

A Classifier for Determining an Artist's Musical Genre Based on Lyrical Content

Josh Hug
CS 294

Introduction

Identifying a musical artist's genre based on lyrical content is an unstudied field. This is unsurprising, of course, as an algorithm which could perfectly perform this task would still be of small practical use. However, I found it interesting from an epistemological point of view to determine how accurately one can extract an artist's genre from the verbal content of their work.

Methodology

Construction of the Data Set

Nineteen musical genres were chosen, shown in Table 1. Genres were chosen arbitrarily, with the restriction that a list of artists from that genre was available on Wikipedia. A list of artists for each genre was then compiled from Wikipedia, for a total of 5689 artists. Next, a list of artists whose lyrics are available on azlyrics.com was compiled, for a total of 2266 artists. The intersection of these lists was then found to form a list of artists for whom we had lyrics and a genre label. In the end, the intersection contained a total of 736 artists, forming our final list of artists.

Alternative Country	Christian Ska	Gothic Rock	Pop Punk	R&B
Alternative Rock	Country	Hip-hop Groups	Progressive Rock	Solo Rap Artists
Bluegrass	Death Metal	Indie Rock	First Wave Punk	Ska
Blues	Gothic Metal	Industrial	Second Wave Punk	

Table 1: List of Genres

All song lyrics for those 736 artists were crawled from azlyrics.com, for a total of 22,531 songs, and the lyrics were converted to a bag of words representation. This feature set was chosen because it seems obvious that certain words should appear with much greater frequency in songs of different genre due to differences of subject matter, cultural and artistic conventions, etc.

The first attempt to count the words, using a linear search on a dynamic list sorted by word frequency, was too slow to compile the words in a reasonable amount of time, as the linear search got very slow as the word list reached the tens of thousands. The second attempt, a binary search on a list sorted by lexicographic word ordering, was also too slow to compile the word count in a reasonable amount of time, as every additional word required a resorting of the data, and the vector data structure in MATLAB, the development environment being used, did not allow for an easy way to insert a new word in the middle of an alphabetic list, and thus a large number of resortings of lists tens of thousand words

long was eventually required. Thus, a static list of words was built by performing the binary search method on a stratified sample of the artists. The list was then truncated at the top 15,000 words.

This static list of words was then used to convert all of the songs into a 15000 dimensional vector representing the word count. Songs were then aggregated by artist to form a final set of 736 15,000 dimensional vectors to be used for training and testing.

Building the Classifier

To decide on a classifier to use in the multiclass case, several binary classifiers were evaluated on hip-hop and indie rock music. The 118 hip-hop artists and 62 indie rock artists were separated into a training and validation set, each composed of 59 hip-hop artists and 31 indie rock artists. Four different classifier types were tried on two different types of word counts on two different sets of words. Classifier types and types of word counts are shown below in Table 2.

Classifier Type	Word Type	Count Type
K Nearest Neighbor	Top N Words Overall	Integer
Perceptron	Top N Words from Relevant Genres	Binary
SVM		

Table 2: Classifier Types Evaluated

Classifiers were judged based on their accuracy and the stability of that accuracy with respect to parameters in a local region [see Appendix A]. Ultimately, the perceptron classifier on the top N words with integral count was chosen with N=2000. This classifier had 90% accuracy when comparing indie rock and hip-hop artists, and performance benchmarking showed relative insensitivity to word count beyond 2000 words. It is somewhat surprising that the perceptron won out over the SVM, though this was probably because I only tried an off-the-shelf implementation with a linear kernel. In any case, the accuracy of the perceptron was promising enough to warrant its use without delving too deeply into building a better SVM.

A multiclass classifier was then built using the all-vs-all technique with majority voting to resolve binary decisions into a final multiclass decision. Because of a relative sparsity of data in the other genres, perceptron classifiers were built only for alternative rock, country, indie rock, hip hop groups, solo rap artists, and R&B. Artists from other genres were kept for validation.

Notes on the Data Collected

Both of our datasets are from publicly curated databases, and as such, are likely to contain many errors. A brief glance at the less frequent words shows that the word count includes words such as “celular” and “paralel”. In the case where the word count is integral, it does not seem that these would have much bearing as they would be drowned out by the other words, but if we had used a binary counter this might have been significant. In that case, however, the typos might have told us some information about the words which a band’s most enthusiastic listeners are likely to be able to spell, so typos might be an interesting feature. Further notes on the data collected can be found in Appendix B.

Multiclass Classifier Testing and Results

The 736 artists were divided into a training and validation set such that each of the six genres of interest had roughly half of its artists in each set. The binary perceptron classifiers were then trained on these 355 artists. The binary classifier was then applied to the remaining 381 artists, and majority voting was used to decide on a multiclass classification, using random tiebreaking. A confusion matrix was then constructed. In the case that an artist belonged to multiple genres, if the multiclass classification was correct, one of the applicable diagonal entries was incremented randomly, and if the multiclass classification was incorrect, one of the applicable nondiagonal entries was incremented randomly. Graphical and textual versions of the confusion matrix are shown below in Figures 1 and 2. Note that some columns are zero. These are genres for which there was no intersection between Wikipedia and azlyrics.com.

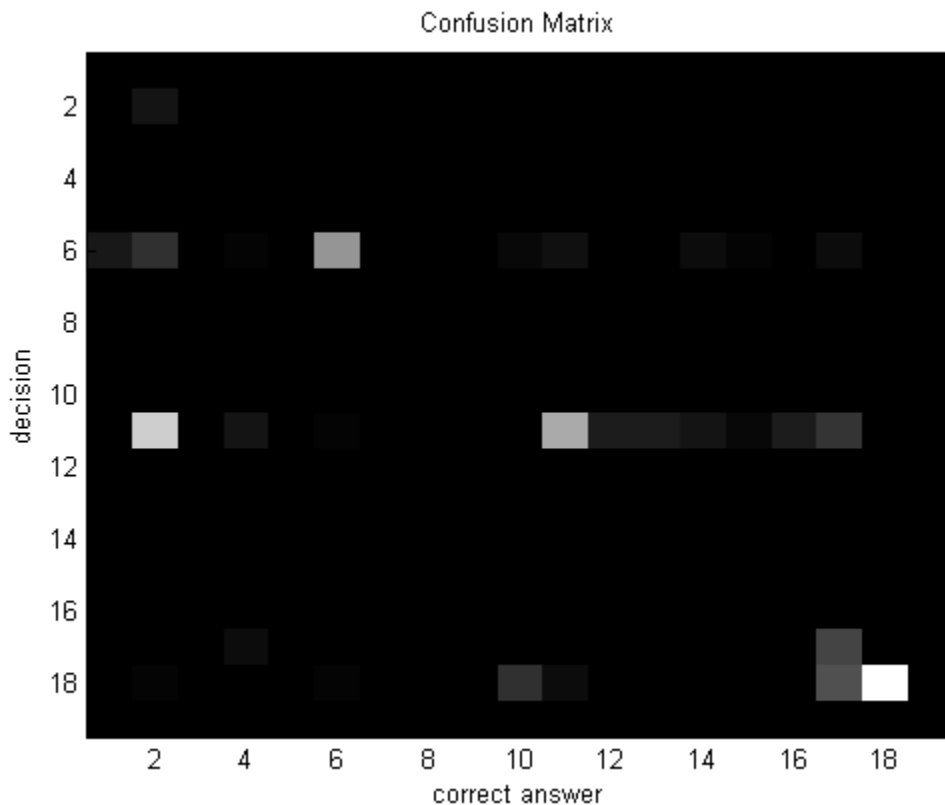


Figure 1: Graphical Depiction of the Confusion Matrix

Confusion Matrix

	Alt Country	Alt Rock	Bluegrass	Blues	Xtian Ska	Country	Deathmetal	Gothmetal	Gothrock	HiphopGrp	Indie Rock	Industrial	Pop Punk	Prog Rock	Wave1Punk	Wave2Punk	R&B	Rappers	Ska
Alt Country	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Alt Rock	0	6	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0
Bluegrass	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Blues	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Xtian Ska	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Country	7	13	0	2	0	39	0	0	0	3	5	0	1	4	2	1	4	0	0
Deathmetal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gothmetal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gothrock	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HiphopGrp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Indie Rock	0	54	0	6	0	2	0	0	0	1	44	8	8	6	3	8	14	0	1
Industrial	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pop Punk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Prog Rock	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Wave1Punk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Wave2Punk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R&B	0	0	0	4	0	0	0	0	0	1	1	0	0	0	0	0	18	0	0
Rappers	0	2	0	0	0	2	0	0	0	13	4	1	0	0	1	0	21	67	0
Ska	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

correct answer

Figure 2: Textual Depiction of the Confusion Matrix

Overall, out of 381 artists in the validation set, 315 belonged to one of the six genres that the predictors covered. Out of these, 174/315 or 55% were correctly predicted. A large majority of the errors occurred between genres generally regarded as similar, e.g. alternative rock and indie rock. For whatever reason, it would appear that the alternative rock classifier was largely overshadowed by the indie rock classifier, and the hip-hop groups classifier was completely eclipsed by the rapper classifier. Though these relationships are reasonable given the similarity of the respective genres, the cause of the outright domination by one genre out of each pair is unknown. A modified voting scheme such as pairwise coupling may have resolved this problem, as the dominated genres did tend to win the majority of their pairwise contests. Out of the four remaining possible decisions, three genres, indie rock, solo rap artists, and country, were all quite distinguishable. R&B did not fare as well, with artists often being misjudged as solo rap artists or slightly less frequently, indie rock artists, but R&B still scored a respectable 18/57.

Conclusion

Though my classification tool is far from mature, it seems that a more focused classifier of greater sophistication would be capable, at the very least, of separating artists into the fairly large bins of rock, country, and hip-hop with very high accuracy based on lyrical content.

Because less common musical genres such as early punk, black metal, death metal, etc. were not represented well on azlyrics.com, it would have been necessary to build more webcrawler/parsers to extract lyrics from specialty lyrics sites such as darklyrics.com to cover the selected genres. Thus, it remains impossible to guess whether sufficient information exists inside song lyrics to build a classifier to accurately resolve all of my genres.

Future Work

The change which would probably make the greatest impact on the six genre classifier would be the implementation of a better scheme for making a multiclass decision from the binary decisions. As mentioned above in the results, the genres which were dominated tended to win their pairwise contests a majority of the time, but were simply outvoted by other classifiers who won their pairwise contests more frequently. Pairwise coupling or an error correcting code inspired approach might have allowed the dominated genres to win the overall vote more frequently. Also important would be an analysis of the binary performance for the other genres. Though our classifier worked very well on indie rock vs. hop-hop, it might not do so well in other settings, and it might be necessary to do more work to get a better binary classifier. The addition of lyrics from other web pages in order to attain coverage of all genres would also be an obvious first step. It might also be interesting to reduce the granularity by moving from artist genre prediction to song genre prediction.

Appendix A – Choosing a Binary Classifier

Several types of classifiers were tested on the data, as described in Table 2. The word type refers to which lists were used to train the classifiers. Top N Words Overall means that only the most common N words across all songs in the database were counted, whereas Top N Words from Relevant Genres means that only the most common words across the two genres being compared were counted.

In all cases, only a small search was done over the parameter space. For the perceptron, an alpha over the range 0.2 to 0.8 had little impact on the results. For the SVM, only a linear kernel was tried.

Classifier Type	Word Type	Count Type
K Nearest neighbor	Top N Words Overall	Integer
Perceptron	Top N Words from Relevant Genres	Binary
SVM		

Table 2: Classifier Types Evaluated

Examples of accuracy scores are shown below vs. number of words:

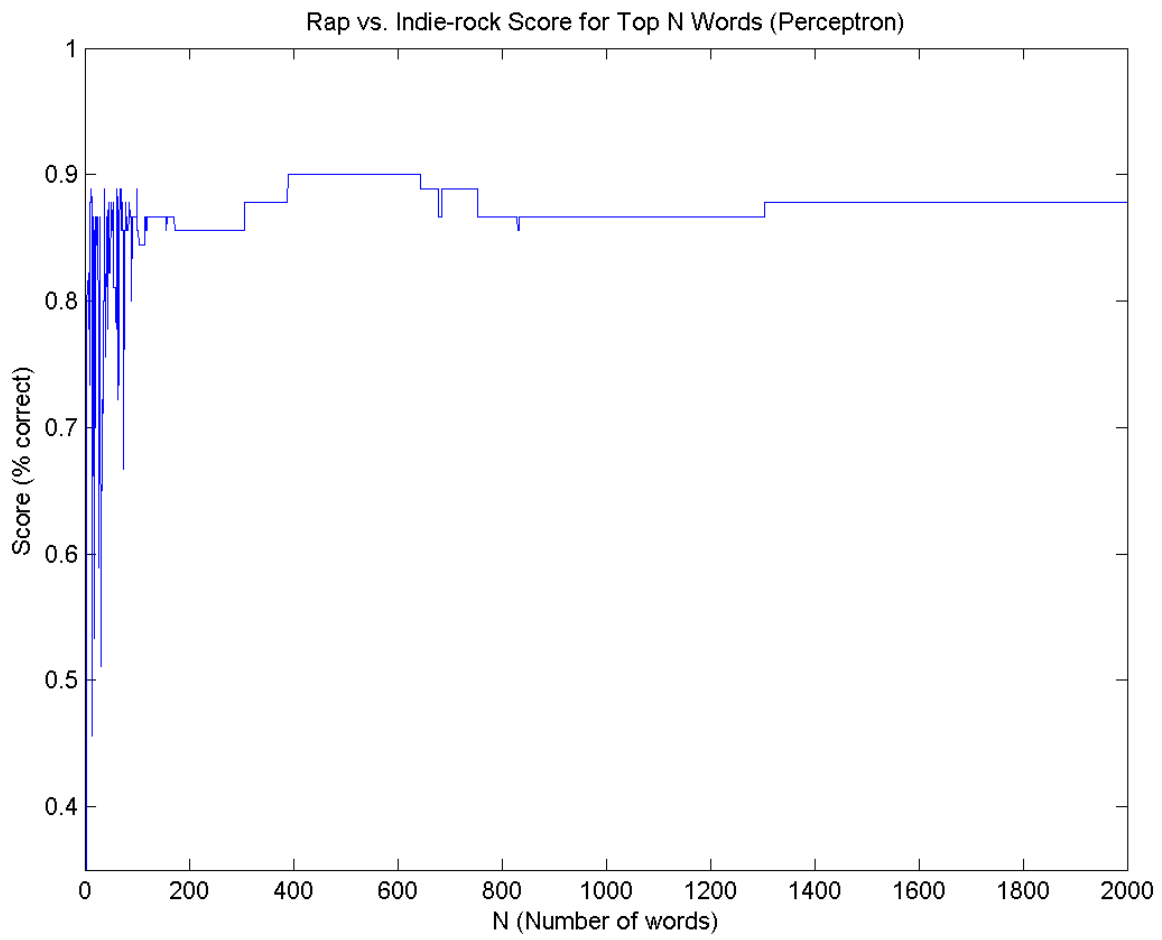


Figure A1: The Winning Classifier

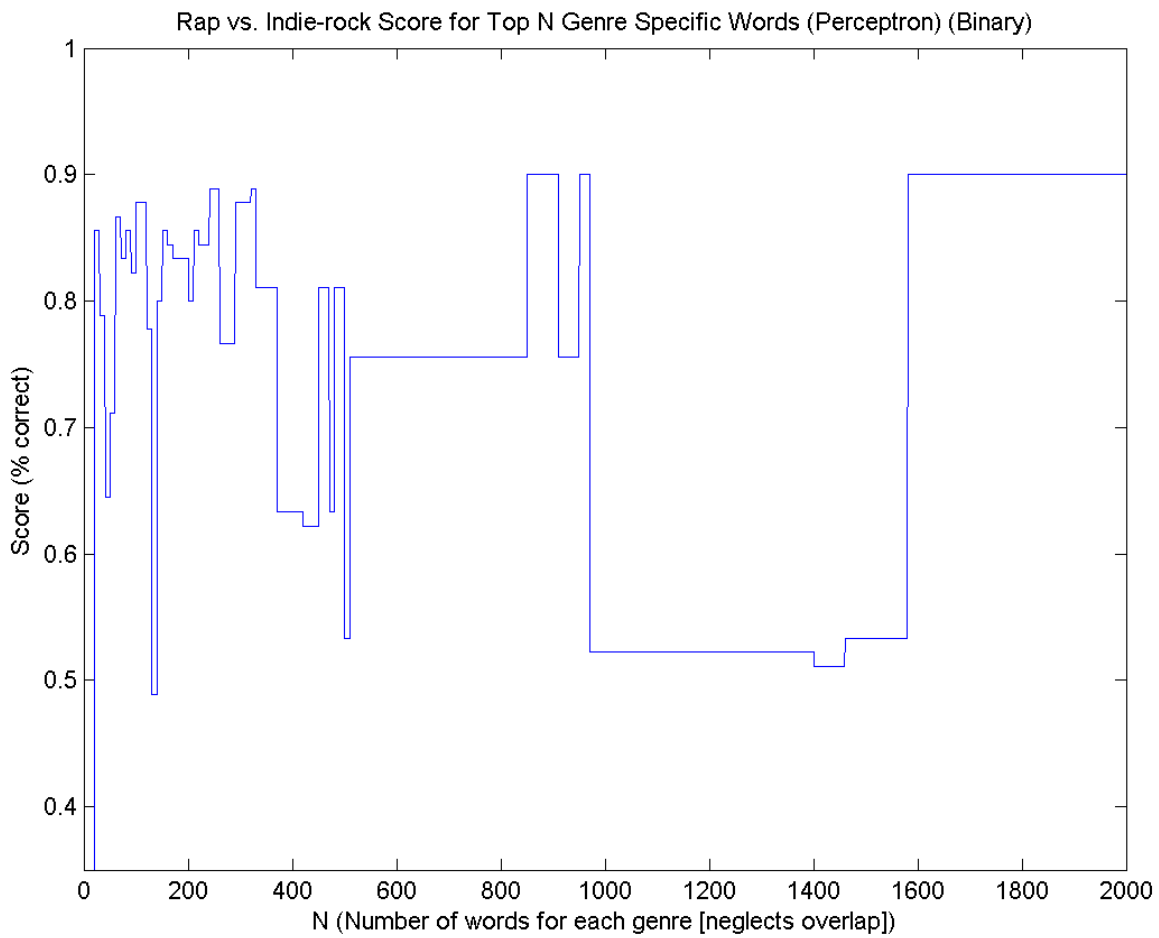


Figure A2: Same Classifier as A1, except with a Binary Count of Genre Specific Words (the large dip is why I chose N=2000 as opposed to a smaller number for the winning classifier)

Appendix B – PCA of the Data

Below in Figures B1 and B2 are shown 2D PCA projections of the data. Indie rock is buried in a small subset intermingled with everything else, while rappers and alt rock are clustered in one area, and R&B and hip-hop groups occupy their own shared distinct space to the top left. The 3D projection is a single orthant oriented towards the viewer. Now, country and rap are separable.

It is interesting that my confusion matrix (Figure 2) matches my intuitions on at least one count for which the 3D PCA projection does not. Specifically, the 3D PCA implies that hip-hop groups and R&B are hard to separate, whereas I'd suspect hip-hop groups would be more closely aligned with solo rap artists in terms of lyrical content.

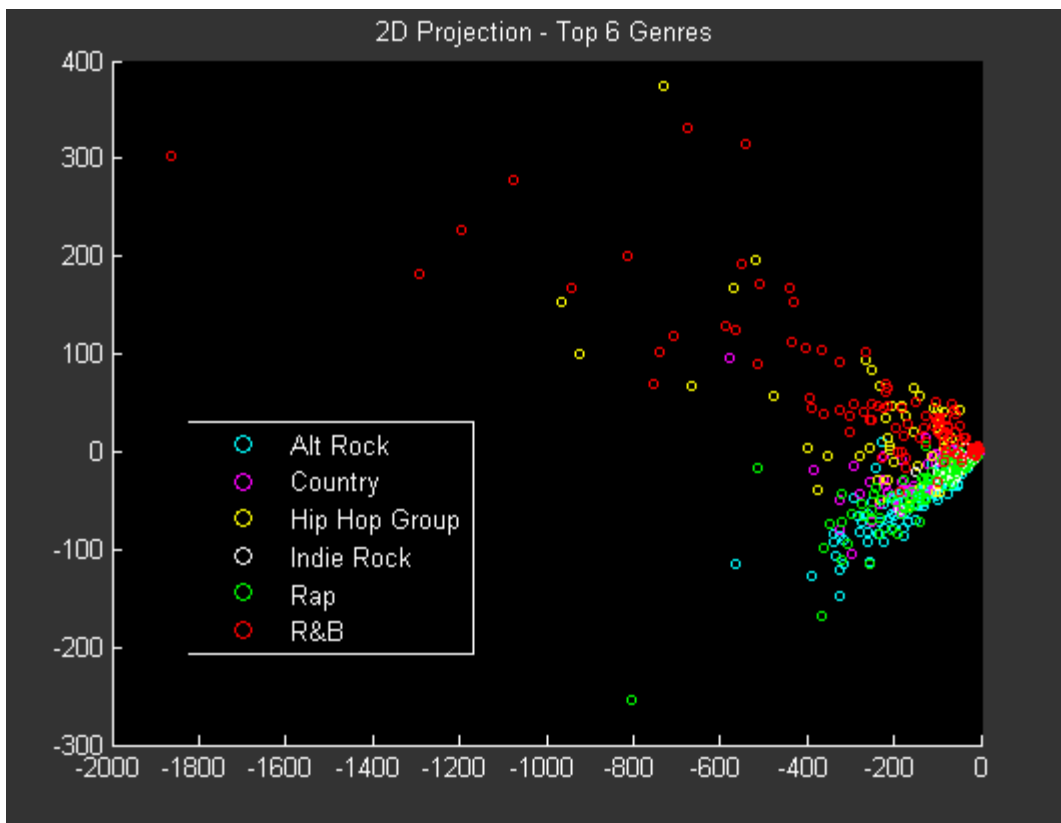


Figure B1: 2D PCA Projection of Word Count on Top 2000 Words

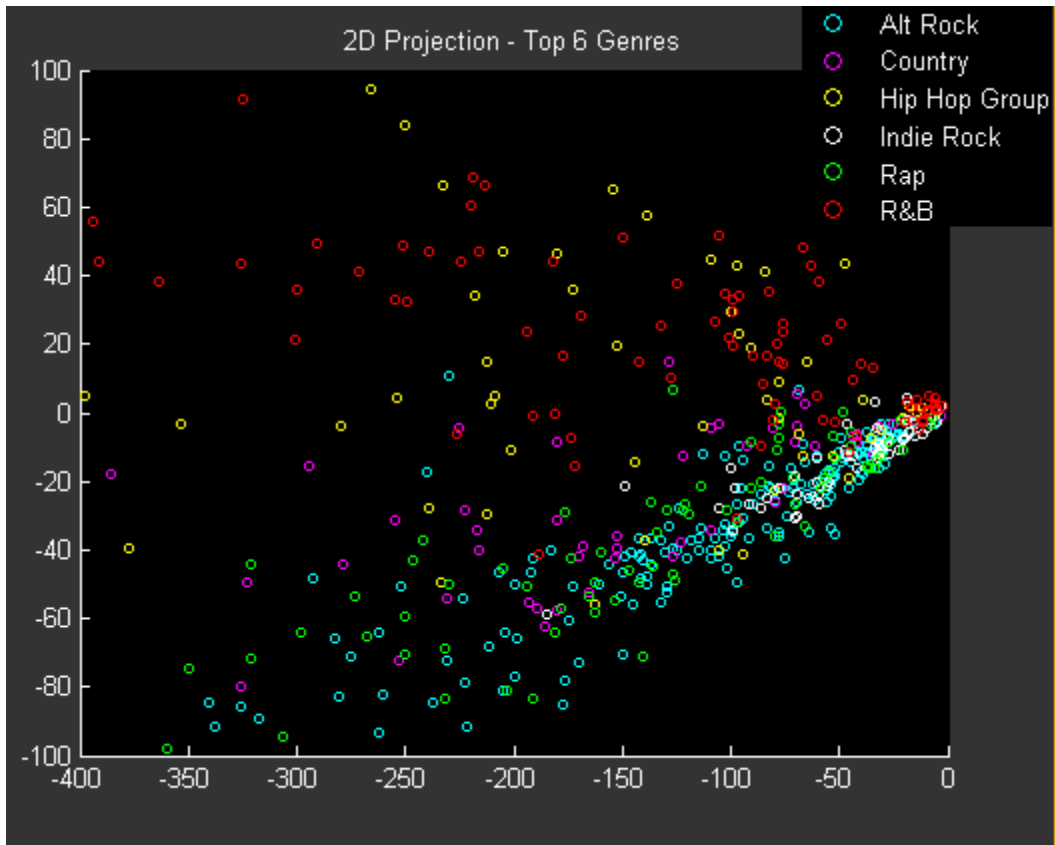


Figure B2: Closer View of PCA Projection of Word Count on Top 2000 Words

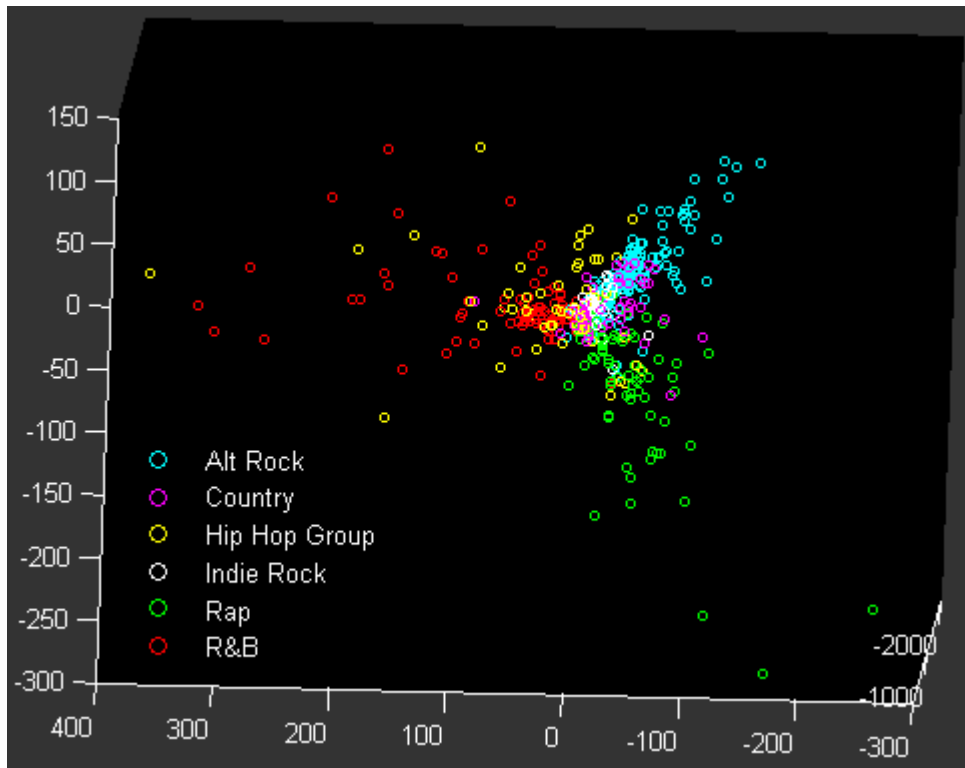


Figure B3: A 3D Projection of Word Count on the Top 2000 Words [orthant facing viewer]