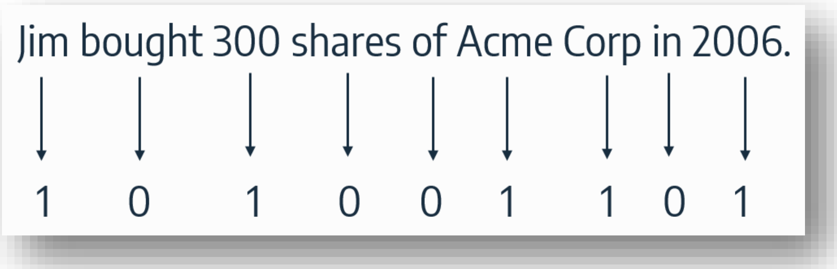
# Natural Language Processing – HW1

In this assignment you will solve the Named Entity Recognition (NER) task using manual features and sklearn.

Named Entity Recognition is a token classification task in which each token needs to be classified to a type of entity. In this assignment we will deal with an easier version where we classify each token to whether it is an entity or not, without classification into entity types.

Example



Jim bought 300 shares of Acme Corp in 2006.

**Instructions**

- Assignment will be submitted in singles or pairs.
- **Basic Part**: Use manual features to reach 90% accuracy on the eval set.
- **Competitive Part**: Train the best model you can while using only 5 manual features. Write your predictions on the test set to “competitive.txt” at the root dir. File should be written in the same format as the other files. Make sure it is exactly in the same format, without trailing spaces of lines.
- You should write readable, modular code. Document your methods. Code is 10% of grading.
- In the assignment’s directory you will find the following files:

    - Data – a directory with all your data files.
    - dataloading.py – Used for loading data and preprocessing.
    - main.py – Used for training and analysis. At submission, running this script should load the data from train.txt, train your model for the competitive part and outputs the predictions to the “competitive.txt” file. The code should run in less than 5 minutes.
- Write a report of up to 2 pages containing
    - What features did you try?
    - What models/hyperparameters/modeling choices did you try?
    - How did you select the top features for the competitive part?
    - Confusion matrix of the base model on the eval set
    - Any other results\plots you think are interesting.
- Submission until Thursday the 23.6 at 23:59 pm. You should submit a zip file with the code and the report in the following format:
    - hw1_id1_id2.zip
        - code
            - main.py
            - dataloading.py
            - competitive.txt
        - report.txt
- Grading
    - 70% base part
    - 10% code
    - 20% competitive results