## Reuters-21578 News classification

## Guidelines

**Expected work**

All data preparation and data analysis tasks you will perform should be thoroughly documented in your code script. You are expected to state all assumptions you may make and provide clear explanations of your choices and results.

**Programming language**

We would prefer that you conduct your analysis using Python

**Assessment criteria**

You will be evaluated on your technical skills and on your ability to clearly present your results.

**Exercise rules**

In order to complete the exercise, you have the possibility to use the Internet and any other source of information you will find useful. However, the case study must be carried out **alone.**

## Context

The Reuters-21578 dataset contains 21 578 financial articles tagged with topics.

There are 135 different topics, but this exercise will focus on only 5 of them:

* Money/Foreign Exchange (MONEY-FX)
* Shipping (SHIP)
* Interest Rates (INTEREST)
* Mergers/Acquisitions (ACQ)
* Earnings and Earnings Forecasts (EARN)

## Analysis to conduct

### Descriptive analysis

Which are the main characteristics of each topic?

### Predictive analysis

You want to classify automatically each article into a category. Which approach should you use? Once implemented, which algorithm scores the best?

## Data

The dataset is composed of 22 sgm files containing the 21 578 Reuters articles as well as some useful attributes of these articles.

You will also find additional txt files that may help you understand the dataset organization better.

You can find the dataset at this address:

<https://archive.ics.uci.edu/ml/machine-learning-databases/reuters21578-mld/reuters21578.tar.gz>