Studies (.16%). Additionally, the percentage in the sample Rap Lyrics did not changed significantly over the years. Second, the overall percentage of words dealing with *Money* was also close to the overall percent in the other LIWC studies (.6% versus .7%). Again, the percent in Rap Lyrics did not changed significantly over the years. Finally, and not unexpectedly, the percentage of *Swear* words in Rap Lyrics (1.6%) was greater than in the other LIWC Studies (.2%). This percentage increased in a curvilinear (wave-like) fashion from .1% in P1 to 2% in P6. Given the general impression of Rap Lyrics among the public, the fact that the percent of *Swear* words was on average only 2% may seem surprising. However, there was quite a bit of variability from song to song. While the maximum percentage for any song in the sample was 14%, over 80% of the songs had 2% or less. Obviously, it's the songs at the upper range that garner the attention.

Summary of ANOVA trends

Reading a detailed statistical analysis of this sort is much like getting the results from a comprehensive series of blood panel tests – there are a myriad of details which can make it difficult to see various relationships or to paint an overall picture. A general summary will be provided after the results of the "topic" analysis have been discussed. For the moment, Table 3 provides a summary of the significant trends and the patterns found in the stylistic analysis. The summary is organized by the possible trends displayed in Figure 2.

	Type			
Direction	**Linear**	**Convex**	**Concave**	**Wave**
Positive	Words Per Line Words per Uniq Line % See % Sexual % Reward	% Pres Tense % 1st Per Pron Ratio 1st/2nd 3rd	% Neg Emot % Body	Word Count Num Lines Uniq Lines % 2nd+3rd Per Pron % Swear
Horizontal	Not applicable	% Perceptual % Hearing	% Power	% Positive % Social % Female % Certainty % Affiliation
Negative	% Male % Health	Lex Diversity % Past % Ratio Past/Pres+Fut		

Table 3. Summary of ANOVA Trends from Analysis of Style in Sample of 1206 Rap Lyrics

Trends in Content: Topic Modeling

Content(s) refers to the subject, topics, or themes of a document or one of its parts. Obviously, for songs the content rests on the topics or themes reflected in the words of the lyrics. *Topic modeling* is a collection of (text mining) algorithms designed to ferret out the themes or topics that occur across a corpus of documents. A recent edition to this collection of tools is *Non-negative Matrix Factorization* (NMF). It has been successfully applied in a range of areas including bioinformatics, image processing, audio processing, as well as text analysis (e.g. sentiment of analysis of restaurant reviews). Technically, NMF is "an unsupervised family of algorithms that simultaneously performs dimension reduction and clustering in order to reveal the "latent" features underlying the data (see Kuang, 2013 and Lee and Seung, 1999) . In this case the data of interest are found in the song-stem tf-idf matrix described in Figure 10 of Part 2 of this analysis.

To understand how NMF works and what the terms "dimension reduction" and "clustering" mean, let's reconsider this matrix of tf-idf values discussed in Part 2. However, this time, instead of focusing on the song-stem matrix for the individual songs, we're going to first combine the stem counts for those songs that were released in the same time periods (according Table 2 in Part 2), and then calculate the tf-idf values for each stem in each of the groups of songs by year (the processes is depicted in Figure 5). In essence we are the treating each group of songs by year as a single song. In this way we can more easily explore the trends in content by year. For purposes of discussion we'll label this **matrix S**.

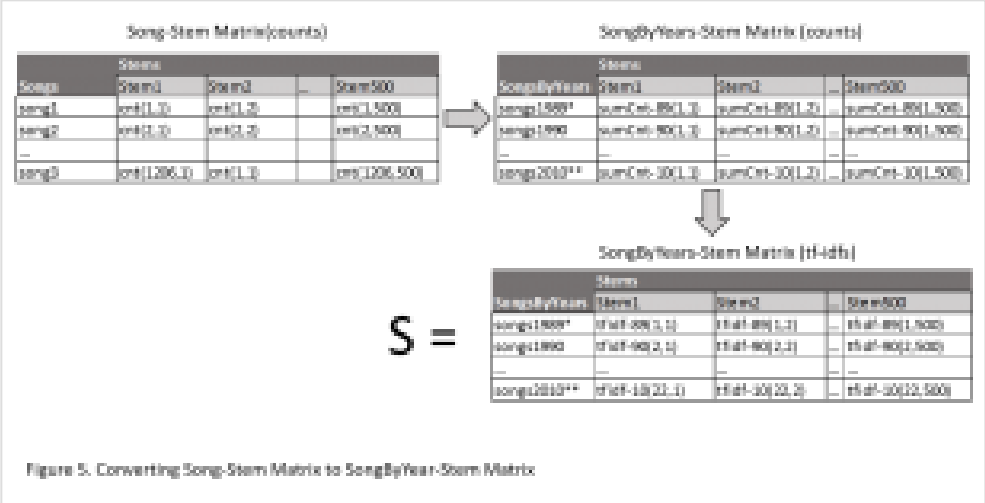

Figure 5. Converting Song-Stem Matrix to SongByYear-Stem Matrix

This matrix is 22 (years) by 500 (stems). In this analysis the stems are sorted in alphabetical order. The matrix is much too large to reproduce in tabular form (at least in this format). However, the following *heatmap* provides a general overview of the values. Each row of the map contains a color coded rendition of the normalized stem frequencies (TFIDFs) for a given year. The colors reflect the relative magnitude of a given TFIDF with the darker colors representing more important terms (see index at the top).

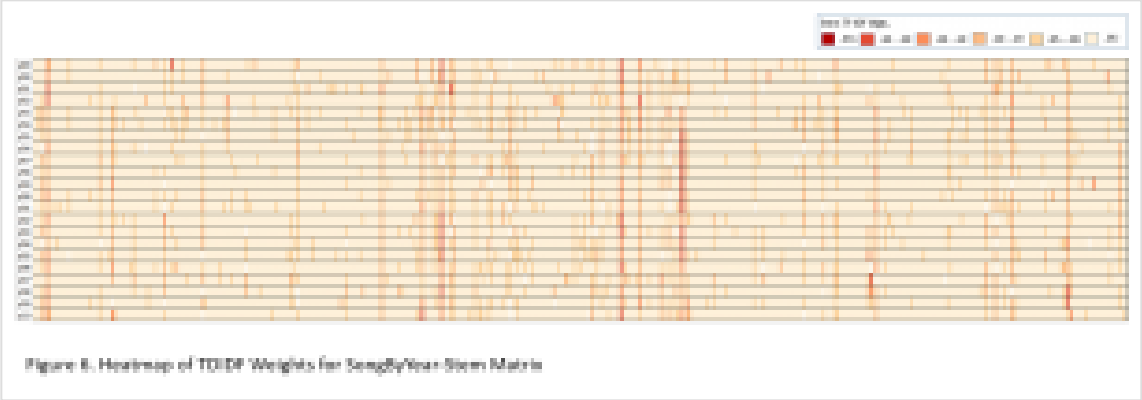

Figure 6. Heatmap of TDIDF Weights for SongByYear-Stem Matrix

The heatmap indicates that:

- For any given year (row), most of the TFIDF values are toward the lower end of the spectrum (.05 or less). These stems are of lesser importance.
- For any given year (row), approximately 10% of the stems (~50) have values toward the higher end of the spectrum (.20 or more). These are the stems that are key to understanding the content for that particular year.
- Looking down the years (across the rows), there are a variety of instances (denoted by the darker lines) where the TDIDF values appear to remain relatively strong from year to year. These instances provide visual clues to the ebbs and flows of content from year to year.

Of course, the problem with the display is it's hard to produce a succinct summary of the patterns and changes in content from year to year. This is true for both the patterns among the stems and among the songs. This is where *topic modeling* and *NMF* come into play.

Factorization: Simplifying the Representation

Mathematical factorization is the decomposition of an object (e.g. number, polynomial, or matrix) into a product of other objects which when multiplied together give the original. For example, the polynomial x2 – 4 can be decomposed into (x – 2)(x + 2). In all cases, a product of simpler objects is obtained… The aim … is usually to reduce something to "basic building blocks" (e.g. polynomials to irreducible polynomials).

This is the basic idea underlying NMF. As the name implies, Non-negative Matrix Factorization (NMF) starts with a non-negative matrix (in this case the songByYear-Stem Matrix S) and factorizes it into two smaller non-negative matrices W and T. In this instance, W is an (N x K) matrix where N is the number of years and K is the number of latent topics discovered by the factorization process. Each entry in the matrix represents the importance of a particular topic for that particular year. In contrast, T is a (K x M) matrix where K is the number of latent topics and M is the number of original stems. Here, the entries represent the importance of a particular stem for a particular topics. So, in NMF the idea is to (iteratively) derive the two sets of weights so that when W and T are multiplied together the product approximates the original matrix S (i.e. S[n x m] ~ W[n x k] * T[k x m].

In carrying out this factorization process, the first goal is *dimension reduction* which revolves around the analysis of the weights in the "T" matrix. The aim of this analysis is to surface a reduced collection of K topics that consists of a small set of highly weighted, distinctive stems that differs from topic to topic and represents a relatively coherent theme (not just a seemingly random collection of words). The second goal is *clustering* which revolves around the analysis of the weights in the "W" matrix. The aim of this analysis is to determine the similarities and difference in topics among the various time periods in order to surface trends in the changes in topics across the years.

Calculating and Analyzing Topics

There are a series of papers that detail and illustrate the specific steps and calculations used in performing NMF topic modeling (Blei, 2012; Das, 2015; Green, 2015; and Riddle, 2015A). Virtually all of these papers begin the process with a "raw" corpus, detail the NLP steps needed to turn the corpus into a document-term matrix of frequencies, then describe the steps needed to turn the doc-term matrix of frequencies into a matrix of TFIDF values, and finally submit this matrix of TDIDF values to the NMF decomposition process to yield the W and T matrices. With these latter matrices in hand, the papers finally demonstrate various procedures for analyzing the results of the decomposition.

To simplify the steps (and to avoid the necessity of having a comprehensive understanding of the underlying math) virtually all of these papers utilize Python programs to carry out the computations and to assist in the analysis. In the case of our sample of rap songs, it was very easy to perform the modeling computations. The reason why is that I already had the matrix of TFIDF values in hand (i.e. the matrix S discussed earlier) which meant that I only needed three lines of Python code to handle the computations:

1. model = decomposition.NMF(init="nndsvd", n_components=ncomp, max_iter=500)
2. W = model.fit_transform(S)

3. T = model.components_

The only real decision that needs to be made is to pre-specify the number of topics (ncomp) to be produced by the process. After a bit of experimentation, I settled on 5 topics because I found little practical difference in the results whether the number was 5, 10, 15, or 20.

Setting ncomp = 5 resulted in a T matrix with 5 rows (topics) by 500 columns (stems). As noted, each cell(i,j) provides a weight for the stem relative to the topic, so by "ordering the weights in a given row and selecting the top-ranked stems, we can produce a description of the corresponding topic." The top 15 stems for each of the topics is provided in Figure 7 (the sensitive words have been masked to protect the innocent):

Topic 0: n*gr*, f**k, s**t, b**ch, as*, m*f*ker, time, yall, hit, love, real, hoe, life, feel, wanna
Topic 1: girl, babi, wanna, love, man, shake, feel, boy, good, n*gr*, thing, gotta, move, call, start
Topic 2: man, love, babi, time, good, girl, check, rock, funki, break, rhyme, brother, play, jump, wanna
Topic 3: love, girl, babi, nothin, rock, money, music, hair, thing, time, b**ch, f**k, watch, hand, friend
Topic 4: shawti, money, wanna, girl, f**k, king, watch, rock, b**ch, night, time, damn, babi, parti, man

Figure 7. Top 15 Stems for Topics Produced by NMF Modeling

Similarly, the top 50 stems for the same 5 topics is shown in a series of wordclouds displayed in Figure 8. The size and color of the words in each of the clouds reflects their relative weights for the associated topic. One of the first things you might notice about the lists and the clouds is that there seems to be overlap from one topic to the next. However, looks can be deceiving. Actually, in this case when you calculate something like the Jaccard Index of Similarity between any two sets of stems, the overlap is relatively

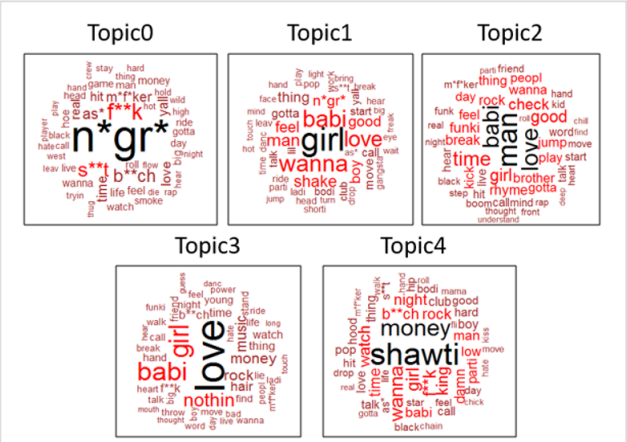

Figure 8. Top 50 Stems for Topics Produced by NMF Modeling

low (i.e. most of the pairwise similarities are around .20 out of a max of 1.0). Except for Topic 0 (which is stereotypical Gangsta), another thing you might notice is that even though there are variations in the top ranked terms, it's fairly difficult to come up with descriptive terms or titles for each of the individual topics. To do this requires a deeper understanding of the trends in the topics across the time and a look at some examples of the individual songs within those time periods.

Trend in Topics

The W matrix provides the key to understanding the associations between the various years and the derived topics. In this case the weights in the individual cells of W represent the importance of a particular topic for a particular year (of songs). By comparing successive years we can gain an understanding of the trends in topics over time. Before comparisons are made, the weights in this matrix are usually normalized so they sum to a value of 1 for each row (year) in the matrix. The results of this normalization are displayed in Table 6.

Year	Topic0	Topic1	Topic2	Topic3	Topic4
y89-	0.000	0.275	0.677	0.048	0.000
y90	0.000	0.000	1.000	0.000	0.000
y91	0.024	0.804	0.011	0.161	0.000
y92	0.000	0.495	0.000	0.505	0.000
y93	0.255	0.106	0.077	0.561	0.000
y94	0.161	0.000	0.149	0.603	0.087
y95	0.498	0.000	0.000	0.502	0.000
y96	0.532	0.088	0.000	0.380	0.000
y97	0.451	0.123	0.035	0.389	0.001
y98	0.563	0.168	0.000	0.268	0.000
y99	0.851	0.031	0.035	0.062	0.021
y00	0.493	0.327	0.000	0.179	0.000
y01	0.519	0.292	0.155	0.000	0.035
y02	0.234	0.440	0.201	0.000	0.126
y03	0.361	0.468	0.069	0.003	0.109
y04	0.375	0.494	0.000	0.000	0.131
y05	0.346	0.557	0.000	0.000	0.096
y06	0.307	0.243	0.128	0.000	0.322
y07	0.000	0.000	0.000	0.086	0.914
y08	0.104	0.443	0.006	0.000	0.448
y09	0.186	0.536	0.000	0.045	0.232
y10+	0.388	0.108	0.302	0.000	0.202

Table 6. Normalized Topic Weights by Year for W from NMF Model

As Table 6 shows, for example, Topic2 had a weight of 1.0 in y91 (1991), which means that the remaining topics (0, 1, 3 and 4) all had weights of 0.0 since the sum of all the weights for any year adds to 1.0. In y92 things were a bit different. Topic 2 had a weight of ~.8 in y92 (1992) followed by Topic3 which had a weight of ~.16. The weights for the other topics (0, 3, and 4) were negligible.

One way to visually represent the data in Table 6 is with a series of radar charts. A *radar chart* is a graph or plot for displaying multivariate data. A radar chart consists of a series of axes equal to the number of variables. The axis emanate from the same origin, the have equal length (i.e. they are normalized), and are separated by the same angles (like the spokes of a wheel). Typically, a point or dot is placed on each axis to represent the value on the associated variable. Finally, all the dots on adjacent axes are connected by lines, resulting in the formation of a closed polygon. Figure 9 shows the radar chart for a single year y95, Figure 10 shows the radar charts for all 22 time periods.

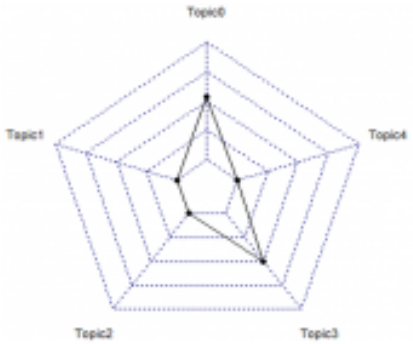

Figure 9. Radar Chart of Normalized Topic Weights from Matrix W for the Year 1995

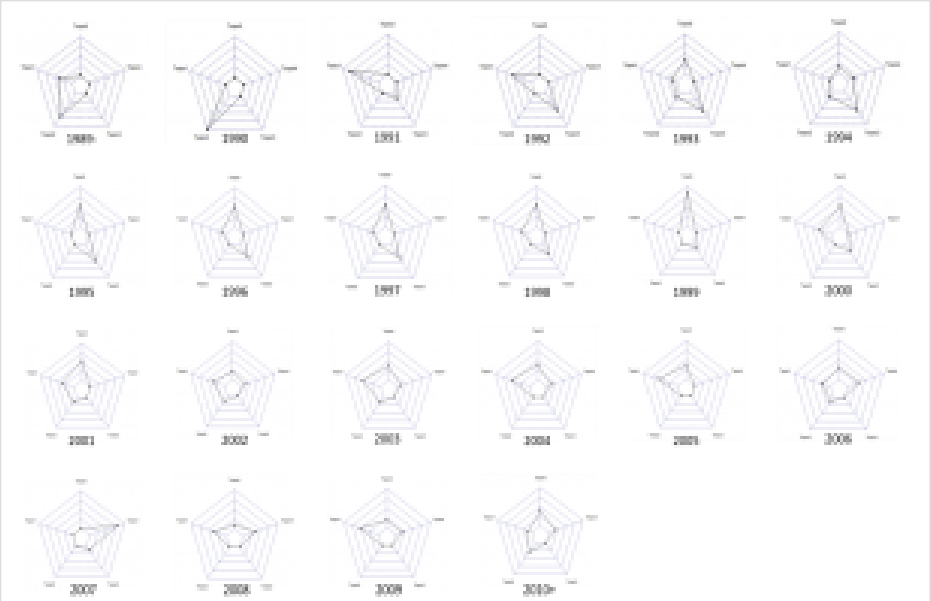

Figure 10. Radar Charts for Normalized Topic Weights from Matrix W for all Years from 1989 to 2010.

The picture that begins to emerge from Table 6 and Figure 10 is that most of the years are dominated by one or two topics (just like 1995) and that the patterns seem to run in streaks. Take, for example, the years from 1995 to 1998. They are almost clones of one another, i.e. for each of these years the weights for Topic0 and Topic3 are both close to .5 while the weights of the remaining 3 topics are close to or equal to 0.

The various streaks are also highlighted in Figure 11, which contains a matrix of correlations between the topic weights for any two years in our sample (e.g. r = 0.9 between y89- and y90). In most cases, the correlations between adjacent time periods are large (between .7 and 1.0), although periodically they aren't (e.g. r = 0.4 between y92 and y93). Blocks of highly correlated adjacent cells signal streaks where the same topics have dominated from year to year. In Figure 11, the box to the lower left summarizes the streaks based on these blocks, as well as the dominant topics in those years.

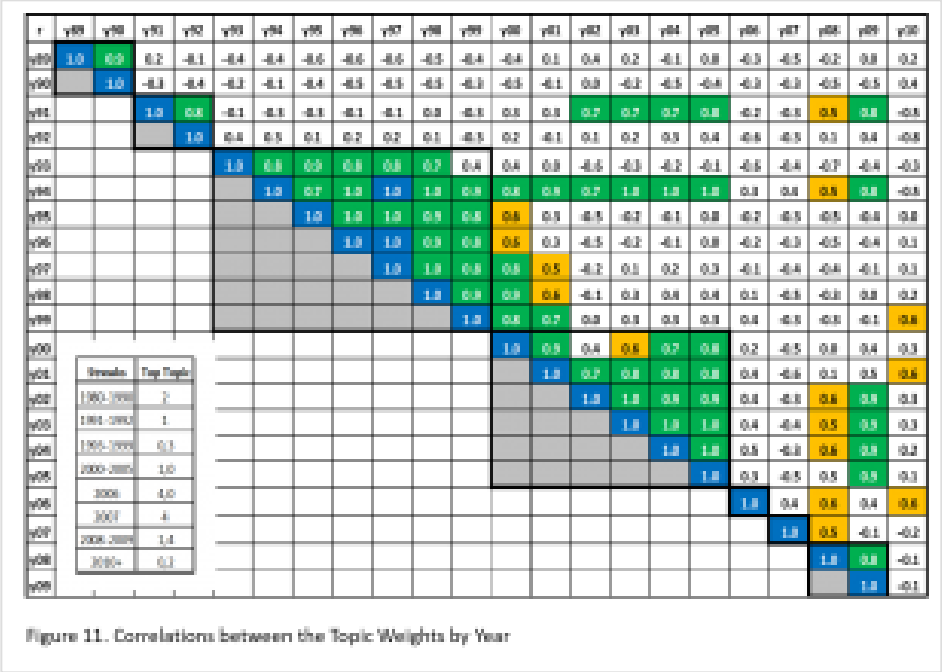

Figure 11. Correlations between the Topic Weights by Year

There are other analysis that can be run to refine the notion of streaks (e.g. variants of hierarchical clustering). In looking at the topic modeling results discussed here (as well as running a number of other analyses I won't be discussing here), it appears that the evolution of topics and themes of rap lyrics has followed the same general pattern as the evolution of music audio features found by Mauch et al. (2015). More specifically, their analysis found that, "while musical evolution was ceaseless, there were periods of relative stasis punctuated by periods of rapid change." For them there were three major revolutions within the continuous evolution – 1964 (Motown and the British Invasion), 1983 (New Wave, Disco and Hard Rock) and 1991 with rise of Rap. In the case of the evolution of rap lyrics, the shift in lyrical topics seems to have followed a similar pattern – continuing evolution punctuated by abrupt changes brought about by the rapid appearance of key sub-genres, i.e. the shifts from (1) Old School Funk and Soul to (2) Golden Age East Coast, New School, Rap Rock, and Pop Rap to (3) Gangsta G-Funk, West Coast and East Coast Hardcore and back to the future with (4) mainstream Pop Rap, Dirty South, Crunk to to West Coast to Crunk and now to more mainstream).

Analytical Summary

Sometime after starting this analysis (which has been going on forever), I began to think that it would suffer the same fate as Thompson's earlier analysis of *The Evolution of Pop Lyrics* (2015). In Thompson's words:

Sadly the results for predicting the musical genre purely based on the lyrics aren't good. It was only correct for 16% of results... The heat chart below shows a cross tab of the main lyric topic (LT) crossed with the musical genre (MG). It's a pretty even spread. Therefore it wouldn't matter what classification model was used these features aren't going to predict genres... I may return to this dataset and try some different language processing techniques to try and create some different features and hopefully improve on the model accuracy...

In other words I had the sinking feeling that when I was finally finished, the analysis would only be good for the *Journal of Non-Significant Findings* or the *Journal of Negative Results*. Those are actually real journals, but unfortunately not in my fields of endeavor. So, not only would the results be useless, but they couldn't even be published in a journal about useless results.

Nearing completion I'm a little more optimistic than that — for a couple of reasons. First, many of the results are statistically significant (at very high levels of probability). There are a number of stylistic properties in rap lyrics that vary significantly across time– like lexical diversity which has been on the decline (a concept that was the genesis of this analysis in the first place). Similarly, the data indicate that lyrical content also varies in significant ways across time. Second, unlike Thompson's analysis, the results from this analysis seem to mirror the patterns found in Mauch et al.'s (2015) study of *The Evolution of Popular Music: USA 1960-2010,* even though the subject matter differs. I already discussed their notion of "continuous evolution musical with abrupt changes" and the usefulness of this for understanding the results from the NMA topical modeling analysis. In the same vein, they also noted that, "The frequency of topics in the Hot 100 varied greatly: some topics became rarer, others became more common, yet others cycled (see Figure 12 which is reproduced from their study). By topics, they were referring to a series of harmonic and timbral features in music. The patterns they were referring to are summarized in Figure 12 (a copy of a key figure in their study). Essentially, this is the same thing that occurred with the various stylistic properties examined in this study – some went up, some went down, some cycled and some stayed the same. Finally, as they noted in their closing, "we have not addressed the cause of the dynamics that we detect." Like biological evolution, a causal account of musical evolution must "ultimately account for how musicians imitate, and modify, existing music when creating new songs, that is, an account of the mode of inheritance, the production of musical novelty and its constraints." The same can be said of this analysis.

Hindsight: Round up the Usual Suspects

Any results I've reported have to be tempered by a few caveats:

1. Sample — There a couple of problems here. First, the songs in this study come from a list generated from the Hot rap songs on Billboard. It's a source used by many of the published studies dealing with lyrics of all sorts. Obviously, these are the songs that have enjoyed commercial success and, as a consequence, are probably biased in a number of ways that can potentially impact the style and content of the lyrics. For instance, one known way (that I recently discovered when I came across on article on *Quantifying Lexical Novelty in Song Lyrics* (Ellis et al. 2015) is that *lexical novelty* is significantly lower for the songs and artists on the Billboard Top 100. Obviously, this has a number of implications for both the general content and style. Second, the sample size of 1206 songs was probably too small for a really detailed study of the trends in rap over a 35 year period. This was especially true for the songs of the 80s , as well as the songs from 2010 to the present (both around an n =40). Given this smaller sample size, it is very easy for the songs of a handful of artists to skew the results, especially those dealing with the analysis of lyrical style.
2. Independent Variable – In this study the focal point was on *time*, which seemed natural since I was interested in the *evolution* of rap. However, as noted above, the key to this analysis rests with artists and writers. So a better starting point might be to begin with a list of artists

who are representative of various sub-genres, locations and eras and build a sample of songs from the albums of these artists. In this way, trends could be viewed from a multi-dimensional rather than a unidimensional perspective.

3. Data Preparation – Among all the musical lyric genres, rap may be the easiest to recognize with automated NLP but one of the hardest, if not the hardest, to analyze with NLP. It is rife with slang, (intentional) misspellings, phonetic spellings, abbreviations, grammatical issues, and the like. It makes it extremely difficult and very time consuming to convert published lyrics into representative tokens of various sorts (e.g. into stems). This is true for both the analysis of style and content. Surprisingly, when LIWC was applied to the sample of 1206 rap songs lyrics in their original lowercase form, 85% of the words were found to be in one or more of the LIWC dictionaries which means that 15% weren't in the dictionaries. 15% may seem like a lot, but this is exactly the average percentage found in the dictionary for the earlier studies. In terms of analyzing content, quite a bit of work went into substituting one version of a word for all its different forms before any NLP was done. For example, the word m*f*ker had enumerable spellings but was translated into its correct spelling before processing. Of course, the problem is that it is virtually impossible to catch all these ahead of time, so it take a number of iterations to handle the corrections.
4. Dependent Variables – In this analysis we used very crude measures for handling rhymes. There are a number of paper devoted solely to this subject. In subsequent analyses, it would be preferable to investigate this particular aspect in more detail. Similarly, this analysis only focused on unigrams and, like a number of earlier studies, completely ignored bigrams or trigrams. The notable exception was Fell and Sporleder study of Lyric-Based Analysis and Classification of Music (referenced in Part 2 of the analysis) which found that n-grams (n<= 3) were the most important factor in classification tasks over and above a range of other stylistic and content variables. The use of bigrams or trigrams might assist with another major caveat – interpreting the topics and themes.
5. Interpreting Topics – NMF is a type of unsupervised machine learning technique. That is, it "infers a function to describe hidden structure from *unlabeled* data. Basically, it cranks the numbers and assigns a weight to a generic topic label (like the weight for Topic0). In contrast, with supervised learning the topics (and their meanings) are known ahead. The sample on which the model is built is divided into various training and testing sets. The goal is to use the training sets with the learning technique to arrive at an algorithm which when used with the testing sets will correctly classify or predicted the known topic for a given object. As noted, a major issue with unsupervised techniques like NMF is that once we've arrived at the weights for the various topics it is usually very difficult to arrive at a meaningful label (unless you know some of the possibilities *a priori*).
6. Other Modeling Techniques – There are a number of other techniques that can be used for topic modeling besides NMF. I chose it primarily because it is easy to understand the meaning of the decomposed factors (i.e. matrices W and H). Latent semantic analysis (LSA) and Latent Dirichlet allocation (LDA) are two other unsupervised techniques I've used in the past for other topic modeling projects. It is possible that one of these other techniques might be preferable (that's work for another time).
7. Visualizing the Results – There are a variety of static visualizations that can be used with this sort of trend and topic analysis. This analysis only employed a few including standard bar charts, heatmaps, word clouds, and radar charts. Although they weren't used here, I also created a variety of dendogram, other heatmap types, stacked area charts, streamgraphs,

themerivers, and denogram/heatmap combinations. They were constructed with a combination of R, Python and Excel. For those who are interested, take a look at Visualizing Topic Models at a site devoted to Text Analysis with Topic Models for the Humanities and Social Sciences (Riddell 2015B). These programs are written in Python. If you prefer R, then the Tutorial on Area Charts at FlowingData.com is worthwhile. This is Nathan Yau's site, he's the author of a Visualize This and Data Points. It costs a bit of money to access the tutorials but it's worth it if you are a frequent user of R.

What I haven't tried is any interactive visualizations with this data set. These are especially useful if you're trying to develop applications that enable users to interactively browse, search and retrieve song lyrics based on the results of real-time analysis of lyric content and style. These sorts of visualizations provide one entry way into more practical applications of this sort of research.

Applications of Analysis

The Ellis (2015) article cited above, begins with the intriguing claim that, "From 2004 through 2013, both U.S. and worldwide Google searches for "lyrics" outnumbered searches for 'games', 'news', and 'weather', as computed by Google Trends." Actually there's an update. The relative positions of the frequency of search for these 4 terms has shifted somewhat since 2013 (Figure 12) so that "lyrics" no longer dominate. Yet, the sentiment still remains the same. The interest in lyrics is still very high and it continues to motivate "several explorations for translating a song's lyrics into "queryable" features: for example, by topic, genre, or mood."

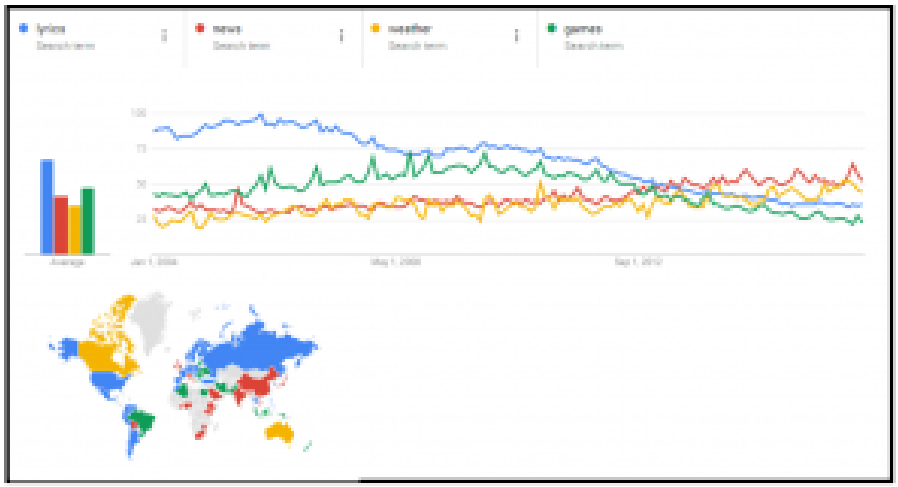

Figure 12. Google Trends Results for Queries about the terms Lyrics, Games, News, and Weather from 2004-16

This interest in the queryable features of lyrics is really a small part of the larger interest in the area of *music information retrieval* (MIR). While the research side of *MIR* covers a range of topics including "classic similarity retrieval, genre classification, visualization of music collections, or user interfaces for accessing (digital) audio collections" (Schedl 2014),

[It]... is foremost concerned with the extraction and inference of meaningful features from music (from the audio signal, symbolic representation or external sources such as web pages), indexing of music using these features, and the development of different search and retrieval schemes (for

instance, content-based search, music recommendation systems, or user interfaces for browsing large music collections)…

Table 7, which comes from Downie (2009) and has been modified somewhat, details some of the common tasks undertaken by those with an interest in domains as varied as "digital libraries, consumer digital devices, content delivery and musical performance."

Specificity	Task
High	Music Identification
	Rights Management, Plagiarism Detection
	Multiple Version Handling
	Melody Extraction And Retrieval
	Performer or Composer Identification
	Recommender Systems including Playlists
	Predicting Hits
	Style, Mood, Genre Detection
Low	Music-speech Segmentation

Table 7. Common Tasks in MIR

MIR is really in its infancy (for an overview of the field see Downie 2009 and Schedl et al. 2014). *MIR* only started 10 to 15 years ago and really has only one major international society (ISMIR) solely devoted its study. Regardless, it has certainly captivated the interest in and investment from of a number of major players in the world of commercial music including companies involved in obtaining, storing, indexing, identifying, streaming, recommending, and delivering to music products and services end consumers.

Much of the early interest and activity both from a research and commercial standpoint has revolved around the audio side of music life. More recently, the span of interest seems to be widening to include the lyrical side. How this might playout will be the subject of a future posting.

Resources

References

Blei, David. Probablistic Topic Models. April, 2012. Communications of the ACM.

Bradley, A. and A. Dubois. The Anthology of Rap. 2014. Yale University Press.

Das, Sudeep. "Finding Key Themes from Free-Text Reviews: Topic Modeling." January, 2015.

Downie, J. "Music Information Retrieval." 2009.

Ellis, R. et al. "Quantifying Lexical Novelty In Song Lyrics." Proceedings of the 16th ISMIR Conference, Malaga, Spain, October 26-30, 2015.

Greene, D. “NMF Topic Modeling with scikit-learn.” March, 2015.

Hirjee, H. and D. Brown. “Using Automated Rhyme Detection to Characterize Rhyming Style in Rap Music.” October, 2010. Empirical Musicology Review.

Jaime, M. *Hip Hop Legends: 55 Game Changing Artists.* 2015. Engaged Media.

Kuang, D. et al. “Nonnegative matrix factorization for interactive topic modeling and document clustering.” 2013.

Lee, D. and Seung, H. “Learning the parts of objects by non-negative matrix factorization.” October 1999.

McNulty-Finn, C. The Evolution of Rap.” April, 2014. Harvard Political Review.

Pennebaker, J. et al. “LIWC Linguistic Inquiry and Word Count: LIWC2015.”

Malmi, E. “Algorithm That Counts Rap Rhymes and Scouts Mad Lines.” February, 2015.

Riddell, A. “Topic Modeling in Python.” 2015A. Text Analysis with Topic Models for the Humanities and Social Sciences.

Riddell, A. “Visualizing topic models.” 2015B. Text Analysis with Topic Models for the Humanities and Social Sciences.

Schedl, M. et al. “Music Information Retrieval:Recent Developments and Applications.” 2014. Foundations and Trends in Information Retrieval.

Places (Virtual and Real)

International Society of Music Information Retrieval (ISMIR).

Tools and Modules

Basic Radar Chart in R – “fmsb” package.

One-way Analysis of Variance (AOV) with Contrasts – R “stats” package.

Linguistic Inquiry and Word Count (LIWC).

Line plot for Means and Standard Error – R “ggplot2” package with geom_line and geom_errorbar.

Non-Negative Factorization Matrix (NMF) Topic Modeling in Python. Also, see Sudeep Das’ article.

Word Clouds in R – “wordcloud” package.

This entry was posted in **Uncategorized** by **daveking63@gmail.com**. Bookmark the **permalink [http://dataffiti.com/2016/06/20/long-time-between-posts/]** .