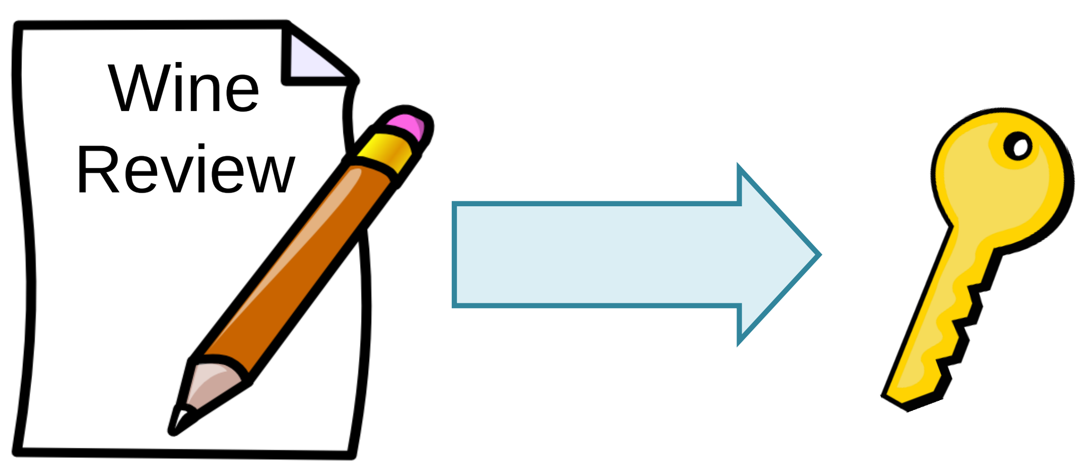
Wine Review Keyword Prediction

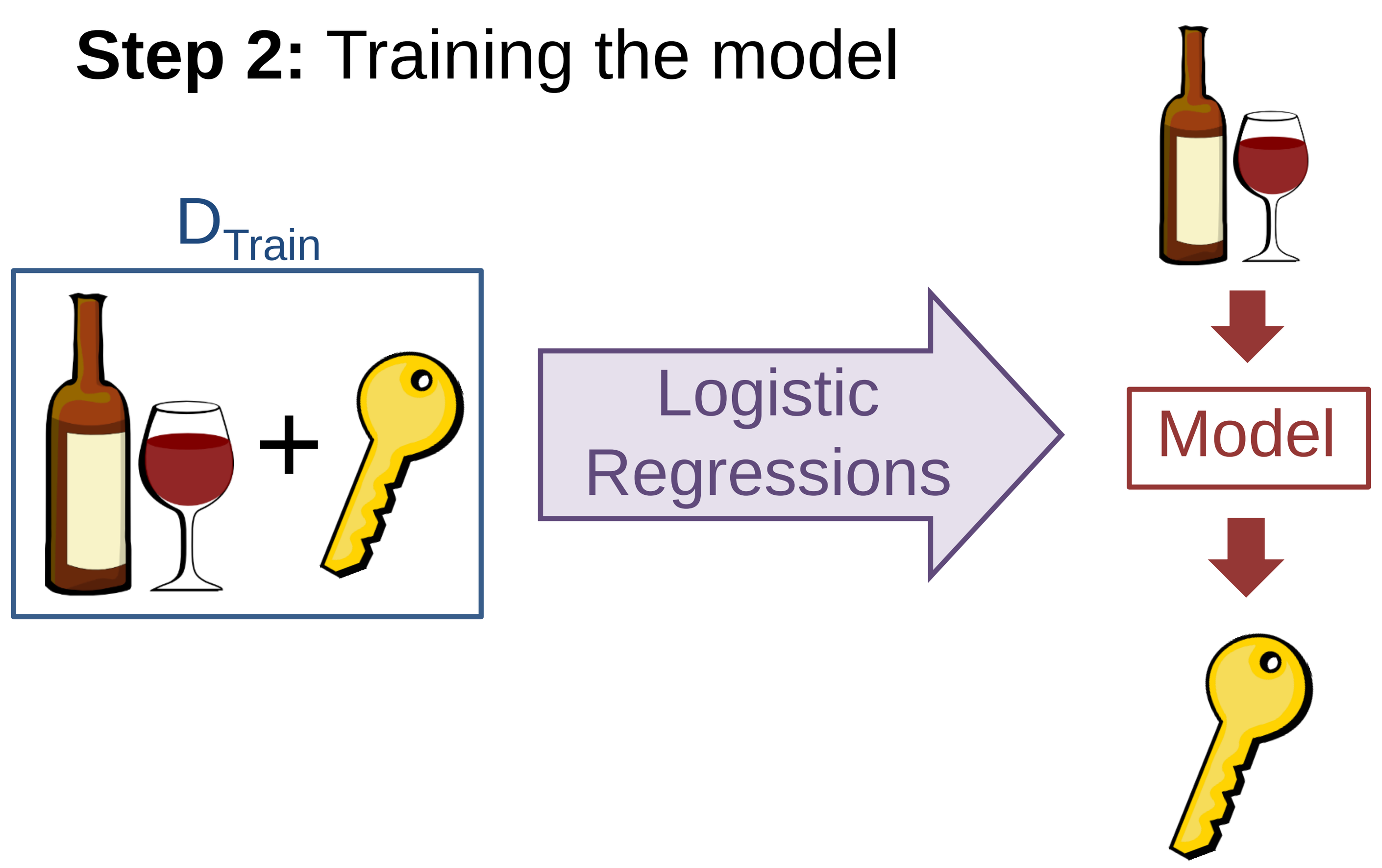
Noam Habot, Mackenzie Pearson, Shalini Ranmuthu
CS221 – Artificial Intelligence, Stanford University

Problem and Motivation

Step 1: Building dictionary of keywords



Step 2: Training the model



Motivation: To explore correlations between wine attributes and keywords found in reviews

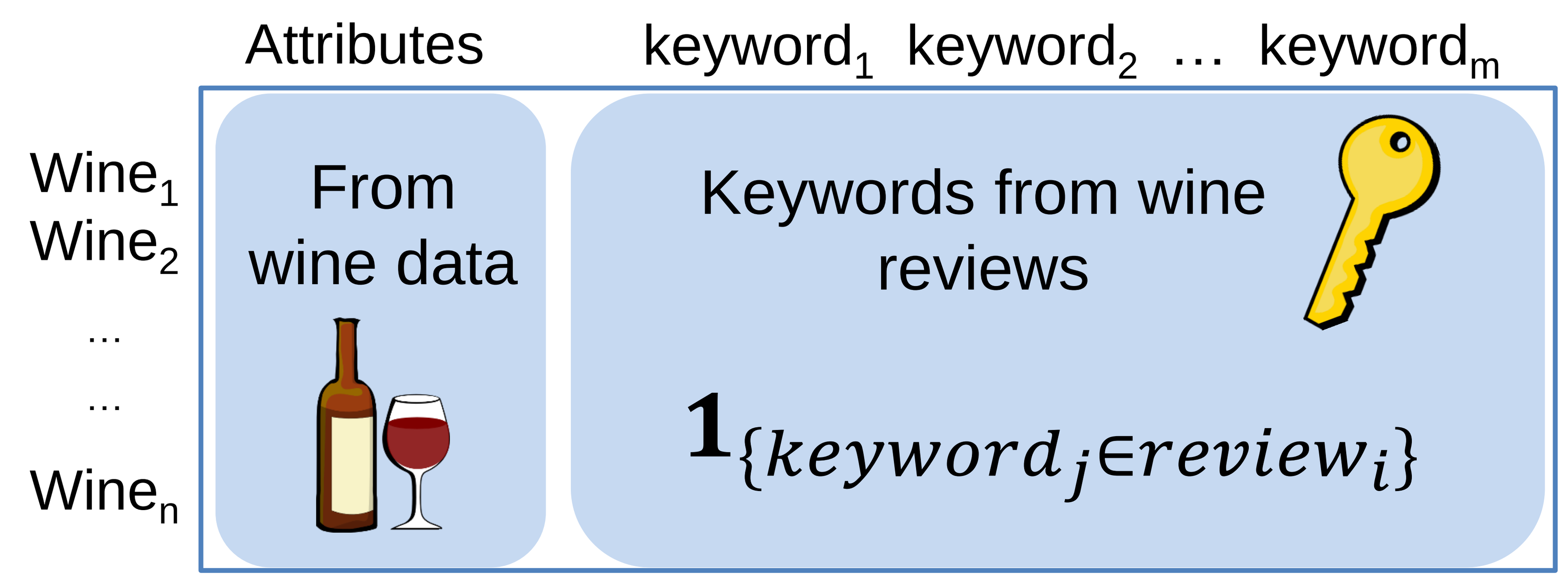
Approach

Step 1: Take most frequent words from each set to get keywords

$$\begin{pmatrix} \text{and} \\ \text{the} \\ \text{aroma} \\ \text{to} \\ \text{citrus} \\ \dots \end{pmatrix} \setminus \begin{pmatrix} \text{the} \\ \text{and} \\ \text{to} \\ \text{you} \\ \dots \end{pmatrix} = \begin{pmatrix} \text{aroma} \\ \text{citrus} \\ \dots \end{pmatrix}$$

From wine reviews — From generic book — Final keywords

Step 2: Build D_{train}



- Reduce cardinality of categorical features to increase predictive power of feature labels
- Create *m* logistic models

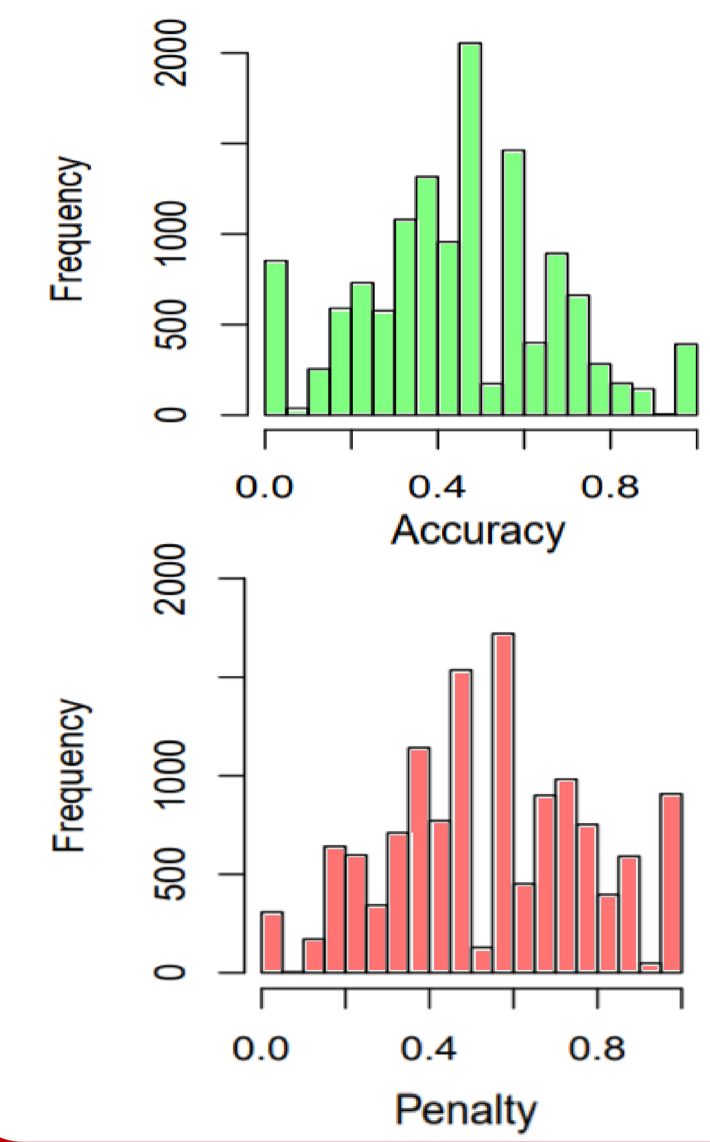
$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

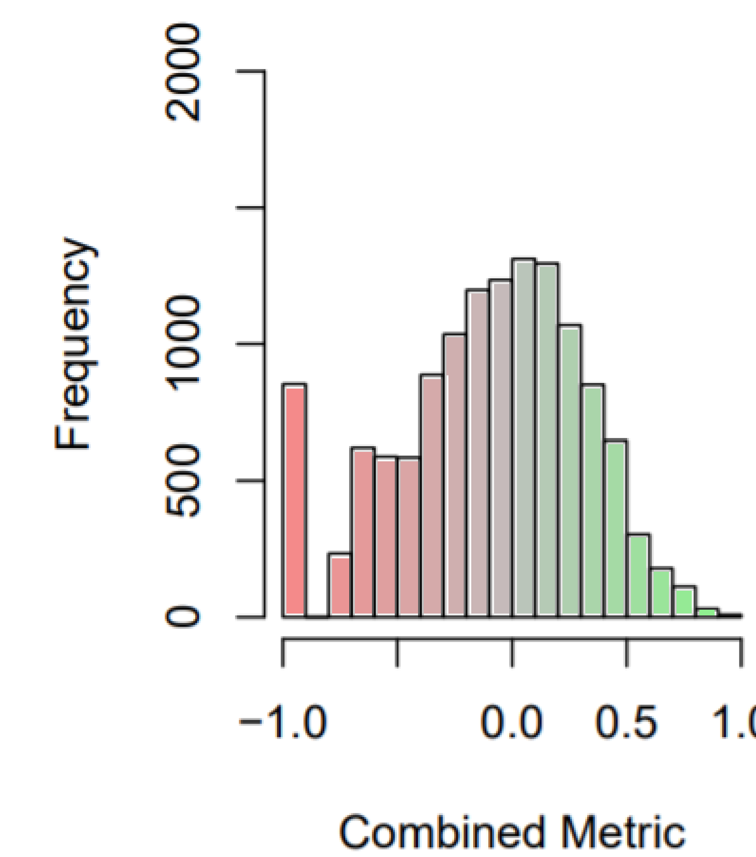
$$L(\beta_0, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

Results

The following evaluation metrics can be calculated for each wine (observation)

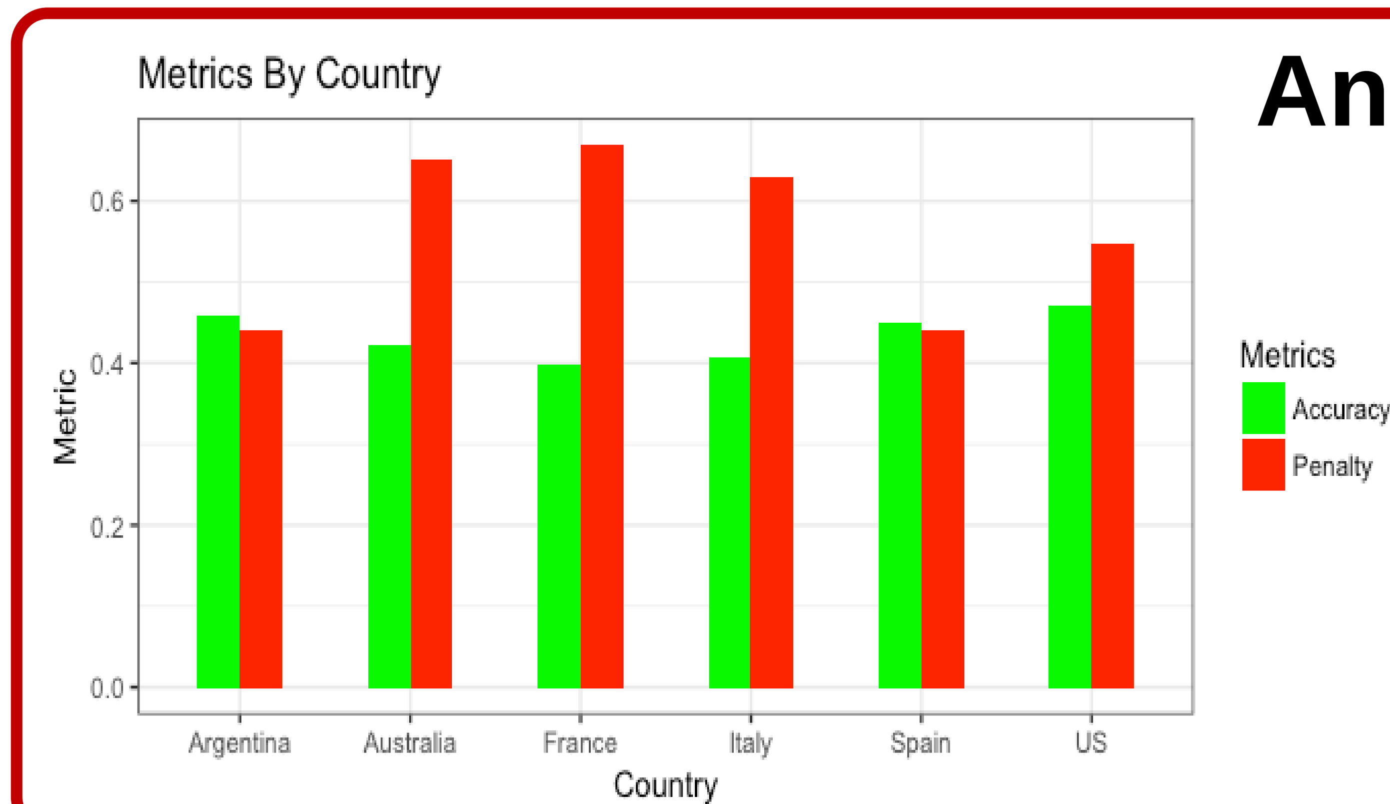
1. **Accuracy:** $\frac{\#\ keywords\ correctly\ predicted}{\#\ total\ keywords\ in\ review} \in [0,1]$
2. **Penalty:** $\frac{\#\ keywords\ incorrectly\ predicted}{\#\ total\ keywords\ predicted} \in [0,1]$
3. **Combined:** $(Accuracy - Penalty) \in [-1,1]$





Challenges and Implementation

- Storing *n* x *m* D_{Train} for quick reload (n=150,930 wines, m=6342 keywords)
- Computation time (building D_{Train}, creating models, and computing metrics)
- Parallelizing construction & prediction of logistic models for each keyword
- Previously only feasible to predict on top m=100 keywords (plan to reach m=1000 for final evaluation)
- Testing on the 100 most frequent keywords may mean that the keywords are not strongly correlated with certain wine features



Analysis

- On average, our model predicts half of the correct words, but ~50% of the predicted words are extraneous
- ~7% of the reviews in the test set did not contain keywords used in the model (combined metric = -1)