Predicting Stackoverflow tags
02807 Final project

November 19, 2016

1 Introduction

In this project a model predicting tags associating a given text will be constructed. The problem will be handled as a classification problem, hence a classifier will be trained.

The classifier will be trained on posts from Stackoverflow where associated tags have been given.

Initially only the top 20 frequent tags and posts containing these tags will be used for training and prediction.

Finally the model will be evaluated using all tags and posts.

During implementation and debugging a subset of the data will be used.

2 The data

The dataset consists of two XML files, one containing all possible tags and their corresponding counts, and one containing posts with *title*, *body*, *tags* and some meta data.

The total size of the files are approximately 49GB in uncompressed format.

3 Methods

For modelling the distribution of the post tags an unsupervised approach is used.

The data will be presented using feature-hashing. Each post will be transformed to a fixed-size feature vector resulting in each post being a point in a given feature space.

A training set of posts will be clustered into K clusters where K is the number of distinct tags in the data, resulting in K cluster centroids.

When predicting a given post the tags of the closest n centroids will be assigned the post.

The cluster algorithm *K-means* will initially be used, and due to the size of the dataset a normal trivial implementation of the algorithm will not suffice, hence a parallel version will be implemented.

4 Future plan

For the final two weeks of the project I will be working on the following:

- Fix memory issue of parallel K-means implementation
- Find proper feature-hashing (currently using simple trivial hashing)
- Train and evaluate model
- Finish the report