Lab Help

N-grams Corpus preprocessing

The input corpus in this week's assignment is a continuous text that needs some preprocessing so that you can start calculating the n-gram probabilities.

Some common preprocessing steps for the language models include:

- lowercasing the text
- remove special characters
- split text to list of sentences
- split sentence into list words

Can you note the similarities and differences among the preprocessing steps shown during the Course 1 of this specialization?

In [1]:
```
import nltk                     # NLP toolkit
import re                       # Library for Regular expression operations

nltk.download('punkt')          # Download the Punkt sentence tokenizer
```

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

Out[1]:
```
True
```

Lowercase

Words at the beginning of a sentence and names start with a capital letter. However, when counting words, you want to treat them the same as if they appeared in the middle of a sentence.

You can do that by converting the text to lowercase using [str.lowercase] (https://docs.python.org/3/library/stdtypes.html?highlight=split#str.lower (https://docs.python.org/3/library/stdtypes.html?highlight=split#str.lower)).

In [2]:
```
# change the corpus to lowercase
corpus = "Learning% makes 'me' happy. I am happy be-cause I am learning! :)"
corpus = corpus.lower()

# note that word "learning" will now be the same regardless of its position in the sentence
print(corpus)
```

```
learning% makes 'me' happy. i am happy be-cause i am learning! :)
```