Ling 131A
Introduction to NLP with Python

Tokenization

Marc Verhagen, Fall 2017

Contents

- Python sessions
- Final project
- Assignment 4
- Regular Expressions in Python
- Tokenization

Python sessions

- From the registrar:
  - Shapiro Science Center GL 14 has been reserved for LING 131a review sessions on Mondays and Wednesdays from 1-1:50pm during the semester while classes are in session with the exception of Nov 21 (Thanksgiving).
- Start tomorrow

Final Project

- Start thinking about a project
- Project proposal will be due on Tuesday November 26th
  - Submit to me by email
- Groups of up to 4 people
- Should involve some serious coding beyond what you have done so far
- Delivery vehicle is again GitHub
- Graded on code and final report

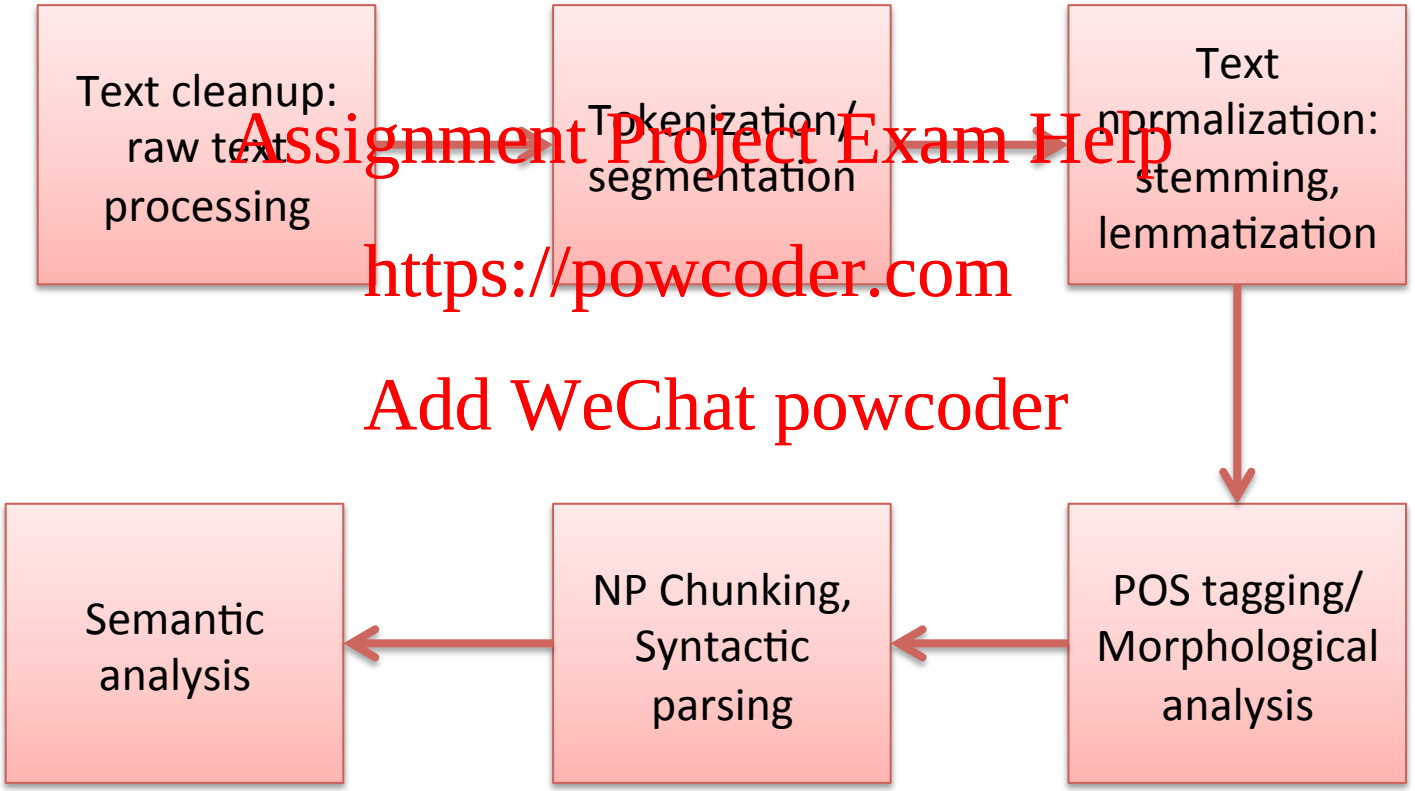
Final Project Topics

- Project does not need to involve NLTK
- Programming heavy
  - Create your own tokenizer, lemmatizer, POS tagger, syntactic parser, word aligner, entity extractor or some other NLP module
- Linguistics heavy
  - Analyze a corpus of data or compare a couple of corpora
  - Should go beyond running some NLTK code

Final Project Topics

- Examples from last year
  - Natural language interface to relational database
  - Haiku bot
  - Twitter classification
  - Text normalization (currencies)
  - Detecting challenging or uncommon words
  - question-answering system
  - Spanish chat box
  - Generating rhyme patterns

NLP Pipeline



Tokenization

- First level of abstraction
- What are the basic units in your text
- Knowing that '*the*' is the same thing whether it occurs as '*the*' or '*The*'
- Usually does not include normalization
  - *The man in the High Tower.*
  - Tokenized as
    - *The man in the High Tower .*
  - Not as
    - *the man in the high tower .*

Penn Treebank Tokenization

- Lorrillard Inc., the unit of New York-based Loews Corp. that makes Kent cigarettes, stopped using crocidolite in its Micronite cigarette filters in 1956.
- Typically, money-fund yields beat comparable invest-ments because portfolio managers can vary maturities and go after highest rates.

- Periods and hyphens are ambiguous
- Hyphens originally not considered single tokens in PTB, but later revised to context dependent tokenization.

Penn Treebank Tokenization

- Lorrillard Inc. , the unit of New York - based Loews Corp. that makes Kent cigarettes , stopped using crocidolite in its Micornite cigarette filters in 1956 .
- Typically, money-fund yields beat comparable investments because portfolio managers can vary maturities and go after highest rates .

- Periods and hyphens are ambiguous
- Hyphens originally not considered single tokens in PTB, but later revised to context dependent tokenization.

Penn Treebank Tokenization

- In the new position he will oversee Mazda **'s** U.S. sales , services , parts and marketing operations .
- We did **n't** have much of a choice .
- U.S. trade officials said the Philippines and Thailand would be the main beneficiaries of the president **'s** action .
- Anything **'s** possible -- how about the new Guinea Fund ?

- Contractions are separated out

Penn Treebank Tokenization

- Assets of the 400 taxable funds grew by $1.5 billion during the latest week.
- Exports in October stood $5.29 billion, a mere 0.7% increase from a year earlier, while imports increased sharply to $5.39 billion, up 20% from last year.

Penn Treebank Tokenization

- Assets of the 400 taxable funds grew by $ 1.5 billion during the latest week .
- Exports in October stood $ 5.29 billion , a mere 0.7 % increase from a year earlier , while imports increased sharply to $ 5.39 billion , up 20 % from last year .

- Punctuation marks are their own tokens, and not just periods and commas

Penn Treebank Tokenization

- The federal government suspended sales of the U.S. savings bonds because Congress hasn't lifted the ceiling on government debt.
- The Treasury said the U.S. will default on Nov. 9 if Congress doesn't act by then.

Penn Treebank Tokenization

- The federal government suspended sales of the U.S. savings bonds because Congress has **n't** lifted the ceiling on government debt **.**
- The Treasury said the U.S. will default on Nov. 9 if Congress does n't act by then .

- Contractions are separated out

Stemming and Lemmatization

- The process of reducing inflected words to their word stem or root form (aka lemma)
- Strongly related, but there are differences
  - Stemmer usually does not take the context into account and does not care about a word's category
  - Lemmatizer includes dictionary lookup of wordforms and uses the context of a word
    - for example, *meeting* should be mapped to *meet* if it is a verb but not if it is a noun
  - Stemming is faster compared with lemmatization, but can't ensure the resulting words are legitimate

Stemming and Lemmatization

- Regular expressions often used for suffix stripping, but
  - Can't elegantly handle irregular patterns, e.g., women → woman
  - Off-the-shelf stemmers (Lancaster and Porter) use many special rules to handle these irregularities

Tokenization in NLTK

```
from nltk import word_tokenize

text1 = """Lorrillard Inc., the unit of New York-based Loews
Corp. that makes Kent cigarettes, stopped using crocidolite in
its Micornite cigarette filters in 1956."""

print(word_tokenize(text1))
```

Stemming and Lemmatizing

```
from nltk.stem.porter import PorterStemmer
from nltk.stem import WordNetLemmatizer

words = ['cars', 'feet', 'cats', 'dies', 'runs', 'denied',
         'died', 'agreed', 'owned', 'humbled', 'sized',
         'meeting', 'stating', 'siezing', 'itemization',
         'sensational', 'traditional', 'reference',
         'colonizer', 'plotted']

stemmer = PorterStemmer()
print([stemmer.stem(w) for w in words])

lemmatizer = WordNetLemmatizer()
print([lemmatizer.lemmatize(w) for w in words])
```