Automatic Product Categorization for Anonymous Marketplaces

Project Overview

Anonymous marketplaces are a rapidly growing segment of online illegal drug sales. However, due to their clandestine nature, it can be difficult to extract information about product listings without manual intervention.

In this project, we built a machine learning algorithm to extract listing type and category from public listing text.

This information provides valuable insight about marketplace trends to law enforcement and researchers.

Categorization Results

Learning Algorithms

- **Baseline Heuristic Model**
  - No machine learning. Categories chosen by substring matching against a known dictionary of product names.

- **SVM with Unsupervised Features**
  - Word features were projected onto a low dimensional (~300) subspace, chosen using principal component analysis on the large (~100,000) unlabeled training set. Trained using stochastic gradient descent with L1 regularization and a linear loss function.

- **SVM**
  - Trained like the SVM with Unsupervised Features, but using labeled features (~30,000) instead.

- **Multinomial Naive Bayes**

Future Work

Our training and testing used data from only a single market. The algorithm could be made more robust by including data from more sources. Test data drawn from a broader source would provide a better generalization estimate.

Our category labels in conjunction with another learning algorithm. For example, we could use our classifications along with vendor history to predict anomalous behavior.

Model hyperparameters were chosen somewhat arbitrarily. Each should be chosen in a more principled way using cross validation.