Large Scale Hierarchical Text Classification

Classify Wikipedia documents into one of 325,056 categories

119 teams · 3 years ago

Training Data

7 files

- AllZerosBenchmark.zi...
- hierarchy.zip
- knn-baseline.tar.gz
- test-remapped.zip
- test.zip
- train-remapped.zip
- train.zip

hierarchy.zip

Download File

File size 4.03 MB

Data Introduction

File descriptions

- **train** - Training set
- **test** - Test set
- **hierarchy** - Wikipedia hierarchy
- **AllZerosBenchmark** - example submission file
- **knn-baseline** - A simple flat kNN baseline
- **train-remapped, test-remapped** - Training and Test sets reformatted per this forum thread

Hierarchy

The hierarchy file contains the information regarding the hierarchy of classes. Each line of this file is a relation between a parent and a child node. For example, the line:

```
897 67
```

is to be read as node 897 is parent of node 67

Data

The format of each data file follows the libSVM format. Each line corresponds to a sparse document vector and has the following format:

```
label, label, label ... feat:value ... feat:value
```

label is an integer and corresponds to the category to which the document vector belongs. Each document vector may belong to more than one category. The pair feat:value corresponds to a non-zero feature with index feat and value value. **feat** is an integer representing a term and **value** is a double that corresponds to the weight (tf) of the term in the document.

For example:

```
545, 32 8:1 18:2
```

corresponds to a document vector whose features are all zeros except feature number 8 (with value 1) and feature number 18 (with value 2). This document vector belongs to categories 545 and 32. Each feature number is associated to a stemmed word.

The labels of the test document vectors are set to 0.