DATA SCIENCE CHALLENGE

[dstl] Government Office for Science SECURITYSERVICE MI5 SECRET INTELLIGENCE SERVICE MI6

Challenge Guide
Growing Instability: Classifying Crisis Reports

The following content is supplementary content to the challenge content on the website.

Data Specification

Data Files

The data for this challenge has been acquired from a major international news provider, the Guardian. The training data represents the past historical record, and the test data represents new documents that require classification.

This dataset consist of:

- **Training data** (TrainingData.zip): All the news articles published between 1999 and 2014. [2.4GB]

- **Test data** (TestData.zip): A sample of the news articles published between 2015 and 2016. [14MB]

- **Topic dictionary** (topicDictionary.txt): A list of topics for classifying articles that improve awareness of the developing crisis. [3KB]

- **Sample submission** (sampleSubmission.zip): A sample submission file with the correct format but random topic predictions. [141KB]

Training data

The training data is contained within the zip file TrainingData.zip. The contents of this zip file are a set of 32 JSON formatted files, e.g. 2012a_TrainingData.json (2 files for each year ranging from 1999 to 2014). Each file contains multiple articles and each article is separated from the other articles by its unique reference key (see example below).

The filenames refer to the year the articles contained within the file were published, followed by an ‘a’ or ‘b’ to denote whether they were published during the first or last six months of that year respectively. The large training dataset has been split into these smaller files for convenience, to allow participants to manage and select data.

An example of one of articles within a JSON formatted training file is shown below:

```
“2014a_TrainingData_36617”: {
        “webPublicationDate”: “18-02-2014”,
        “topics”: [
            “health”,
            “society”,
            “politics”,
            “uk”,
            “police”
        ],
        “bodyText”: “The number of times police cells are used as a place of safety for people
having a mental health crisis is intended to be halved under a far-reaching agreement…”
}
```

There are multiple articles with the same format in each training file. Each article has a unique reference key (in this case 2014a_TrainingData_36617) and contains information for three further fields as follows:

- webPublicationDate [DATE dd-MM-yyyy]. This refers to the date the article was published on the The Guardian web site

- topics [STRING LIST] This refers to the topics assigned to the article by The Guardian journalist who wrote the article

- bodyText [STRING] This refers to the text content of the article. Please be aware of unicode characters in the text (e.g. for punctuation characters) which you may need to be convert

The training data comprises the full set of articles published during the period 1999-2014. In total there are of the order of1.6 million articles.

Test data

The test data is contained within the zip file TestData.zip. The zip file contains a single JSON-formatted file called TestData.json.

This JSON file has the same structure as the JSON-formatted training files. The articles are presented in the time order of their publication. Each article has a unique reference key (e.g. TestData_08249). The ‘topics’ field for each test article is blank (“topics”: [ ]) as the challenge is to predict topics for the test articles.

The test set is composed of a sample of articles published during the period 2015-2016. In total there are 7581 test articles. This sample relates broadly (though not exclusively) to the growing instability theme about which we would like to improve understanding.

Topic dictionary

The topic dictionary is contained within the plain text fileTopicDictionary.txt. This file contains a list of topics that are of interest in helping to understand the developing crisis.

There are 160 topics in the dictionary and these are listed in alphabetical order. Each topic is of equal interest and this is reflected in the scoring.

Further information

This section is for background information only and provides further detail about the data we have prepared for the challenge in relation to the original Guardian source data.

What we have referred to as topics here, are referred to by the Guardian as keyword tags. All Guardian content is manually categorised using keyword tags. These have the form A/B, specifying a coarser (A) and a more refined (B) categorisation of an article e.g. environment/recycling. In some cases A=B.

The topics in our topic dictionary represent a very small subset of B-level tags from the Guardian's extensive keyword tag dictionary. We have not renamed or altered the original Guardian tag names, with the exception of multi-term tag names (e.g. organised crime) which we have represented in the topic dictionary as single terms (e.g. organisedcrime).

For each article in the training dataset the provided topics correspond to either the Guardian's B-level or A-level tags. If A ∈ {world, UK, politics, law, technology, global-development, science} then B-level tags are provided, otherwise A-level tags. This was done to provide refined categorisation in the training data for only the coarse-level categories that are expected to relate to articles relevant to the growing instability and crisis theme.

The Guardian provides its webPublicationDate meta-data to the nearest second of publication, e.g. 2014-02-17T12:05:47Z. We have coarsened this meta-data to the nearest day of publication, e.g. 2014-02-17, which should be sufficient accuracy for any solution that wishes to exploit temporal information.

The Guardian provides a significant amount of other meta-data with its articles. In this challenge we are only providing you with the web publication date, topics / keyword tags (training data only), as well as the article text. Omitted meta-data includes: other tag fields (e.g. contributor/author, tone, and series) and various content fields (e.g. web title and web URL). For further information see open-platform.theguardian.com/documentation.

Scoring

Submitted solutions for the test articles will be evaluated with respect to the ground truth for those articles, which is exactly known. This assessment will be done using the well-known F_1 score to produce an overall measure of performance.

The F_1 score for a classifier is the harmonic mean of the precision and recall metrics. Specifically, for document classification in which a document has to be classified by labelling it with respect to pre-defined topics from a dictionary, the F_1 score we will use to score submissions will be calculated as follows:

$$\text{Precision } (p) = \frac{\sum_{i=1}^{T}(tp)_i}{\sum_{i=1}^{T}(tp)_i + (fp)_i}; \qquad \text{Recall}(r) = \frac{\sum_{i=1}^{T}(tp)_i}{\sum_{i=1}^{T}(tp)_i + (fn)_i}; \qquad F_1 = \frac{2 \cdot p \cdot r}{p + r}$$

tp, *fp*, and *fn* denote true positives, false positives, and false negatives, respectively. T is the number of topics in the topic dictionary.

This is known as the micro-averaged F_1 score. It regards each article and each topic of equal importance. The score ranges between 0 and 1, with 1 denoting perfect prediction.

Guidance

The most common approach to predicting topics for text documents is to use machine learning (ML). Specifically, this challenge falls under the ML area of multi-label classification, because each document can be classified with multiple topics.

Supervised topic modelling is a common approach. Here you would use the input (text article) and output (topic labels) pairs for each of the training articles to build a classification model, and use that model to predict the labels for the test articles.

Unsupervised topic modelling approaches, such as those which use dimensionality reduction and clustering, may also be used. However, the predicted topics must match the topics in the topic dictionary.

It is usual to pre-process text before it is classified. This typically involves tokenising the text, removing stop words, and applying a stemmer.

The next step is often some form of vectorization of the data. Note that this can produce a very large data matrix. You will need to be clever about how you represent this to reduce your memory footprint, although there are no specific solution restrictions per se.
Data scientists often use Python. The packages you may find most useful are nltk for pre-processing and Scikit-Learn for machine learning.

The Challenge Master is our expert for the challenge and will be available to offer guidance through the challenge forum.