Lab Help

Word Embeddings: Ungraded Practice Notebook

In this ungraded notebook, you'll try out all the individual techniques that you learned about in the lecture. 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 is made of two main parts: data preparation, and the continuous bag-of-words (CBOW) model.

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

In [1]:

```
import sys
!{sys.executable} -m pip install emoji
```

```
Collecting emoji
  Downloading emoji-0.6.0.tar.gz (51 kB)
     |████████████████████████████████| 51 kB 14.2 MB/s eta 0:00:01
Building wheels for collected packages: emoji
  Building wheel for emoji (setup.py) ... done
  Created wheel for emoji: filename=emoji-0.6.0-py3-none-any.whl size
=49715 sha256=e5aa797e44c426f28df55ea35ba7905e8da2c5a6602e481b4468165
6e76d35a3
  Stored in directory: /home/jovyan/.cache/pip/wheels/4e/bf/6b/2e22b3
708d14bf6384f862db539b044d6931bd6b14ad3c9adc
Successfully built emoji
Installing collected packages: emoji
Successfully installed emoji-0.6.0
WARNING: You are using pip version 20.1; however, version 20.2.2 is a
vailable.
You should consider upgrading via the '/opt/conda/bin/python -m pip i
nstall --upgrade pip' command.
```

In [2]:

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

from utils2 import get_dict

nltk.download('punkt')  # download pre-trained Punkt tokenizer for Engl
ish
```

```
[nltk_data] Downloading package punkt to /home/jovyan/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
```

Out[2]:

```
True
```

Data preparation