# Lab work 3: Text sentiment analysis

## 1. Goal

Study and assess the performance of a text sentiment analysis sample employing the TF-IDF [2] in conjunction with various classifiers. The accuracy will be the only parameter for the assessment. You will be using a large movie review data set provided by kaggle [1] and the necessary scikit learn libs. The Kaggle data set comprise about 50k movie reviews with its sentiment positive or negative.

[1] https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

[2] https://en.wikipedia.org/wiki/Tf%E2%80%93idf

## 2. After completion you have learned

- The impact of the "max_features" parameter for the tfidf technique in conjunction with the employed classifier, upon the performance of a text sentiment classification.
- Working with the TfidfVectorizer() class and classifier using the scikit learn libs
- Applying k-fold cross validation and parameterizing the logistic Regression and the SVM classifier provided by the scikit learn libs.

## 3. Tasks

**A group consist of up to two students.**
**If a single person builds a group, assignments c) and d) are not required.**

a) Study the obtained **test data accuracy** using the above kaggle data by varying different parameters. In particular, **draw curves** of the obtained **accuracy versus the C parameter**, employed logistic regression classifier. Thereby varying the **max_feature parameter** of the TfidfVectorizer() class in the range of {100, 1000, 10000, 100000}. For each parameter setting perform 5 runs and plot the obtained **average accuracy**.

**Remark for advanced programmer (optional):**
You may also employ the GridSearchCV and Pipeline objects solution carry out gridsearch for both tfidfvectorizer and classifier, e.g.,

```
pipeline = Pipeline([tfidf, estimator,….])

parameters = {….max_features: (100, 1000, 10000,
1000000),estimator__...)}

tune = GridSearchCV(pipeline, parameter,…)

….
```

b) Briefly discuss the results of the plot (parameterized curves of C parameter versus accuracy) and make a conclusion regarding the optimal parameters (C-parameter and max_features parameter).

c) Carry out the test accuracy as stated in a) but this time employ the SVM classifier. Choose the kernel and its parameters appropriately **(group of two only).**

d) Briefly compare the results with the logistic regression classifier **(group of two only)**.

**Pre-settings:**

- From the source file use **only a fraction of 0.2** of the entire source file.
- Set stop_words='english'
- Use for min_df , max_df and ngram_range the default settings
- use an 80/20 train / test split
- use e.g., a 10 fold for cross validation
- use the source file: imdb.csv

**4. Submission/presentation**

**Each group** submits a **small report (pdf) via email** to the lecturer **1 days before presentation**, containing:

a) source file
b) curves showing C parameter vs Ø accuracy with different max_feature parameter.
c) Conclusion statement
d) Present proper curves or table of obtained results which compare the logistic regression classifier vs the SVM classifier (**group of two only**).
Conclusion statement (**group of two only**)
e) The running code is presented and explained by each group.

**Literature**

[1] https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html
[2] https://www.datacamp.com/tutorial/stemming-lemmatization-python
[3] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
[4] https://scikit-learn.org/stable/modules/feature_extraction.html#stop-words
[5] https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction