Homework 2: Document Classification

Form:	Jupyter notebook file
Language:	English
Requirements:	The report should be clear, readable and include all code documented
Submission:	Ipynb uploaded to moodle
Contact:	Simo Hanouna simohanouna@gmail.com
Deadline for submission:	**December 31, 2019**

Students will form teams of two people each, and submit a single homework for each team. The same score for the homework will be given to each member of the team.

Submit your solution in the form of an Jupyter notebook file (with extension ipynb). Images should be submitted as PNG or JPG files. The whole code should also be submitted as a separate folder with all necessary code to run the questions separated in clearly documented functions. Python 3.6. should be used.

The goal of this homework is to let you practice with document text pre-processing, dataset exploration and document classification with python.

Submission: Submission of the homework will be done by uploading the Jupyter notebook to Moodle. Please include also screen shots of the results and the saved models, so I'll not have to retrain the models. The homework needs to be entirely in English. The deadline for submission of Homework 2 is set to December 31, 2019 end of day Israel.

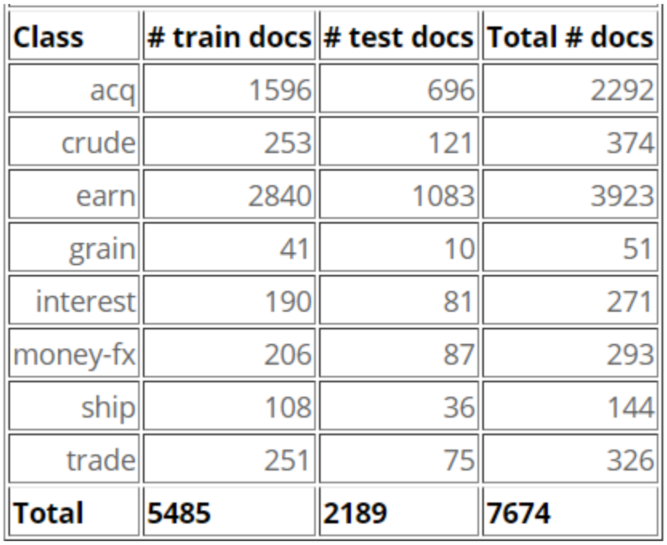
Task

The Reuters-21578 text categorization dataset is a widely used for text categorization research. More details about this dataset are available here. A modified dataset derived from the original dataset is provided for this homework. The homework dataset includes only documents which have a single label belonging to the top 8 categories in the original dataset. Table 1 below describes the 8 categories labels and the number of documents for each label in the provided train and test set.

Download the modified dataset train, test files.

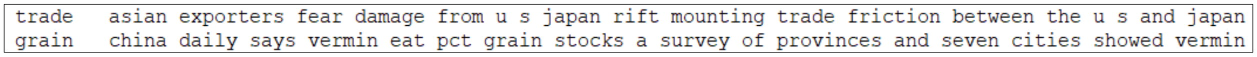
Use the external libraries and resources presented in the class for task implementation. Please set a variable at the beginning of the exercise, with the dataset folder.

The first word in each line is the document label follows by the document. Figure 1 presents the first two lines of the test set. The label of the first document is trade and the label of the second one is grained. The first word of the first document is “asian” and the first word of the second document is “china”. The different steps of the task are described below.



Class	# train docs	# test docs	Total # docs
acq	1596	696	2292
crude	253	121	374
earn	2840	1083	3923
grain	41	10	51
interest	190	81	271
money-fx	206	87	293
ship	108	36	144
trade	251	75	326
Total	**5485**	**2189**	**7674**

Table 1 – Dataset description



```
trade   asian exporters fear damage from u s japan rift mounting trade friction between the u s and japan
grain   china daily says vermin eat pct grain stocks a survey of provinces and seven cities showed vermin
```

Figure 1 – Line example

1. **Text pre-processing and exploration:**

- Download the corpus.
- Split to train and test
- Clean and normalize the text (e.g. tokenization, lower case, stop words removal, stemming)
- Explore the dataset (#of categories, #of docs from each category, terms distribution per category). Present a table of top 10 words per category.
- Explain the expected challenges (e.g. top words which are common to multiple categories)

2. **Document classification:**

Here, you should test combinations using 2 feature extraction methods and 3 machine learning models to train a classification model. Test the impact of changing at least one parameter per feature extraction and machine learning model on classification result.

- Implement feature extraction (Bag of words, n-grams, TF-IDF, any other feature - optional)
- Classify using machine learning methods (e.g. SVM, Naïve Bayes)
- Tune each model parameters, as well as pre-processing and parameters steps to optimize the results
- Use accuracy metrics to compare between the different models

- Use the best model selected in the previous steps for prediction on the test set. Present the accuracy of the model and the challenges.
- Describe the task challenges, and explain effective solutions

Good luck