# Assignment 4 - Part of Speech (POS) Tagging

In this assignment you will implement a Hidden Markov Model for Part of Speech tagging. The objectives of this assignment are:

1. Better understanding of POS
2. Better understanding of HMM and generative sequence models.

**Submission deadline**: 23:59, Tuesday, June8, 2021



## 

## Data and documentation

Extracted tokens and POS tags are taken from the English Web Treebank via the Universal Dependencies Project (<http://universaldependencies.org/>).

Tagset overviews:

\* Universal: <http://people.cs.georgetown.edu/nschneid/p/UPOS-English.pdf>

\* Penn (full new-style tagset): <https://spacy.io/docs/usage/pos-tagging#pos-tagging-english>

\* Penn (examples): <http://surdeanu.info/mihai/teaching/ista555-fall13/readings/PennTreebankTagset.html>

Data for training and testing: [training](https://www.dropbox.com/s/dmmeo0froqsw2sf/en-ud-train.upos.tsv?dl=0) and [development](https://www.dropbox.com/s/xzz799pp541z2t4/en-ud-dev.upos.tsv?dl=0) datasets.

## API and requirements:

A code skeleton with the API, documentation and hints can be found here: [tagger.py](https://www.dropbox.com/s/mq9midp83vsqav5/tagger.py?dl=0)

You can use all standard modules and packages (not requiring special installation) and the modules imported in the API file.

### **Submission guidelines**:

1. You should submit one **tar.gz** file with all relevant code files (at least tagger.py, supporting the specified API)
2. You should use Python 3.9 for the coding part.

## Integrity

As always - you could (and should) consult each other, but don’t share code.