# Lyrical Distinction

**Jesse Bartola** jbartola@
**Phil Michalowski** pmichalowski@
**Nischal Tamang** ntamang@
**Max Berezin** mberezin@

## 1 Introduction

Music is an important part of our everyday lives. Technological advances in modern society have made it easier than ever for new artists to distribute their songs over the Internet. Unfortunately, the massive volume of songs that are regularly released into the public can make it difficult to discern which song is made by which artist. Moreover, a more important question arises: Are the most popular artists the ones with the most distinct music?

Aside from sound, the words an artist uses in a song are perhaps the most important determining factor in how it will appeal to the public. Some of the greatest artists of all time are renown for their unique use of lyricism. Our project will aim to determine if (a) Given a set of lyrics, we can predict with a high degree of certainty who the artist is, and (b) Given our trained prediction model, we can rank artists by their relative lyrical uniqueness and use that to predict their popularity.

## 2 Related work

Have others worked on this idea or related ideas? Clearly describe the approaches of at least two other papers, along with some pros and cons. To look for relevant papers, check out the top NLP conferences (e.g., ACL, EMNLP, NAACL, TACL). Make sure to properly cite them. You can cite a paper parenthetically like this (Andrew and Gao, 2007) or use the citation as a proper noun, as in "Borschinger and Johnson (2011) show that..." If you're not familiar with LaTeX, you'll have to add entries to *yourbib.bib* to get them to show up when you cite them.

## 3 Your approach

How do you plan to solve the problem you chose? Will you approach it differently from previous work or do you plan to try to replicate an existing paper?[1] Remember that this project should take $\sim$ 2 months of work!

**What baseline algorithms will you use?** A baseline algorithm is one that is very simple and trivial to implement. For example, predict the most common class, or tag all capitalized words as names, or select the first sentence in the document. Sometimes it can be difficult to get a fancy algorithm to beat a baseline. Always ask yourself, Whats the simplest experiment I could do to (in)validate my hypothesis? Talented researchers have a knack for coming up with simple baselines.

### 3.1 Milestones & Schedule

Divide your project into subtasks and estimate how much time each will take. If your group plans to divide subtasks amongst itself, also write who will be responsible for each milestone. If you plan to work on everything together, please say so here. Definitely budget some time for writing the progress report and final report, as well as performing an in-depth analysis of any models you build and/or data you collect. Sample schedule below:

1. Acquire and preprocess data (1 week)
2. Build models for task (3 weeks)
3. Write progress report! (due Nov. 16)
4. Analyze the output of the model, do an error analysis (2 weeks)
5. Work on final report and presentation (2 weeks)

[1] If you choose replication, remember that you have to implement the majority of the code yourself! Do not just copy the authors' Github code, because we will find out.

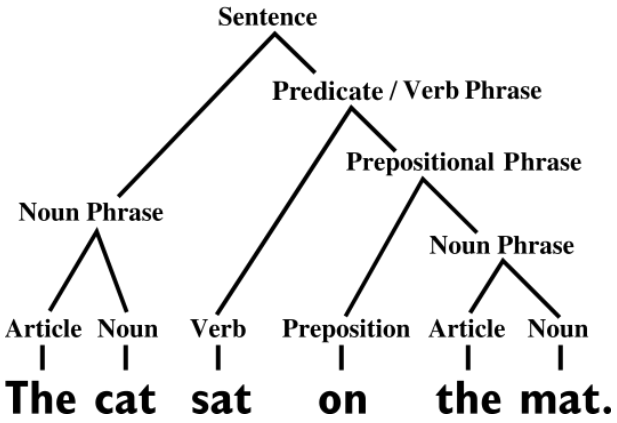


Figure 1: Please feel free to include figures! If you want your figure to span both columns, use *figure** instead of *figure*.

## 4 Data

There are many online datasets available containing song lyrics with their respective artists.

What text data do you plan to use in your project? Where will you get it from? Will you be annotating text yourselves? Convince us that it is available for you, and that you can easily get it, and that it is appropriate for the task and research questions you care about.

## 5 Tools

What existing libraries or toolkits are you going to use? Some questions to think about: will you be doing any preprocessing of your data such as tokenization or parsing? Will you be training logistic regression models? Will you be using deep learning libraries? Will you need to use any services for GPUs?[2] Do you need to use crowdsourcing?

## References

Andrew, G. and Gao, J. (2007). Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Borschinger, B. and Johnson, M. (2011). A particle filter algorithm for Bayesian wordsegmentation. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 10–18, Canberra, Australia.

[2] if so, check out https://colab.research.google.com!