Goals:

* Classify unstructured text using machine learning
* Create way to predict types of products that people have bought, as well as sentiment with films.

3 requirements:

* Unstructured data
* You’ve categorized some of it yourself
* You have a threshold of acceptable time/acceptable error

Review of Machine Learning terms:

* **Machine learning:** inferring a property of a new piece of data based on prior experience with similar types of data.
* Algorithm/model approach
* **Classification**
* **Training set**
* **Test set**
* **Confusion matrix**
* Demo 1: classifying twitter categories
  + What we’re doing here: Naïve Bayes
    - Step 1: Uploading to HDFS
    - Step 2: Vectorizing data (this is where it gets difficult)
    - Step 3: Training Model
    - Step 4: Testing on the training set
    - Step 5: Testing on test set
* Demo2:
  + Scaling to larger data set
    - Advantages of Naïve bayes
    - Disadvantages of Naïve bayes: small, less-well defined datasets perform less well… need a LARGE amount of training data
* Demo3: SGD
  + Larger IMDB data set
    - SGD computes the loss function continuously of using a set of features with weights attached to it
    - Has a list set of functions for classifying and computes the average of the performance of that function on all training examples so far
    - Greedy algorithm—classifies with the function that has the greatest average success.
    - It is stochastic in that the input is stochastic.
  + Disadvantages:
  + SGD requires a number of hyperparameters such as the regularization parameter and the number of iterations.
  + SGD is sensitive to feature scaling.

Oh but WAIT! Our input is NOT stochastic

It CAN be stochastic if we actually randomize it… or don’t overfit.

* Demo 4: SGD on the Hobbit versus Naïve Bayes on Hobbit

What’s next?

* Synonym mapping with Wordnet. I have a means of determining language, but not of adding synonyms to the tokenstream. I’ve been told on forums that that can improve the model by 5-10% percent instantly.
* Random Forest. The holy grail for small feature set.