# Practice IV

Text classification

# Specifications

- Individually do the following
  - Load the 20 newsgroup corpus
  - Obtain the train and test sets
  - Apply text normalization processes
  - Create different text representations of the corpus
  - Use different machine learning methods to train a model and predict test instances
  - Evaluate predictions of models

# Specifications

- Text normalization
  - For this processes you can use:
    - Tokenization
    - Text cleaning
    - Stop words
    - Lemmatization
  - You should try different step combinations or versions in order to improve the classifier performance

# Specifications

- Text representation
  - For this processes you can use:
    - Binarized
    - Frequency
    - TF-IDF
    - Embeddings
  - You could try SVD to generate a different version of text representation

# Specifications

- Machine learning methods
  - For this processes you can use any classifier that supports multi-class classification
  - It would help if you tuned the algorithm parameters to improve the results

# Evidence

- Source code
- A report in PDF format describing the following:
  - Task to be solved
  - Text normalization process
  - Text representations
  - Machine learning methods

# Evidence

## A table describing the experiments performed

| Experiment | Text normalization | Text representation | Machine learning method | Average f-score |
|---|---|---|---|---|
| 1 | Tokenization + stopwords + lemmatization | binarized | Logistic regression | 0.85 |
| 2 | Tokenization + stopwords + lemmatization | frequency | Logistic regression | 0.88 |
| ... | ... | ... | ... | ... |
| n | Tokenization + text_cleaning + stopwords + lemmatization | Tf-idf + svd | Multilayer perceptron | 0.9 |