Python Tutorial 3

LING-381-Language Technology and LLMs

Instructor: Hakyung Sung

September 18, 2025

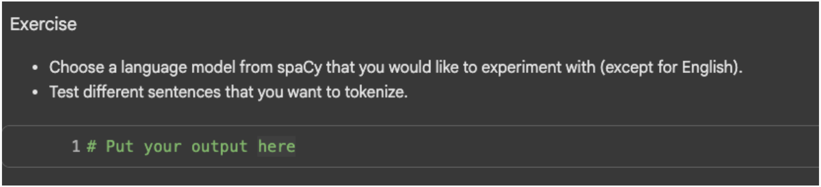
Review: Tutorial 2

- Tokenization
- Lemmatization
- Frequency calculation
- Concordance

Review: Tokenization

- What is tokenization?
- What were the approaches that we’ve used in the tutorial?

Review: Tokenization-Exercise



Exercise

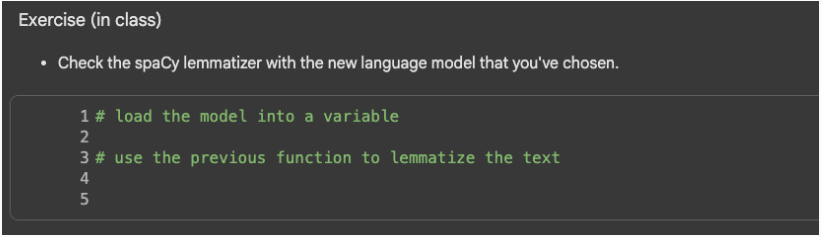
- Choose a language model from spaCy that you would like to experiment with (except for English).
- Test different sentences that you want to tokenize.

```
# Put your output here
```

Review: Lemmatization

- What is lemmatization?
- What were the approaches that we’ve used in the tutorial?

Review: Lemmatization-Exercise



Exercise (in class)

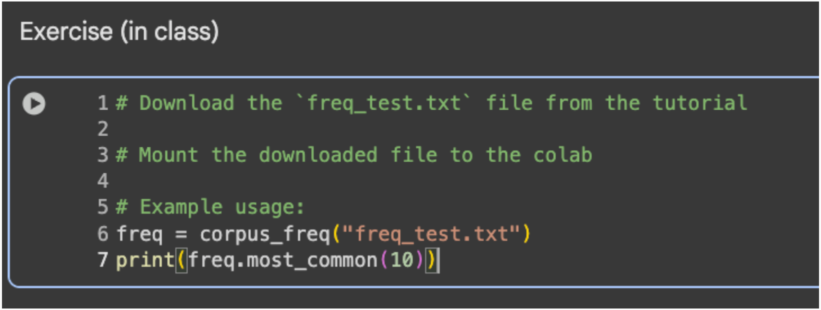
- Check the spaCy lemmatizer with the new language model that you've chosen.

```
# load the model into a variable

# use the previous function to lemmatize the text


```

Review: Frequency calculation

Word frequency calculation is most often the very first step of the text analysis.

Review: Frequency calculation-Exercise

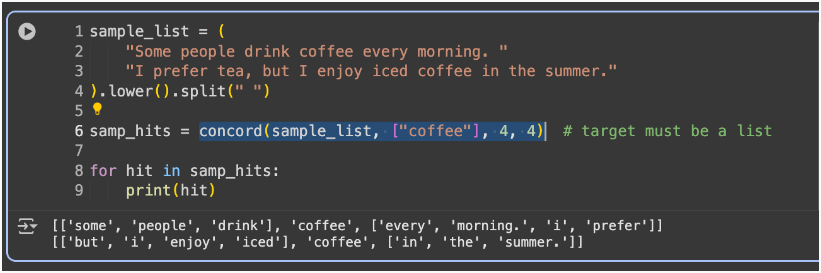


Exercise (in class)

```
# Download the `freq_test.txt` file from the tutorial

# Mount the downloaded file to the colab

# Example usage:
freq = corpus_freq("freq_test.txt")
print(freq.most_common(10))
```

Review: Concordance

Concordancing is a method in corpus linguistics that involves identifying and displaying occurrences of a target linguistic item along with its surrounding context.



```python
sample_list = (
    "Some people drink coffee every morning. "
    "I prefer tea, but I enjoy iced coffee in the summer."
).lower().split(" ")

samp_hits = concord(sample_list, ["coffee"], 4, 4)  # target must be a list

for hit in samp_hits:
    print(hit)
```

```
[['some', 'people', 'drink'], 'coffee', ['every', 'morning.', 'i', 'prefer']]
[['but', 'i', 'enjoy', 'iced'], 'coffee', ['in', 'the', 'summer.']]
```

Let’s check the exercise together!

Goal for today’s class

- Done with Tutorial 2 →Work with Collocation (Bonus 2 points)
- Want to fix your previous submission Tutorial 2 →Redo and resubmit
- Want to work on Tutorial 2 →Do and Submit

Submission deadline: Tomorrow 11:59PM

Preview: Collocation analysis

- **What:** Collocations = word pairs/phrases that co-occur more than chance

Preview: Collocation analysis

- **What:** Collocations = word pairs/phrases that co-occur more than chance
- **Why:** Studying collocations via n-gram analysis helps to (1) assess native-like fluency (e.g., *make a decision* vs. *do a decision*) and (2) extract informative n-grams for information retrieval.

Preview: Collocation analysis

- **What:** Collocations = word pairs/phrases that co-occur more than chance
- **Why:** Studying collocations via n-gram analysis helps to (1) assess native-like fluency (e.g., *make a decision* vs. *do a decision*) and (2) extract informative n-grams for information retrieval.
- **Break down steps**:

Preview: Collocation analysis

- **What:** Collocations = word pairs/phrases that co-occur more than chance
- **Why:** Studying collocations via n-gram analysis helps to (1) assess native-like fluency (e.g., *make a decision* vs. *do a decision*) and (2) extract informative n-grams for information retrieval.
- **Break down steps**:
    - Tokenize the text

Preview: Collocation analysis

- **What:** Collocations = word pairs/phrases that co-occur more than chance
- **Why:** Studying collocations via n-gram analysis helps to (1) assess native-like fluency (e.g., *make a decision* vs. *do a decision*) and (2) extract informative n-grams for information retrieval.
- **Break down steps**:
    - Tokenize the text
    - Generate word frequency across the entire corpus

Preview: Collocation analysis

- **What:** Collocations = word pairs/phrases that co-occur more than chance
- **Why:** Studying collocations via n-gram analysis helps to (1) assess native-like fluency (e.g., *make a decision* vs. *do a decision*) and (2) extract informative n-grams for information retrieval.
- **Break down steps**:
  - Tokenize the text
  - Generate word frequency across the entire corpus
  - Generate word frequency within contexts

Preview: Collocation analysis

- **What:** Collocations = word pairs/phrases that co-occur more than chance
- **Why:** Studying collocations via n-gram analysis helps to (1) assess native-like fluency (e.g., *make a decision* vs. *do a decision*) and (2) extract informative n-grams for information retrieval.
- **Break down steps**:
  - Tokenize the text
  - Generate word frequency across the entire corpus
  - Generate word frequency within contexts
  - Calculate association strength between word pairs using statistical measures (e.g., Mutual information [MI])