Nikolas Wolfe

11-761 Language and Statistics

Assignment 1

1/23/2015

**Question 1**

For each of the following two problems (A and B), formulate it in the source-channel paradigm:

1. Given satellite images of crops in a field one might hope to determine which, if any, disease is currently affecting the crops.

**Input**: Image **Output**: Classification, e.g. {‘disease’, ‘locusts’, etc.}

The source-channel paradigm assumes we have a *source* signal x, a *receiver* of the signal, and a channel which performs some kind of transform on x to get x’, which is read by the receiver. To model this, we are going to flip the above.

Receiver: A crop image

*C*

Signal: A Disease

*D*

Channel

*Converts D into an in image C, maximizing the probability of C given D.*

P(C|D)

The posterior, **D’**, is the argmax of all possible diseases that maximizes the likelihood of the observed image **C.** We can write this as follows:

**D’ = argmax P(D|C) = argmax P(D) P(C|D)**

The prior, **P(D)**, is the likelihood that a crop has a given disease **D** a priori. We then find the disease that maximizes the prior times **P(C|D)** which is the likelihood of a crop image *given* a particular disease. To model **P(D)** there’s a variety of things we can do, but it is probably most important to be able to tell how likely it is that the image is not to be confused with 1.) Another disease and 2.) A false alarm. A crop image is akin to, say, an HIV blood-test, or a TB skin test. The test itself is error prone, even if we are 99% sure that we are correct. For instance:

|  |  |  |
| --- | --- | --- |
| **Outcomes** | P(Crop has disease D) | P(Does not have disease D) |
| P(Image C shows disease D) | True Positive | False positive |
| P(C not shows disease D) | False Negative | True Negative |

The prior in this case is the marginal probability of the cases where the crop has disease **D**, regardless of the image, which in this case is the sum of the true positive and false negative. Furthermore, the prior is just a guess. Once we go through a trial, the posterior **D’** becomes our new prior, which accounts for the new evidence presented by practical experience. So the prior will get better with time.

1. Choose a language technology on your own that was not discussed in class.

**Problem: Music recommendations based on one’s listening history.**

**Input**: List of a person’s previously heard songs

**Output**: A new song that the person is likely to enjoy

The source-channel paradigm assumes we have a *source* signal x, a *receiver* of the signal, and a channel which performs some kind of transform on x to get x’, which is read by the receiver. To model this, we are going to flip the above.

Receiver: A list of songs

*L*

Signal: The recommendation

*R*

Channel

*Converts R into L, maximizing the probability of L given R.*

P(R|L)

The posterior, **R’**, is the argmax of all possible songs that maximizes the likelihood of the observed playlist **L.** We can write this as follows:

**R’ = argmax P(R|L) = argmax P(R) P(L|R)**

The prior, **P(R)**, is the likelihood of a person listening to a song **R** a priori. We then find the song that maximizes the prior times **P(L|R)** which is the likelihood of a playlist *given* **R**. The prior in this case is the marginal probability of the cases where a person listens to **R**, regardless of the person’s listening history. Once we go through a trial, the posterior **R’** becomes our new prior, which accounts for the new evidence presented by practical experience. So the prior will get better with time. Furthermore, if the person liked the song, that song **R** will become part of **L**, which can help with later recommendations.

**Question 2**

**Motivation – A Commentary**

Popular music is one of the most divisive forms of social self-identifcation. I remember, growing up in the stratified and racially tense cultural environment of South Florida’s public school system (in the 90s, no doubt), the music you listened to determined most of your social life. People used to self-authenticate with such expedient terms as “baser” (meaning you listened to rap and hip-hop) or “metalhead” (meaning you listened to rock and metal.) You either conformed to a specific genre and were included in the so-concerned social circles, or you were excluded. The more strictly you adhered to a particular sub-genre of music, the more likely you were to be liked and respected by your friends. The lines in the sand were fixed. One often faced severe social pressure for even trying to cross over. If desired, one might even choose to commit social suicide with their current group of friends (and take up with an entirely new set of people) simply by declaring that the type of music they liked had changed, and appropriately switching out the stickers on their backpack. And, if one chose not to follow this process of rigid self-imposed segregation you wound up being implictly forced into a special group known as the nonconformists.

It was stupid. But we were just kids.

Today, now that the people who grew up in that day and age are starting to fill the ranks of so-called established social commentary, I am beginning to see the real effect that this type of childhood environment has had on my generation. Far beyond “getting over it,” I notice there’s a real dearth of variety in the musical preferences of many people my age. And, as we in the computer science field know all too well, much of technology today is being built specifically to accommodate people in their comfort zones to further isolate themselves both culturally and musically, and to find other people who think and do likewise more efficiently than ever before. People assume there’s some choice in the matter, but many social psychologists would argue there’s likely a well-hardened childhood bias at work here as well.

In typical fashion, as people grow older they try to intellectualize their inherently irrational preferences, and that’s what leads people like Virgil Griffith (a now-famous software developer and social commentator from Caltech) to observe some very trite statistical correlations between the publicly listed musical preferences of people on Facebook, the publicly listed universities they attend, and the average SAT scores of people at those universities. But wait, it gets better! Mr. Griffith “published” his findings on his personal website using a graph generator that laid out a continuum of musical preferences plotted against an axis of average SAT scores. Well, that’s fun, right?  
  
Turns out this post of Mr. Griffith’s went viral and was reposted many times over by some of the biggest news organizations in the world, including but not limited to the Washington Post, the UK Telegraph, The Wall Street Journal, and pop culture site TMZ. If you Google the words “intelligence and musical taste” the entire first page of results is a series of high-profile reposts of Mr. Griffith’s graph. Suggestively baited titles like, “Can the music you listen to determine how clever you are?” quickly earn clicks from lay Internet surfers looking for validation. And boy do they find it! “Smart People Listen to Radiohead and Dumb People Listen to Beyoncé, According To Study” reads a blog post on Metro.co.uk.

All this could be chalked up to a “nuh uh!” and moving on with our lives, but there’s another theme at work here. This is not the first “study” to conclude that musical preferences are a good predictor of intelligence. But why is it that they always seem to “discover” that rap and hip-hop are for “dumb” people? Why is it that black musicians are always the first on the chopping blocks of social commentary?

It’s not a coincidence. When jazz music first began gaining popularity in the U.S., as Amiri Baraka writes in a 1960 critique, most of the critics were white, and most of the musicians were black. It took a while before white Americans “warmed up” to jazz music, just as it took a while before there were any white hip-hop artists. There is a long historical tradition in America of inventive new trends in music coming from black musicians, and white Americans criticizing them. The type of juvenile commentary offered by Virgil Griffith’s “study” is just another edition to this long, racist social commentay built on nothing.

# But there’s a new devil in the details of Mr. Griffith’s commentary: Math. There is a strong social bias towards pseudo-intellectually based arguments in America, and a severe lack of critique in response to social commentary which builds its legitimacy not on the real arguments it intends to make (e.g. “Rap is crap!”), but rather a bluster of words which are related to statistics in some way. On one level it’s a smart bait-and-switch. I have no doubt that the correlations Mr. Griffith calculated from the data he was working with are, so to speak, “mathematically” sound. But data is easy to fudge in this way. For example, there is a 0.993 correlation between the divorce rate in Maine and the per-capita consumption of margarine in the U.S. since the year 2000 (source: http://www.tylervigen.com/). Does that mean I can estimate the likelihood of a successful marriage in Maine this year based on the stock price of I-Can’t-Believe-It’s-Not-Butter?

# No. Of course you can’t. Neither can you say that someone who listens to Beyoncé is dumb or any more likely to be dumb *as a result* of this musical preference. With enough data I can say pretty much anything I want about Beyoncé’s cancer-curing fans, and pass it off with the legitimacy of faux statistics. In this situation, however, I think there’s another way. What can we actually say about the critiques of rap and hip hop *using* statistics? One of the most common nonsensical arguments about rap is that it requires no musical, literary, or lyrical talent, that its content is repetitive and vapid, and that all that’s going on is a flurry of expletives padded by animal sounds. Can we show that this is not the case? Well, you could, y’know, just listen to some hip-hop with an open mind, but that’s too much to ask of the lay (typically white) rap-hater.

It would be fair, I think, to show that music typically listened to by white, affluent Americans is no more linguistically “interesting” than rap and hip hop – and there are perfectly appropriate empirical things we can calculate in this respect without getting into subjectives – but maybe we can even do better than that. Who is the greatest lyricist – the greatest writer – of all time in the English language? Tupac? No, silly! Shakespeare! With a vocabulary of roughly 27379 word types (after stop word removal), Bill Shakespeare’s sonnets and plays are among the most lexically rich writings in the English Language. I wouldn’t be so bold in my hypotheses to predict that any single rapper has a greater lexical inventory than Shakespeare, but in this study we will unequivocally run the exact same analyses on the complete works of Shakespeare that we will on the likes of the most popular rap, hip-hop, rock, metal, and country artists of that legendary musical decade – the 1990s. This will ground our numbers with some anchor in the real world. How rich *are* any particular artist’s works compared to the greatest of all time? This seems like a useful sanity check. Anyway…

**My Hypotheses**

I intend to calculate lexical statistics on a representative sample of artists from the 1990s in what have commonly been referred to as polar opposite genres: Rap, hip-hop, rhythm & blues (R&B) – and rock, grunge, heavy metal, country, etc. One might suggest that we’re trying to lump white artists and contrast them with black artists, but I will assert no such thing. These are self-evidently considered by society to be “opposite” genres in competition for fans, whose listeners seldom exhibit much cross-over. At least, that’s how it seemed in the 90s. But why the 90s? Both of these genres also experienced so-called “golden ages” in the early-to-mid 90s. Almost all of the canonical artists in these genres got their start in the music scene of the 1990s. With 15 years of hindsight on an interesting decade at this point, it is clear that there were some extremely rich developments in hip hop and heavy metal in tandem in the early 1990s. It would therefore stand to reason that something interesting can be said about both of these genres by examining corpora of their respective lyrics from that time. But what will we find?

My prediction is that we will observe that hip-hop and rap (HHR) music are more lexically diverse than rock, metal, and country (RMC) music, and that HHR music is actually comparable as a whole to the works of Shakespeare in terms of its linguistic features and lexical variety.

By demonstrating this, I intend to show that the very vector in which HHR music *seeks* to excel – language – it both succeeds and also does a lot more with the ever-changing English language than it’s often credited. And when history looks back on this period it will appreciate better than it currently does the contributions of HHR music to human expression – just as we do with that wordsmith Shakespeare today. Shakespeare, for all of his celebrated talents, was never one to shy away from pushing the English language in a direction it didn’t necessarily want to go. He wrote things – often – as they were actually spoken, commuted consonants and the like – and made people speak differently by writing things in new ways. He did, in a way, what rap music is doing with English today. Fo rizzle.

This will not prove anything about the intelligence of the listeners of any particular genre. It will, however, lay to rest any nonsensically lingering notions of lyrical or lexical impoverishment in HHR music – precisely the domain it as a genre seeks to explore. There’s a reason heavy metal lyrics are few and far between – it’s about the guitar solos, dude. Likewise, it doesn’t make sense to criticize HHR music on the basis of sampling, repetitive “musical” content, or “playing” an instrument of any kind. Playing an instrument is not the point of HHR – it’s about wordplay. And play with words they do, in amazing ways.

**Notes on Methodology**

When I started this project I looked for existing corpora of music lyrics. I found none. Furthermore, I was interested in a dataset for the music that typified popular culture in the 1990s. While websites offering lyrics for songs are mostly free, their content is still proprietary and, as such, *not* free. The site azlyrics.com (which I believe my IP address is now banned from) offers a front end to the Musixmatch database, which contains lyrics for more than a million songs. One can freely access this database via several APIs with limitations on the number of requests you are allowed to make in a day. You can pay if you want to remove this limitation.

So I had to build a corpus myself. The assignment called for a million words or more, so I set about finding crowd-sourced lists of the most popular bands, songs, and albums of the 1990s. I found many such lists on Ranker.com, and various other music blogs. I also culled Billboard.com for historical lists of the biggest hits for each year in the 1990s, and I scraped Wikipedia’s lists of musicians to find more names. I made cursory attempts at making sure each artist was popular at some point in the 90s, though with 600-800 artists it was a tad messy. For the most part the artists either got their start in the 90s or were around beforehand and remained popular during the 90s. Some artists have also continued since the 90s and have remained popular until even today, e.g. Jay-Z. I did not try to account for whether the lyrics themselves came from the 90s except in the case of the song-based type rankings (and even then I was lenient). If an artist made a splash in the 90s at all, I downloaded their “entire” corpora, i.e. whatever I could scrape. Granted, this resulted in a lot of songs from the 70s, 80s and 2000s, but, very obviously pre-90s artists like The Beatles, Jimi Hendrix, The Doors, Elvis Presley, Led Zeppelin et al are *not* in my corpus. My intention in doing all this was twofold: 1.) Capture the popular music of the era, and 2.) Get enough words to do some real number-crunching.

When I concatenated the artist lists I assembled it was on the order of 12,000 lines in total. From this I chopped it down to unique names, collapsing obvious dupes (like 2pac / Tupac, NWA and N.W.A., Motorhead / Motörhead, Björk / Bjork, Queenryche / Queensrÿche, Rage, Rage Against the Machine, RATM, etc.) and made ranked lists of artists based on their mentions. I then went in order and began scraping various lyrics websites (mostly azlyrics.com and mldb.org) for as many lyrics as were publicly available. I then cross-compared the lyrics available on several sites and came to the conclusion that lyrics websites are mostly copied from one another and no particular site is preferable, other than the fact that there are unwritten caps on the number of requests any anonymous IP address can make to their databases. Thus, for every 300 song lyrics I downloaded I had to get a new IP address. Open VPN came in very handy.

In the end I had a corpus of .txt files which was roughly 71 MB in size. I then ordered the artists whose lyrics were found online by the number of songs they each had. I took the top 300 most prolific artists in each genre, so my parallel corpora was thus a comparison of the available lyrics for 300 artists in two genres, hip-hop/rap (HHR), and rock/metal/country (RMC). In total, stripping for punctuation, removing the top 21 stop words and using simple word/line separation rules, my corpus thus consisted of 9279811 tokens, or roughly 9.3 million words.

The hard part was getting the data. Crunching the numbers was fairly straightforward. I spent more time making these graphs than actually writing the statistics-gathering code. Everything, including the scripts to find popular artists and download their lyrics amounted in all to about 1200 lines of Python.

**Results**

My hypotheses were confirmed – big time. Shakespeare turned out to be a little better than I expected, but the general result is the same: HHR music blows RMC out of the proveribal water in terms of its lexical richness. I will start with a discussion of the Zipf’s law behavior exhibited by the respective HHR and RMC corpora upon examination, and proceed with a more in-depth discussion of the results. Then, I will discuss a new means of representing lexical variety in terms of type/token curves normalized by document length and plotted in logarithmic space. Finally, I will examine in more detail a few high-scoring individual songs from both HHR and RMC artists. But enough talk. Most of the data speaks for itself, if you feel so inclined to just flip through the charts….

Here is the plot of the Zipf’s law behavior, plotted for both corpora:

**Discussion of Zipf’s Law Behavior**

With a large corpora of words we expect the graph of the log of the rank of each word versus the log of its frequency to exhibit the behavior above, with a relatively smooth curve in the center of the graph and this “step behavior” on the extreme ends. This demonstrates the constant inverse proportionality between the rank of a word type and its token frequency. The hip-hop/rap corpus has a very long tail of single-digitons formed from a large vocabulary of slang / invented words and various semantic encodings via what appear to be “misspellings.” In quite a few cases, the spellings are unique to a particular artist.

**Examples:**

Rank 26529 (7-ton): Ice Cube’s spelling of “AmeriKKKa” – This is quite clearly intended to convey a particular defintion of “America.” It cannot be considered the same type as the word “America,” regardless of how much the innate spell-checker inside us all dislikes it.

Rank 92787 (singleton): Outkast’s “southernplayalisticadillacfunkymuzik” is a variation on the name of the 1994 album “Southernplayalisticadillacmuzik.” Again, we see here a wealth of semantic information packed into one word, which can stand on its own, perfectly distinct from the individual sub-words of which it is clearly comprised. It is linguistic inventiveness like this that allows the tail of the blue curve to exhibit such interestingly “smooth” behavior.

Rank 92863 (singleton): “supercalifragilisticexpialidocious” – Mary Poppins’s own creation, used by the same artist in the same song.

Rank 93698 (singleton): “hiphopgesetze” – I cannot claim to know what this means exactly, but it is yet another example of the type of wordplay that typifies hip hop.

The curve goes on like this to rank 102502. From rank 18033 down to 102502, the last 84468 word types are mentioned 10 times or less in the entire corpus. At the other end, there are a lot of words that might be considered by some to be stop words but, for example, in positions 3, 2, 1 are the words “all”, “get”, and “up”, respectively. These are not stop words, because they are crucial to the context of hip hop music, which usually involves dancing, (e.g. a chorus for a famous Afrika Bambaataa hit from the 1980’s is, “Y’*all* just *get* *up* and dance!”) There is a short vocabulary of words at the top of the word frequency list which are important in hip-hop for things like vamping and rhythmic word filler, but they differ from stop words in that they carry a lot of cultural, contextual, and semantic (i.e. “attitude setting”) information. Though it needs no explanation, in position 20, (well within the range of most heuristic stop-word removal windows) is the word “nigga,” which is right around the beginning of the step function behavior. These are clearly features of the Hip-Hop/Rap curve and not just stop words.

The Rock/Metal/Country graph has sharp uptick at the top, in contrast to the Hip-Hop/Rap curve, because rock music has relatively low lexical variety for the vast majority of songs, e.g. in position 5 of the ranking is the word “love,” which accounts for an entire 0.8% of the words in the entire Rock/Metal/Country corpus.

**Discussion of Lexical Variety in Corpora**

This issue of word variety is important because it provides one of the main sources of distinction between the Hip-Hop/Rap HHR corpus and the Rock/Metal/Country (RMC) corpus. The statistics gathered on the genre level pretty much say it all:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Genre** | **Artists** | **Songs** | **Types** | **Tokens** | **MALL** | **Unigram** | **Bigram** | **Trigram** |
| HHR | 300 | 16206 | 102502 | 5430793 | 5.1 | up | dont\_know | la\_la\_la |
| RMC | 300 | 27886 | 70339 | 3849018 | 4.46 | all | dont\_know | la\_la\_la |
| W.S. | 1 | ~194 | 27379 | 652812 | 6.78 | not | thou\_art | exeunt\_all\_but |

In extremely rough terms, the number of songs in the RMC corpus is on its way to being double the number of songs in the HHR corpus. At first, my suspicion was that the RMC artists were that much more prolific than those in HHR and therefore comparing these corpora would not be interesting. But not only are there 32163 (46%) *more* types in the HHR corpus, with only 16206 songs compared to 27886 in RMC, there are also more than 1.5 *million* more words in HHR! The mean average line length (MALL) per verse in both corpora are about the same, and interestingly the top bigram and trigram words in HHR and RMC are identical. Furthermore, it is interesting that the mean average line length hovers around 5. This may be related to the historically common convention of using iambic pentameter in poetry, though without further investigation it would be impossible to say this for sure. It is an interesting idea, however.

Clearly there is something going on in HHR to account for this discrepancy. With so few songs by comparison (and mind you these are only the songs I could scrape from the web), the diversity of words per song in HHR must be far greater than that of RMC. Lexically, hip hop and rap are much more interesting than rock, metal, or country. To visualize this I have created following graph:

This graph shows a comparison of the type/token ratio between HHR and RMC normalized per song on a log scale. Comparing these in normal space is a bit unwieldy due to the outliers, i.e. artists for which lyrics for only one or two songs were found online but which had an extremely high type/token ratio. Furthermore this better demonstrates the separation of the data. On the horizontal axis we see the log of the type/song ratio. Vertically we have the log of the token/song ratio. Data points lying to the upper right of the graph exhibit greater lexical variety on a per song basis. As you can see, almost *all* HHR artists appear to exhibit greater lexical variety than their counterparts in RMC. This makes somewhat intuitive sense given the fact that hip hop and rap are heavily driven by lyrical content as opposed to musicality. (Of course, this is the 90s and we’re talking about bands like Nirvana and Pantera here. “Musicality” is a loose term.)

If we were to simply compare type/token ratios between HHR and RMC we would see a lot of numbers in the 10%-20% range, with a few crucially misleading values. Neil Young, the most prolific rock artist in the RMC list (who oddly showed up in a bunch of 90s rock band lists, for some reason), has a type/token ratio (TTR) of around 9%, which is lower than most. This turns out to be a reasonable number, though it is seemingly contradicted by the fact that William Shakespeare – the most lexically prolific single artist of all time – has a type/token ratio of 4%. (By comparison, Britney Spears’ TTR is actually 9% as well.) If greater TTR is an indication of greater linguistic variety, then this doesn’t seem to make sense. Normalizing for the lexical variety on a per-document basis and putting things on a log scale better clarifies the separation of the data and shows us, as we’d expect, Neil Young in the lower lexical ranks (which doesn’t necessarily mean he’s lexically impoverished, just that he writes a lot of songs with similar words), and William Shakespeare flying high above the rest (Bill is marked as the isolated green dot in the upper right-hand corner).

My argument is that this type of comparison makes a lot of sense in terms of demonstrating the lexical diversity of comparable corpora, and in this case specifically HHR music. Excluding the odd outliers, (which are flukes because we only have one or two songs by each of the floating blue dots high on the right) the data here lays out as expected. Tupac Shakur is easily distinguished by his barrage of lyrical fury, but for all of his lexical variety he also has an extremely large number of common so-called ‘filler’ words. The top 1, 2, and 3 bigrams for Tupac are “2pac ya’ll”, “that’s right”, and “ain’t nothin,” and one of his most common trigrams is “fuck all ya’ll.” So, on a per song basis, Tupac’s token counts are high though his type count is comparatively low. Of course, examining individual songs for their statistically interesting features is also worth a look. The table below is a comparison of the top ten songs in HHR and RMC ordered by type counts.

**Hip Hop / Rap** *…Give Ras-Kass a listen, if you get the chance!*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ARTIST** | **SONG** | **TYPES** | **TOKENS** | **ALL** | **UNIGRAM** | **BIGRAM** | **TRIGRAM** | **TTR** |
| The-Roots | The-Session-Longest-Posse-Cut-In-History | 902 | 1891 | 5.68 | ('have', 73) | ('have\_fun', 50) | ('have\_fun\_have', 40) | 48% |
| Ras-Kass | Nature-Of-The-Threat | 688 | 1052 | 5.88 | ('was', 14) | ('his\_name', 3) | ('around\_2000\_bc', 2) | 65% |
| The-Sugarhill-Gang | Rappers-Delight | 684 | 1987 | 5.78 | ('ya', 50) | ('n\_n', 23) | ('n\_n\_n', 15) | 34% |
| Kanye-West | Last-Call | 681 | 1812 | 7.37 | ('was', 59) | ('jay\_z', 9) | ('was\_gonna\_do', 4) | 38% |
| Busta-Rhymes | Flipmode-Squad-Meets-Def-Squad | 673 | 1023 | 5.98 | ('up', 20) | ('def\_squad', 6) | ('lord\_have\_mercy', 3) | 66% |
| Beastie-Boys | B-Boy-Bouillabaisse | 641 | 1047 | 5.98 | ('as', 12) | ('new\_york', 4) | ('his\_fists\_against', 3) | 61% |
| Blackalicious | Release-Part-12-3 | 612 | 941 | 4.61 | ('all', 14) | ('get\_focus', 4) | ('accelerate\_never\_wait', 4) | 65% |
| Lootpack | Episodes | 603 | 1093 | 6.21 | ('ya', 26) | ('hip\_hop', 7) | ('la\_la\_la', 4) | 55% |
| Wu-Tang-Clan | Triumph | 595 | 775 | 5.46 | ('from', 12) | ('wu\_tang', 5) | ('ol\_dirty\_bastard', 2) | 77% |
| Company-Flow | Collude | 572 | 804 | 6.76 | ('they', 12) | ('will\_fall', 4) | ('timewarner\_will\_fall', 2) | 71% |

**Rock / Metal / Country** *This is why ‘Cradle of Filth’ was marked on the previous graph!*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ARTIST** | **SONG** | **TYPES** | **TOKENS** | **ALL** | **UNIGRAM** | **BIGRAM** | **TRIGRAM** | **TTR** |
| Dream-Theater | Six-Degrees-Of-Inner-Turbulence | 481 | 915 | 4.12 | ('she', 29) | ('he\_was', 7) | ('his\_solitary\_shell', 5) | 53% |
| Cradle-Of-Filth | Beneath-The-Howling-Stars | 413 | 560 | 3.94 | ('her', 22) | ('beneath\_howling', 3) | ('beneath\_howling\_stars', 3) | 74% |
| Neil-Young | Sixty-To-Zero | 403 | 763 | 3.93 | ('he', 31) | ('he\_was', 5) | ('put\_hose\_down', 2) | 53% |
| Bob-Dylan | Brownsville-Girl | 383 | 775 | 8.07 | ('was', 22) | ('brownsville\_girl', 12) | ('world\_brownsville\_girl', 4) | 49% |
| Nirvana | The-Priest-They-Called-Him | 382 | 700 | 8.33 | ('he', 23) | ('get\_out', 3) | ('back\_his\_room', 2) | 55% |
| Savatage | Turns-To-Me | 373 | 706 | 3.92 | ('time', 14) | ('all\_those', 5) | ('all\_those\_moments', 4) | 53% |
| Cradle-Of-Filth | Thirteen-Autumns-And-A-Widow | 358 | 466 | 4.31 | ('her', 22) | ('when\_she', 3) | ('light\_so\_when', 2) | 77% |
| Cradle-Of-Filth | A-Gothic-Romance-Red-Roses-For-The-Devils-Whore | 350 | 431 | 4.63 | ('her', 12) | ('velvet\_enrobed', 2) | ('must\_know\_art', 1) | 81% |
| Mekong-Delta | Dances-Of-Death | 334 | 605 | 3.46 | ('all', 12) | ('days\_betrayal', 7) | ('days\_betrayal\_days', 5) | 55% |
| Lou-Reed | A-Dream | 331 | 651 | 5.05 | ('was', 26) | ('was\_so', 4) | ('was\_very\_cold', 2) | 51% |

**NOTE**: I had to “massage” the data in the RMC table because many artists which may have had hits in the 90s have been around for quite some time, e.g. Bruce Springsteen, Bob Dylan, Rush, etc. Some of the songs in this table are “near” the 90s, admittedly, e.g. Sixty-to-Zero was popular in the 90s even though it came out in 1989. But it is quite interesting to give the songs themselves a listen to see their lexical variety for yourself. Many of the rock songs are extremely long, which likely accounts for their Type/Token counts. I also had to clean up some songs which were too far out of the 90s to be a part of this study, and there were some hip-hop artists in there which I was unable to automatically clean out – after all there are 600 artists in this study and I do not know them all by name!

In the previous chart I marked Cradle of Filth as being an interesting outlier. I suspected a priori that they would be somewhere near the top. Their TTR for 117 songs is 26%, which is a high ratio for an artist with 100+ songs. The metal band ‘Sodom’ appears high in this category as well with a TTR of 29% but they have a much lower type count than Cradle of Filth. Unsurprisingly, Cradle of Filth have *3 songs* in the top ten most lexically rich RMC songs from the 1990s era.

The reader is recommended to give the links in this list a listen. They are all, interestingly, songs which I was not even aware of for the most part prior to doing this study. The most interesting part, at the end of the day, is the fact that these findings more or less confirm what has often written subjectively about these artists by music critics and other pop culture observers.

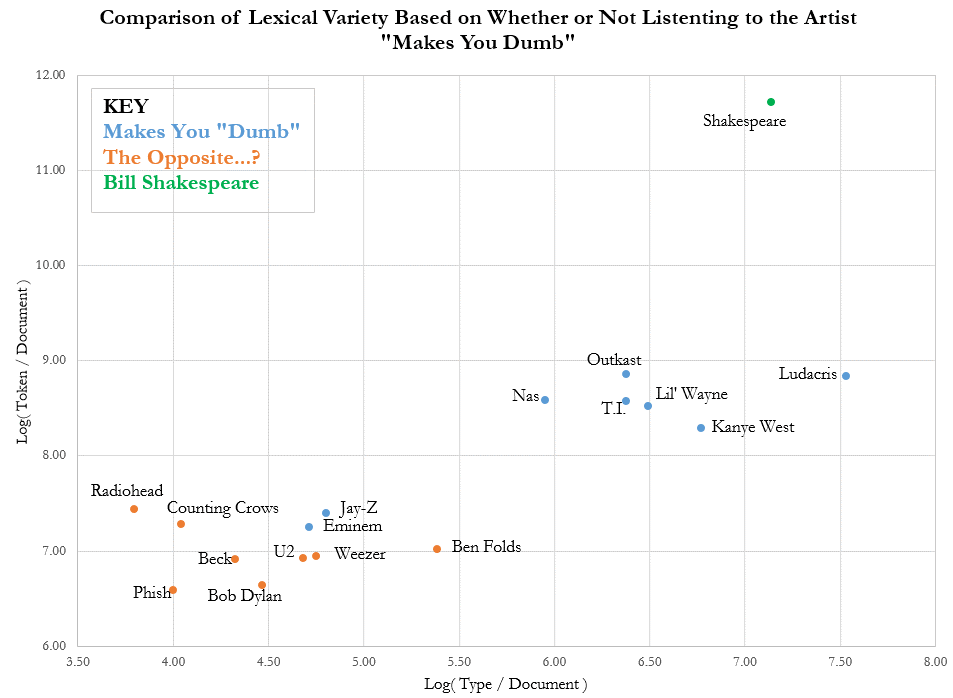
**Conclusions**

I don’t intend to rehash the discussion above, but in the final analysis, the results of this study fly somewhat in the face of the supposed “findings” by Virgil Griffith and his notions of “music that *makes* you dumb.” Given the fact that, again, his chart was picked up and published by such notable publications as The (Pulitzer-prize winning) Guardian, the Wall Street Journal, Yahoo! News, The Washington Post, The Times of India, the New Yorker, The Canberra Times, and the UK Telegraph, it is not quite “just fun” to make such claims. The tagline on the website itself says the following:

**“Yes, I'm aware correlation ≠ causation. The results are hilarity incarnate regardless of causality.   
You can stop sending me email about this distinction. Thanks.”**

Hilarity incarnate? Hardly. Mr. Griffith is promoting racist stereotypes in musical preferences using the aura of statistical inference to gain legitimacy in the minds of uncritical audiences. Such actions can be extremely derisive, offensive, misleading and destructive. Millions of people have likely seen this graph. It’s arguably no coincidence (whether the author acknowledges his implicit bias or not) that the artists with the “dumbest” listeners are all black, and likewise no coincidence that the artists on the opposite end of the spectrum are all white. Empirically, the data may be a manifestation of some underlying sociological phenomenon, but to brazenly make assertions about human intelligence in this format is quite frankly unworthy of someone bearing a @caltech.edu email address.

So, let’s examine some of the artists on opposite sides of Mr. Griffith’s chart. In the Rap / Hip Hop section (the section with the supposed lowest SAT-scoring listeners by far), Mr. Griffith lists Outkast, Eminem, Kanye West, “Rap”, Nas, Akon, Ludacris, Jay-Z, “Hip Hop”, T.I. and Lil Wayne. In the section listing artists whose listeners have apparently higher than 1156 scores on the SAT, we see Norah Jones, U2, Counting Crows, Bob Dylan, Beck, Weezer, The Shins, Radiohead, Ben Folds, Sufjan Stevens, Guster, and Beethoven. Some of these artists have been excluded in my study because they got their start after the year 2000, so I have no data on them. Without further discussion, I conclude with this comparative graph sorted by the logs of the word type / document ratio.



**Appendices**

**Top-20 Artists in HHR sorted by Type Counts**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ARTIST** | **SONGS** | **TYPES** | **TOKENS** | **MALL** | **UNIGRAM** | **BIGRAM** | **TRIGRAM** |
| Tupac-Shakur | 338 | 10338 | 165195 | 5.34 | ('they', 1639) | ('rock\_body', 173) | ('body\_rock\_body', 142) |
| Ll-Cool-J | 208 | 10259 | 89313 | 5.11 | ('up', 873) | ('cool\_j', 642) | ('ll\_cool\_j', 538) |
| Nas | 155 | 9773 | 66155 | 5.72 | ('they', 647) | ('chorus\_nas', 72) | ('bap\_bap\_bap', 48) |
| Wu-Tang-Clan | 117 | 9744 | 54076 | 5.53 | ('up', 443) | ('wu\_tang', 203) | ('ol\_dirty\_bastard', 38) |
| Jay-Z | 210 | 9674 | 92315 | 5.53 | ('this', 805) | ('jay\_z', 616) | ('uh\_huh\_uh', 99) |
| Method-Man | 134 | 9505 | 57547 | 5.37 | ('up', 616) | ('method\_man', 295) | ('yo\_yo\_yo', 47) |
| Eminem | 158 | 9418 | 72362 | 5.8 | ('up', 788) | ('slim\_shady', 96) | ('ha\_ha\_ha', 38) |
| Snoop-Dogg | 250 | 9121 | 104350 | 5.48 | ('up', 1433) | ('snoop\_dogg', 494) | ('d\_o\_double', 93) |
| Ghostface-Killah | 120 | 8646 | 47355 | 5.87 | ('up', 520) | ('ghostface\_killah', 234) | ('ghostface\_killah\_yo', 34) |
| Cypress-Hill | 176 | 8506 | 60747 | 5.33 | ('up', 745) | ('b\_real', 397) | ('la\_la\_la', 84) |
| Lil-Wayne | 190 | 8383 | 73905 | 5.52 | ('got', 774) | ('lil\_wayne', 350) | ('wee\_ooh\_wee', 72) |
| The-Roots | 100 | 8354 | 40537 | 5.42 | ('get', 403) | ('black\_thought', 204) | ('zen\_zen\_zen', 74) |
| Busta-Rhymes | 142 | 8086 | 61255 | 5.45 | ('shit', 801) | ('busta\_rhymes', 383) | ('yo\_yo\_yo', 78) |
| Ice-Cube | 170 | 7967 | 65056 | 5.17 | ('get', 747) | ('ice\_cube', 557) | ('chorus\_ice\_cube', 49) |
| Bone-Thugs-N-Harmony | 179 | 7878 | 94286 | 6.6 | ('up', 1436) | ('krayzie\_bone', 207) | ('bone\_bone\_bone', 101) |
| Public-Enemy | 198 | 7813 | 61380 | 4.27 | ('they', 715) | ('chuck\_d', 126) | ('give\_up\_give', 47) |
| Twista | 111 | 7281 | 55350 | 5.77 | ('up', 817) | ('w\_w', 96) | ('w\_w\_w', 80) |
| Gang-Starr | 119 | 7174 | 42777 | 5.83 | ('so', 482) | ('dj\_premier', 71) | ('dj\_premier\_cuts', 35) |
| South-Park-Mexican | 102 | 7094 | 37841 | 5.09 | ('got', 364) | ('wiggy\_wiggy', 136) | ('wiggy\_wiggy\_wiggy', 108) |
| Dmx | 165 | 7040 | 70203 | 5.64 | ('what', 1020) | ('dont\_know', 111) | ('uh\_uh\_uh', 47) |

**Top-20 Artists in RMC sorted by Type Counts**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ARTIST** | **SONGS** | **TYPES** | **TOKENS** | **MALL** | **UNIGRAM** | **BIGRAM** | **TRIGRAM** |
| Celine-Dion | 325 | 7217 | 78332 | 4.39 | ('love', 1017) | ('la\_la', 179) | ('la\_la\_la', 133) |
| Bob-Dylan | 340 | 7208 | 65049 | 5.36 | ('he', 651) | ('no\_more', 74) | ('when\_gonna\_wake', 27) |
| Bruce-Springsteen | 378 | 6633 | 68148 | 5.64 | ('just', 619) | ('la\_la', 120) | ('la\_la\_la', 102) |
| Elvis-Costello | 330 | 6203 | 45155 | 4.9 | ('but', 513) | ('dont\_know', 84) | ('la\_la\_la', 28) |
| Rush | 187 | 6008 | 29707 | 4.1 | ('all', 277) | ('brollic\_brollic', 32) | ('what\_what\_what', 20) |
| Cradle-Of-Filth | 117 | 5683 | 22207 | 3.85 | ('her', 476) | ('nymphetamine\_nymphetamine', 26) | ('no\_no\_no', 16) |
| Bad-Religion | 212 | 5224 | 24893 | 5.19 | ('all', 301) | ('no\_one', 40) | ('come\_join\_us', 24) |
| Jethro-Tull | 212 | 5098 | 24894 | 4.99 | ('no', 206) | ('watching\_watching', 27) | ('step\_no\_step', 17) |
| Rem | 241 | 5004 | 32165 | 4.65 | ("it's", 348) | ('la\_la', 96) | ('la\_la\_la', 86) |
| 311 | 113 | 4903 | 24264 | 4.83 | ('but', 250) | ('throw\_down', 49) | ('time\_throw\_down', 43) |
| Live | 162 | 4781 | 33569 | 4.65 | ('all', 419) | ('brother\_marquis', 129) | ('fresh\_kid\_ice', 97) |
| Rage Against The Machine | 208 | 4709 | 30580 | 4.46 | ('they', 300) | ('p\_wagner', 132) | ('by\_p\_wagner', 112) |
| Neil-Young | 458 | 4659 | 54417 | 3.26 | ('love', 652) | ('he\_was', 76) | ('wanna\_wanna\_wanna', 44) |
| Lou-Reed | 196 | 4606 | 35946 | 4.63 | ('no', 419) | ('na\_na', 354) | ('na\_na\_na', 338) |
| Tori-Amos | 231 | 4514 | 30048 | 3.59 | ('this', 292) | ('dont\_know', 39) | ('just\_two\_us', 21) |
| Rod-Stewart | 274 | 4470 | 41590 | 4.54 | ('love', 610) | ('long\_time', 57) | ('when\_mans\_love', 36) |
| Marillion | 129 | 4462 | 21893 | 4.53 | ('all', 210) | ('marillion\_lyrics', 52) | ('music\_marillion\_lyrics', 52) |
| Red-Hot-Chili-Peppers | 165 | 4339 | 24936 | 3.79 | ('all', 256) | ("i've\_got", 48) | ('magik\_sex\_magik', 32) |
| Sting | 170 | 4269 | 27006 | 4.75 | ('all', 285) | ('one\_day', 40) | ('only\_only\_only', 27) |
| Barenaked-Ladies | 164 | 4135 | 26156 | 4.61 | ("it's", 301) | ('la\_la', 88) | ('la\_la\_la', 80) |

**Top-20 Artists in HHR found from web scraping**

|  |  |
| --- | --- |
| **ARTIST** | **FREQUENCY** |
| Dr. Dre | 22 |
| Arrested Development | 19 |
| Naughty By Nature | 18 |
| A Tribe Called Quest | 16 |
| Ice Cube | 16 |
| Bone Thugs-N-Harmony | 15 |
| Public Enemy | 15 |
| The Notorious B.I.G. | 15 |
| Gang Starr | 14 |
| Cypress Hill | 14 |
| Ll Cool J | 14 |
| Digital Underground | 13 |
| Tupac Shakur | 13 |
| Ice Cube Feat. Das Efx | 13 |
| Snoop Dogg | 13 |
| Redman | 13 |
| Salt-N-Pepa | 12 |
| Nas | 12 |
| Epmd | 12 |
| Geto Boys | 12 |

**Top-20 Artists in RMC found by web scraping**

|  |  |
| --- | --- |
| **ARTIST** | **FREQUENCY** |
| Aerosmith | 62 |
| Nirvana | 61 |
| Depeche Mode | 61 |
| The Cure | 60 |
| Rush | 59 |
| The Black Crowes | 59 |
| Red Hot Chili Peppers | 57 |
| U2 | 57 |
| Concrete Blonde | 57 |
| R.E.M. | 56 |
| Eric Clapton | 56 |
| Faith No More | 56 |
| Damn Yankees | 56 |
| Garth Brooks | 55 |
| The Jesus & Mary Chain | 55 |
| World Party | 55 |
| Van Halen | 55 |
| The Church | 55 |
| The Psychedelic Furs | 55 |
| Peter Murphy | 55 |

**Top 20 Unigrams, Bigrams, and Trigrams in HHR**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **UNIGRAM** | **FREQUENCY** | **BIGRAM** | **FREQUENCY2** | **TRIGRAM** | **FREQUENCY3** |
| up | 51659 | dont\_know | 4413 | la\_la\_la | 1666 |
| get | 48993 | aint\_no | 3810 | oh\_oh\_oh | 1514 |
| all | 43110 | know\_what | 3802 | yeah\_yeah\_yeah | 1493 |
| so | 42625 | yeah\_yeah | 3472 | no\_no\_no | 1267 |
| dont | 42311 | oh\_oh | 2539 | na\_na\_na | 1241 |
| got | 41767 | no\_more | 2417 | dont\_give\_fuck | 995 |
| this | 40278 | no\_no | 2257 | ha\_ha\_ha | 974 |
| but | 39418 | uh\_huh | 2253 | yo\_yo\_yo | 866 |
| know | 38622 | la\_la | 2250 | da\_da\_da | 762 |
| what | 37361 | dont\_wanna | 2204 | uh\_uh\_uh | 740 |
| no | 35805 | verse\_2 | 2034 | do\_do\_do | 721 |
| it's | 34184 | dont\_want | 2031 | dont\_know\_what | 706 |
| they | 33748 | yo\_yo | 2026 | ll\_cool\_j | 599 |
| when | 32995 | verse\_1 | 2008 | aint\_got\_no | 550 |
| out | 32211 | ha\_ha | 1977 | as\_long\_as | 541 |
| now | 30182 | it's\_all | 1924 | hey\_hey\_hey | 472 |
| do | 29661 | know\_how | 1859 | what\_what\_what | 458 |
| just | 29335 | what\_do | 1826 | it's\_all\_about | 456 |
| nigga | 28500 | this\_shit | 1818 | know\_what\_sayin | 434 |
| if | 27644 | dont\_stop | 1795 | bang\_bang\_bang | 422 |

**Top 20 Unigrams, Bigrams, and Trigrams in RMC**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **UNIGRAM2** | **FREQUENCY3** | **BIGRAM** | **FREQUENCY2** | **TRIGRAM** | **FREQUENCY32** |
| all | 42289 | dont\_know | 4605 | la\_la\_la | 1878 |
| dont | 33118 | no\_one | 2921 | na\_na\_na | 1839 |
| but | 31537 | i've\_been | 2902 | oh\_oh\_oh | 1221 |
| so | 30592 | i've\_got | 2542 | yeah\_yeah\_yeah | 1192 |
| love | 30160 | oh\_oh | 2420 | no\_no\_no | 926 |
| it's | 30152 | la\_la | 2379 | dont\_know\_what | 842 |
| know | 27837 | know\_what | 2364 | do\_do\_do | 561 |
| just | 27787 | dont\_want | 2264 | dont\_know\_why | 521 |
| no | 27669 | yeah\_yeah | 2250 | there\_aint\_no | 500 |
| this | 26147 | there's\_no | 2130 | as\_long\_as | 498 |
| when | 25291 | na\_na | 2122 | hey\_hey\_hey | 485 |
| what | 24971 | let\_go | 1861 | dont\_know\_how | 432 |
| up | 22233 | it's\_all | 1844 | doo\_doo\_doo | 376 |
| can | 21729 | no\_more | 1786 | love\_love\_love | 373 |
| out | 21347 | you've\_got | 1762 | ah\_ah\_ah | 357 |
| now | 21185 | no\_no | 1734 | ha\_ha\_ha | 350 |
| down | 21124 | can\_see | 1716 | come\_come\_come | 333 |
| do | 20656 | what\_do | 1667 | all\_night\_long | 329 |
| got | 20196 | it's\_not | 1655 | it's\_too\_late | 322 |
| if | 19942 | dont\_wanna | 1588 | dont\_know\_where | 317 |

**Top 20 Unigrams, Bigrams, and Trigrams in Shakespeare**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| UNIGRAM | FREQUENCY | BIGRAM | FREQUENCY | TRIGRAM | FREQUENCY |
| not | 8501 | thou\_art | 538 | exeunt\_all\_but | 91 |
| his | 6853 | will\_not | 532 | exeunt\_scene\_ii | 83 |
| this | 6585 | do\_not | 526 | exeunt\_scene\_iii | 69 |
| but | 6266 | no\_more | 510 | know\_not\_what | 54 |
| he | 6250 | king\_henry | 386 | act\_v\_scene | 50 |
| have | 5881 | thou\_hast | 369 | as\_well\_as | 49 |
| as | 5733 | he\_hath | 349 | exeunt\_scene\_iv | 47 |
| thou | 5478 | exeunt\_scene | 346 | do\_not\_know | 47 |
| him | 5194 | they\_are | 339 | as\_much\_as | 46 |
| so | 5035 | if\_thou | 334 | would\_not\_have | 44 |
| will | 4968 | more\_than | 324 | act\_iv\_scene | 44 |
| what | 4456 | not\_so | 313 | act\_iii\_scene | 43 |
| thy | 4032 | how\_now | 294 | act\_ii\_scene | 42 |
| all | 3912 | would\_not | 285 | what\_art\_thou | 41 |
| her | 3847 | king\_richard | 272 | no\_more\_than | 40 |
| no | 3771 | i'\_th' | 269 | as\_thou\_art | 39 |
| by | 3757 | would\_have | 261 | fare\_thee\_well | 39 |
| do | 3747 | if\_he | 258 | thou\_canst\_not | 39 |
| shall | 3583 | let\_us | 256 | why\_dost\_thou | 39 |
| if | 3488 | re\_enter | 251 | exeunt\_act\_v | 38 |

"