LING 331
Text Processing for Linguists

Week 8

—

Python for Text

Class next quarter, 334

- Diving much more deeply into implementing real algorithms for text processing, applications and analysis
- Let me know ASAP if you want to take!
- You are equipped to do it, though it's a jump in challenge
- (Sorry if I owe you an email on it)

Timeline for the Rest of the Quarter!

- Today:
  - Assignment 5 Review
  - External Libraries in Python
  - Assignment 6 Preview (due 3/3)

- Next Week: Research Chat,
  more on text applications in Python

- 3/7, 3/9: Final project working time

Assignment 5 Review

- Readability - “Pythonic”?
  - ' '.join([c for c in s if c.isalpha()])
  - Could be written out as a for loop, etc
  - Personal preference!
  - (it is mine though)

Assignment 5 Review

- Permissions problem!
    - Everyone please go to your A5 directory, and run:

```
chmod g+w *
```

    - Do the same for A6 please!

Assignment 5 Review

- word_counts
  - Note on Counter, this works too:
    - counts = Counter(tokenize(s))

Assignment 5 Review

- syllable_count, final_syllable
    - Follow the given definition!
      Space-separated phoneme ending in a digit.
    - You can “make it work” other ways,
      but might run into hard-to-foresee edge cases

Assignment 5 Review

- words_rhyme
    - Put the return statement in the deepest point of nesting

```
for p1 in cmudict[word1]:
    for p2 in cmudict[word2]:
        if final_syllable(p1, True) == final_syllable(p2, True):
            return True

return False # [RV: equivalent to saying, “if we haven’t
                    already returned”]
```

Assignment 5 Review

- flesch_reading_ease
  - There was a subtlety here I could have spelled out more clearly.
  - We want to calculate two quantities:
    - average_line_length
    - average_syllables_per_word
  - For average syllables, what do we have to say about words for which we don't have pronunciation?
    - Nothing!

Assignment 5 Review

- flesch_reading_ease
    - Further word on this:
      is this actually “readability”?
    - It is validated to some degree,
      and correlates with human judgments of readability
    - Nevertheless, it is an oversimplification -
      often we say, a particular “operationalization”
    - Always look at these critically!!!
      Are we missing important things in your context?

Assignment 5 Review

- Calculating adjacent lines that rhyme
    - Requires some more complex “spatial awareness”
    - Let’s work through it together

Assignment 5 Review

- Calculating delta dictionaries
    - Use subtract method on Counter!
    - Have to look it up - note that it is “in-place” and does not return anything

External Libraries for Text Processing

- A6 looking at spaCy and NLTK
- All kinds of useful functionality!

NAME	DESCRIPTION
Tokenization	Segmenting text into words, punctuations marks etc.
Part-of-speech (POS) **Tagging**	Assigning word types to tokens, like verb or noun.
Dependency Parsing	Assigning syntactic dependency labels, describing the relations between individual tokens, like subject or object.
Lemmatization	Assigning the base forms of words. For example, the lemma of “was” is “be”, and the lemma of “rats” is “rat”.
Sentence Boundary Detection (SBD)	Finding and segmenting individual sentences.
Named Entity Recognition (NER)	Labelling named “real-world” objects, like persons, companies or locations.
Entity Linking (EL)	Disambiguating textual entities to unique identifiers in a knowledge base.
Similarity	Comparing words, text spans and documents and how similar they are to each other.
Text Classification	Assigning categories or labels to a whole document, or parts of a document.

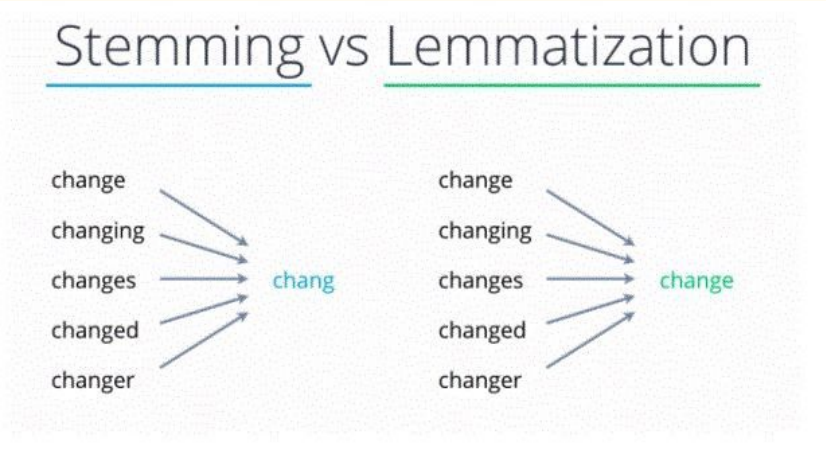
External Libraries for Text Processing

- In this assignment we'll be using their tokenizer, stemmer, lemmatizer, and part-of-speech tagger

- Always important to refer to documentation when using external libraries!

Stemming and Lemmatization

- Forms of text normalization, reduces sparsity



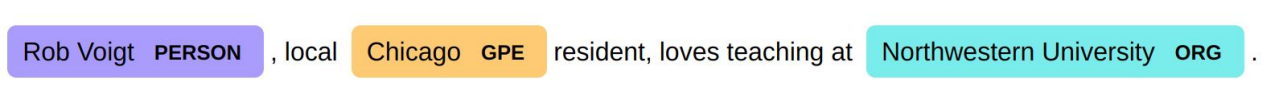
https://www.turing.com/kb/stemming-vs-lemmatization-in-python

Word	Stemming	Lemmatization
information	inform	information
informative	inform	informative
computers	comput	computer
feet	feet	foot

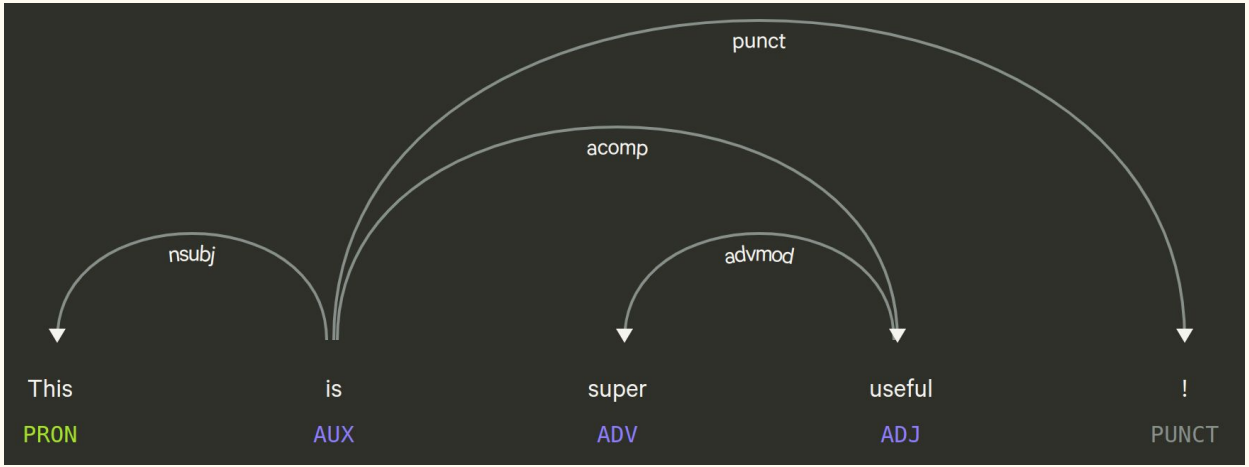
https://studymachinelearning.com/stemming-and-lemmatization/

External Libraries for Text Processing

- More complex functionality often useful, like entity tagging:



- and dependency parsing:



Assignment 6 Preview

- Let’s check it out - tricky install bits to start!