# DAPI Technical Documentation



| Name | Size | Creator | Creation Date | Comment |
|---|---|---|---|---|
| Text File Links.txt | 1.09 MB | laura.tolosi | May 26, 2015 12:01 | |
| PDF File DekelKeSi04.pdf | 129 kB | laura.tolosi | May 26, 2015 11:51 | |
| PDF File MultiLabelHierarchicalPerceptron.pd... | 241 kB | laura.tolosi | May 26, 2015 11:49 | |

## Categorization scheme

. A popular categorization is based on the the IPTC categorization scheme, suitable for news articles. Many of our competitors are basing their categorization on the IPTC standard. Advantage: describes well news. Disadvantage: it is a flat categorization, does not have levels. Too targeted to news.

To date, the categorizations comprizes 17 broad topics: Arts_Culture_Entertainment, Conflicts_War_Peace,Crime_Law_Justice, Disaster_Accident, Economy_Business_Finance, Education, Environment, Health, Human_Interest, Labor, Lifestyle_Leisure, Politics, Religion_Belief, Science_Technology, Society, Sports, Weather.

For the next development versions, it is possible (and desired) to extend the categorization scheme by appending sub-categories, identical or inspired by the IPTC. More refined categories can result in a more specific description of the topic of the document, but can raise problems with model fitting.

B. For unsupervised approaches, where the categories are not specified apriori, one can use ontology terms, such as dbpedia categories, of various degrees of specificity.

## Corpora

A. A corpus consisting of long abstracts from dbpedia of articles that belong to the 17 IPTC categories, as shown here: https://confluence.ontotext.com/display/GSC/Document+Classification+Corpora . The corpus is available in EN and BG.

B. One corpus has been obtained form the ACM classification system . It consists of titles and abstracts of scientific papers published by ACM. Here is the file. Each row starts with CCS, which is the root category of the tree. Tab-separated records specify parths in the tree. The leaves are articles, given as title and abstract, tab-separated. Example of articles in category:

**CCS** -> **General and reference** -> **Cross-computing tools and techniques** -> **Metrics**

| Title | Abstract |
|---|---|
| Measured impact of crooked traceroute | Data collected using traceroute-based algorithms underpins research into the Internet's router-level topology, though it is possible to infer false links from this data... |
| Semantic mining on customer survey | Business intelligence aims to support better business decision-making. Customer survey is priceless asset for intelligent business decision-making.... |
| Predicting software complexity by means of evolutionary testing | One characteristic that impedes software from achieving good levels of maintainability is the increasing complexity of software... |
| Runtime monitoring of software energy hotspots | GreenIT has emerged as a discipline concerned with the optimization of software solutions with regards to their energy consumption.... |
| Structured merge with auto-tuning: balancing precision and performance | Software-merging techniques face the challenge of finding a balance between precision and performance... |

## Approaches

- Multi-label large margin hierarchical perceptron, Woolam and Khan, Int. J. of Data Mining, Modelling and Management, 2008

- Large margin hierarchical classification, Dekel et al, ICML '04 Proceedings of the twenty-first international conference on Machine learning

## Existing models

- Pipeline based on the dbpedia articles corpus (A) already available at S4: http://docs.s4.ontotext.com/display/S4docs/News+Classifier . Only for EN. Has low recall, rarely outputs more than 2 categories.
- An ensemble model, which combines a gazeteer and a classifier. The classifier outputs "yes" or "no" for each category. It is based on a small number of features, up to 30. Some reduced language model that hashcodes words to categories.
- An unsupervised model that works with tagged entities in the documents and tries to find dbpedia supercategories that cover well the entities. Unwanted aspect: very broad supercategories such as "Living_people" are very often output and are unspecific. The approach is promising, but some specificity score of the output categories mush be introduced.

Features:

- We are currently using: stopwords elimination, stemming and a bigram model for feature extraction
  Algorithm:
- The multi-label classification is achieved by training K independent classifiers (perceptron, sigmoid perceptrons), corresponding to the K possible labels. For each classifier, the interpretation is: what is the likelihood that sample x has label l, against the alternative that it does not? After training all K classifiers, for each sample, the top highest likelihoods give the set of labels. A rule of thumb is used for deciding how many labels should be returned.