Text Normalization and Linguistic Features

Lecture 13: Advanced Text Preprocessing

Harito

September 15, 2025

Recap: Tokenization

In Lab 1, we learned about tokenization: breaking text into words or subwords.

- ‘”Running is fun!”‘ $\rightarrow$ ‘[”Running”, ”is”, ”fun”, ”!”]‘

However, words can appear in many different forms (e.g., ”run”, ”running”, ”runs”, ”ran”). For many NLP tasks, we want to treat these as the same underlying word.

Stemming

Definition

Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form.

Characteristics:

- Often involves simply chopping off suffixes.
- The resulting stem may not be a valid word.
- Faster and simpler than lemmatization.

Example (Porter Stemmer):

- "running", "runs", "ran" $\rightarrow$ "run"
- "beautiful", "beauty" $\rightarrow$ "beauti"

Commonly used algorithm: **Porter Stemmer** (available in NLTK).

Lemmatization

Definition

Lemmatization is the process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form.

Characteristics:

- Uses a vocabulary and morphological analysis.
- The resulting lemma is always a valid word.
- More accurate but generally slower than stemming.

Example:

- "running", "runs", "ran" $\rightarrow$ "run"
- "better" $\rightarrow$ "good"
- "geese" $\rightarrow$ "goose"

Often performed using libraries like **spaCy** or **NLTK**.

Stemming vs. Lemmatization

Stemming

- Faster
- Simpler rules
- Output may not be a real word
- Less accurate for complex morphology

Lemmatization

- Slower
- Uses vocabulary & morphology
- Output is always a real word
- More accurate

Choice depends on task: Stemming for IR (search), Lemmatization for deeper linguistic analysis.

Part-of-Speech (POS) Tagging

Definition

Part-of-Speech (POS) Tagging is the process of marking up a word in a text as corresponding to a particular part of speech, based on both its definition and its context.

Examples of POS Tags:

- ‘NN‘: Noun, singular or mass (e.g., ”cat”, ”air”)
- ‘VB‘: Verb, base form (e.g., ”run”, ”eat”)
- ‘JJ‘: Adjective (e.g., ”happy”, ”big”)
- ‘RB‘: Adverb (e.g., ”quickly”, ”very”)
- ‘PRP‘: Personal pronoun (e.g., ”I”, ”you”, ”he”)

Importance:

- Syntactic analysis (parsing)
- Named Entity Recognition (NER)
- Word Sense Disambiguation
- Feature engineering for ML models

Using NLTK and spaCy

NLTK (Natural Language Toolkit) is a leading platform for building Python programs to work with human language data.

```
import nltk
from nltk.stem import PorterStemmer

# Download necessary data (run once)
# nltk.download('punkt')
# nltk.download('wordnet')

stemmer = PorterStemmer()
print(stemmer.stem('running')) # Output: run
```

spaCy is an industrial-strength NLP library.

```
import spacy

# Load model (run `python -m spacy download en_core_web_sm`
nlp = spacy.load('en_core_web_sm')
doc = nlp("running better")
```

Next Steps

Time for Lab 13!

Objective:

- Implement stemming using NLTK.
- Implement lemmatization using spaCy.
- Implement POS tagging using spaCy.