Social Media Users' Notions of Popularity: Applying Machine Learning to Capstone Project

Stephanie Sears

1. **How do you frame your main question as a machine learning problem? Is it a supervised or unsupervised problem? If it is supervised, is it a regression or a classification?**

This capstone examines comments on YouTube to understand sentiment and public perception of popularity. As a text analytics project, the primary machine learning objective is to turn words (YouTube comments) into variables. These variables, in turn, can be employed in machine learning tasks. This project is not so much interested in predictive modeling as it is interested in understanding textual data. As such, this project is interested in natural language processing in order to derive meaning from textual data. Therefore, this is "unsupervised" machine learning. Framing the main question within this methodological context, my main question is: Can we develop analytical models, using text as data, that assess the popularity of R&B women's music on a large scale?

2. **What are the main features (also called independent variables or predictors) that you'll use?**

The main variables that I employ in this project are view counts, comment counts, like counts and dislike counts associated with various videos from identified YouTube channels of interest. The dependent variable is popularity.

3. **Which machine learning technique will you use?**

This project will employ a variation of collaborative filtering to learn the various themes social media users associate with popularity as well as clustering to segment the data into similar groups of thematic interest.

4. **How will you evaluate the success of your machine learning technique? What metric will you use?**

I will evaluate the model using topic classification on an alternate dataset. I will also mix this with sentiment analysis to understand how users feel about these topics. I will then see how these compare with results from the original dataset.