Evaluation of VSM as a classification algorithm using Amazon.com data

Koushik Kumaraswamy

Wednesday, Mar 12, 2015

==========

Introduction

In this report, we evaluate the suitability and performance of the Vector Space Model, (referred to as VSM henceforth), as an item classification algorithm operating on Amazon.com item data. This inquiry is relevant given that VSM has traditionally been employed to solve search problems but has not been employed for solving classification problems given the lack of an underlying theoritical probability model.

The published test results cannot be used as-is to generalize algorithm performance on real-life PO and AP data sets, but instead serve to establish a baseline level of performance given fairly rich input data. This baseline will serve as a comparative datapoint for subsequent analyses that utilize real-life data.

In addition to establishing baseline, another goal of this analysis is to observe changes in VSM performance as number of records as well as classification granularity are increased.

Design of Experiment

This experiment was modeled after a traditional machine learning experiment with by treating VSM as a model for classification . The experiment,designed per reproducible research standards, is outlined below:

On model formulation

The classification schema used in this analysis was UNSPSC with the following classification groups: Level 1:Segments [56], Level 2: Families [411], Level 3: Classes [3713], Level 4: Commodities [46137].

The crux of this approach involves modeling each classification level as a "weighted bag of words" - i.e. every level in the hierarchy is associated with a set of weights (numbers between 0 and 1) relative to a dictionary of words. More formally, a classification level is represented as a vector within n-dimensional vector space, where n is the number of words in the learned dictionary. The classification problem is then solved by representing a incoming new item (query) as another point in N-D space and assigning it to its "nearest neighbour". The angular seperation between the vectors (computed through the vector dot product) is distance measure used in this model. A larger dot product score indicates a higher rank or closeness to category vector.

Raw Data acquisition and cleansing

Raw Data for this analysis came from the data scraping exercise that the Netlink team was contracted to perform. One CSV (comma seperated value) file with more than 399k records was chosen as the raw data source for this analysis. This file was downloaded from the Netlink FTP server, copied , and modified to include the UNSPSC codes for all four levels (by extracting this information from the item specification column). The modified CSV data file was used as the input into R for this analysis.

Data pre-processing

The modified input file was read into working memory as a dataframe object. The rows and Columns that were erroneous or not useful for downstream processing were dropped. The raw Data fields viz. item name , description, and item features were merged into a large text field which to denote a single item within the search space.This data was then segmented into groups of similar categories (based on selected classification granularity).

Training and Test set segmentation

The previous step yielded a collected of objects, each of which denoted a set of items belonging to the same group. These items were randomly distributed between a training set and a test set in the ratio of 80% to 20% . Groups that did not have a minimum records were excluded from the downstream analysis. The intent of this step is to train the model on the *training set* records and to evaluate its performance on the *testing set*.

Training the model

For VSM, training involved the creation of set of artificial documents, each representing a group within the selected classification hierarchy. A Corpus comprising of these artificial documents was then created and normalized through various tranformations (conversion to lower case, stripping white space and removing punctuation). Stemming was not performed during this analysis. A term document matrix was then created using the transformed corpus. This matrix and its associated dictionary represents the trained model for this analysis.

Testing the model

To get around computing power constraints of the analyst's computer, the test bed was created using a random sample of records from the testing data set. The constraint imposed on the sampling was to limit each item group's representation to at most 10 instances inside the test subset. Each item in the test bed was presented to the model for classification and the result recorded. As referenced in the formulation section, the output classification from the model corresponds to category that has the largest value from the dot product between query and categories. The performance of the model was quantified through a simple accuracy measure defined as [(# of correctly classified tests)/(# of total tests)].

The training and testing of the VSM model was performed in sequence

Results

Exploratory data analysis Statistical prediction/modeling Interpret results Challenge results Synthesize/write up results Create reproducible code