**Song-Genre Classification using logistic regressions**

|  |  |
| --- | --- |
| Ziv Levy  The Interdisciplinary Center, Herzliya  Levy.ziv@post.idc.ac.il | Shachar Langbeheim  The Interdisciplinary Center, Herzliya  Shachar.Langbeheim@post.idc.ac.il |

**Abstract**

We’ve attempted to try to classify songs in their respective genres using logistic regressions over their lyrics and additional computations done over their lyrics. We met with middling success which shows that this system could potentially lead to clear classifications, if iterated upon.

**1. Introduction**

Splitting different songs into different genres has been a problem for a while now – genres are not defined according to any clear or exact criteria, and there’s a significant amount of potential parameters – what instruments are used, how fast is the beat, etc., not to mention the fact it’s not clear what exactly should be assigned a genre – can an artist be classified to a genre? Is each track in an album in its own genre or does an album contain a cohesive genre?

We decided to check whether a song’s genre can be estimated from its lyrics and their structure. This comes from presuming that amongst other things, genres are defined by their subject matter [6], and so songs with similar subjects should share a similar lexical space.

We believe that if we could predict, with significant confidence, to what genre a song belongs, we’ll be better able to define the genres of outliers – songs which are on the edge between different genres, but more than that, it’ll set a cornerstone for an objective definitions of genres.

For this experiment we’ve chosen five genres – Pop, Rock, Metal, Rap & Blues, because they bear some musical resemblance between each other, and also some are the historical precursors of the others, and yet are sufficiently distinct from one another that the average listener could differentiate between the mainstream songs of each genre.

**2. Background**

**3. Related Work(?)**

**4. Logistic Regression**

In statistics, logistic regression is a type of regression analysis used for predicting the outcome of a categorical (a variable that can take on a limited number of categories) dependent variable based on one or more predictor variables. The probabilities describing the possible outcome of a single trial are modeled, as a function of explanatory variables, using a logistic function. Logistic regression measures the relationship between a categorical dependent variable and usually a continuous independent variable (or several), by converting the dependent variable to probability scores. We will use Logistic Regressionto predict if the genre of a given song. In addition to it, we will provide confidence percentage for that decision.

The logistic regression methodology is based on a training set. A training set is a set of pre-defined cases of instances that have already been scored. In our case, we will define a set, which include a subsets of genres such that each genre will contain a group of songs belong to the certain genre.

We will use [1] as our guidelines to build a proper logistic regression function.

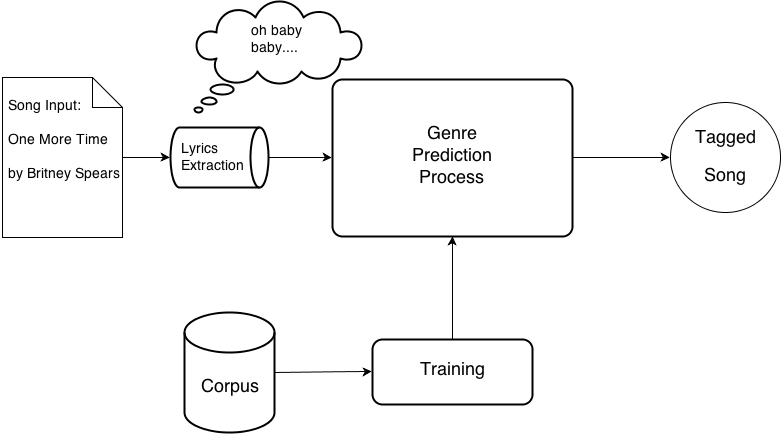
We will define a presence vector in which all members of that vector are the number of occurrences of each word in the song. If a certain word exists in the given inspected song we will mark the number occurrences in its index, 0 otherwise.

Let Y be the result of the algorithm, if the song is considered as belong to the genre n, then Y=n. Moreover, we want to compute the probability that Y=n with a given set of seen words X. This is giving us a probability function:  
*𝑃(𝑌=1|𝑋)*

Moreover, as part of the algorithm result, we also have the confidence level of the result, which means, we have the ability to present how accurate and sure the result of the algorithm is. This is also a great feature that might be used in a later GUI implementation of this mechanism, and of course, for further calculations and statistical measurements.

**4.1. High Level Architecture**

In the following figure the system architecture can be overviewed in the bird’s eye.



Each examined song is being processed in this following pseudo-code:

1. Extract the song’s lyrics while ignoring predefined words.
2. The words are given as input to an already trained model
3. The model calculates its genre prediction regarding the given song.
4. We get a result of the final genre prediction and a confidence level of its prediction.

**4.2. Fields and examples of applications**

Logistic regression is used extensively in numerous disciplines, including the medical and social science fields. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed by Boyd et al. using logistic regression [3]. Logistic regression might be used to predict whether a patient has a given disease (e.g. diabetes), based on observed characteristics of the patient (age, gender, body mass index, results of various blood tests, etc.). Another example might be to predict whether an American voter will vote Democratic or Republican, based on age, income, gender, race, state of residence, votes in previous elections, etc. [4]. The technique can also be used in engineering, especially for predicting the probability of failure of a given process, system or product [5]. It is also used in marketing applications such as prediction of a customer's propensity to purchase a product or cease a subscription, etc. In economics it can be used to predict the likelihood of a person's choosing to be in the labor force, and a business application would be to predict the likehood of a homeowner defaulting on a mortgage.

In each of these instances, a logistic regression model would compute the relevant odds for each predictor or interaction term, take the natural logarithm of the odds (compute the logit), perform a linear regression on the predicted values of the logit, then take the exponential function of the logit to compute the odds ratio. Conditional random fields, an extension of logistic regression to sequential data, are used in natural language processing – and this is exactly our case!

**5. Results**

We’ve run the program over several configurations, and reached somewhat surprising results. Our baseline check was based on a simple word-bucket, and it reached 62.4% correct predictions – the most accurate genres were Metal & Blues, while Rap, Rock & Pop were around 50% accuracy each.

To this we added additional checks – we tested the addition five additional information points:  
a. the song’s length.

b. the average line’s length.

c. and average amount of appearances for all the words in the song – how reused are the words.

d. how diverse are the song’s lyrics – how many different words were in it, in relation to its length.

e. how original are its lyrics – how much were the song’s lyrics used in other songs.   
As it turned out, only checks b and e improved the results, but only to 64.8%:

We tried to play with an additional variable, which is the cost of constraint violation. As part of the algorithm, it considers this value to apply the penalty value of the distance from the given instance to the corpus instances. In this way, we can set how flexible the algorithm is, where the value is defined as follows:  
0 < constraint-violation-cost < INFINITY. The flexibility is expressed by the following: as the value is lower, the algorithm allows itself to “take risks” so it will be more suitable to the real world instances with unseen features. However, as this value goes higher, the algorithm become more and more restricted to the training set (the corpus) and does not allow flexibility and tolerance for unseen features. In order to answer what is the influence of this variable (the cost of constraint violation) over our data, and how does it affect on our results, we’ve run some test where we trained different models, each of the models was trained with different cost value. In the following figures, we show our results of the influence over difference cost-of-constraint values in the range of 1 to 10.

Interestingly, the additional checks reduced the performance of the check – without them results were as high as 65.6%, and never lower than 64% accuracy, while with them results reached as low as 37.6% accuracy, and never above the initial 64.8%.

On the highest accuracy marks – constraint set to 5.0 with no additional checks, we received the following confusion matrix:

From this we can learn the following things:  
a. Metal has an almost unique vocabulary, which characterizes its songs quite well – only a single metal songs was miscategorized by our program.   
b. Pop is a problematic genre (as it is in other discussions, too – defining a music by what’s popular, and not by its musical qualities), and so we had the lowest predictive accuracy with it.   
c. Rap is fairly well defined – rap songs were miscategorized as pop or metal, but almost no other song was miscategorized as a rap song.   
d. rock & blues are two relatively close genres – this is historically true, and since the majority of mistakes in one genre were towards the other, it seems that this connection is still evident in the lyrics of songs today.

We believe that the results shown here, which are consistent with outside knowledge we possess about the genres examined, show that lyrics do hold enough information to classify a genre, and that future works could potentially create a strong genre categorizer based only on songs’ lyrics.

**6. Limitations**

a. we were limited by the amount of songs we collected. We didn’t have effective means for acquiring a significant amount of songs for training and testing, and so our coverage was statistically weak. We believe this was especially harmful when trying to use n-grams instead of individual words, seeing as it significantly dispersed our dictionary, and prevent the creation of cohesive lexical spaces.

b. we relied on web-based lyrics databases, and this caused a hit to our accuracy. The content of these databases is uploaded by users, and so is not formatted in a consistent format, the genre classification usually applies only to the artist and not to individual songs, and the lyrics themselves are inaccurate – many times we had to deal with entries in which a part of a line was replaced with [incomprehensible], and we don’t know how correct the rest of the lyrics are.

c. Even without the subjective limitations of the uploading users, there is no objective song-genre classification. Genres are ill-defined, and so is the songs’ classification. This means that an objective training set doesn’t exist, and that the only meaningful definitions we can aspire to are statistical definitions – which, again, were hampered by our inability to mass-acquire training material.

d. Late into the project we noticed our own formatting was lossy – we removed verse differentiation, which lost us some information which we could’ve used to further analyze the songs.

e. Relying on a few songs, in a single language, uploaded to the internet exposes us up for significant cultural bias, based on the opinions & predispositions of the unknown contributors – for example, the majority of our songs are fairly new, which probably affects their lexical spaces.

**7. Future Work**

We believe that more exact results could be achieved by parsing the lyrics into trees, and examining the genres for repeating tree-structures – and since syllables have weight in a melody, this or other syllable counting methods could be a way to add the songs’ music into the calculation.

We also believe that the strength of rhyming could be an indicator of genre. Other concepts from the field of poetry, such as use of clichés & metaphors, could also be added to the check, as can regular NLP topic analysis.

**8. References**

1. LIBLINEAR: A Library for Large Linear Classification. Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, Chih-Jen Lin, Department of Computer Science, National Taiwan University, Taipei, Taiwan.
2. Categorical Data Analysis (Wiley Series in Probability and Statistics), Alan Agresti, John Wiley & Sons Publications.
3. Evaluating trauma care: the TRISS method. Trauma Score and the Injury Severity Score. Boyd CR, Tolson MA, Copes WS.
4. Logistic regression analysis for experimental determination of forming limit diagrams. M. Stranoa, B.M. Colosimob a Università di Cassino, Dip. Ingegneria Industriale, Italy, Politecnico di Milano, Dip. Meccanica, via Bonardi 9, Milano, Italy
5. Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines. [Sanjay Kumar Palei](http://www.sciencedirect.com/science/article/pii/S0925753508000118), [Samir Kumar Das](http://www.sciencedirect.com/science/article/pii/S0925753508000118), Department of Mining Engineering, Indian Institute of Technology, India.
6. Moore, Allan F. "Categorical Conventions in Music Discourse: Style and Genre". *Music & Letters*, Vol. 82, No. 3