Lab Help

# Word Embeddings First Steps: Data Preparation

In this series of ungraded notebooks, you'll try out all the individual techniques that you learned about in the lectures. Practicing on small examples will prepare you for the graded assignment, where you will combine the techniques in more advanced ways to create word embeddings from a real-life corpus.

This notebook focuses on data preparation, which is the first step of any machine learning algorithm. It is a very important step because models are only as good as the data they are trained on and the models used require the data to have a particular structure to process it properly.

To get started, import and initialize all the libraries you will need.

In [1]:

```
import re
import nltk
import emoji
import numpy as np
from nltk.tokenize import word_tokenize
from utils2 import get_dict
```

# Data preparation

In the data preparation phase, starting with a corpus of text, you will:

- Clean and tokenize the corpus.
- Extract the pairs of context words and center word that will make up the training data set for the CBOW model. The context words are the features that will be fed into the model, and the center words are the target values that the model will learn to predict.
- Create simple vector representations of the context words (features) and center words (targets) that can be used by the neural network of the CBOW model.

## Cleaning and tokenization

To demonstrate the cleaning and tokenization process, consider a corpus that contains emojis and various punctuation signs.

In [4]:

```
# Define a corpus
corpus = 'Who ❤️ "word embeddings" in 2020? I do!!!'
```

First, replace all interrupting punctuation signs — such as commas and exclamation marks — with periods.

In [5]:

```
# Print original
```