

LING 520: Computational Analysis of English

Semester: FALL '16

Instructor: Sowmya Vajjala

Iowa State University, USA

8 September 2016

Class outline

- ▶ Regular expressions review
- ▶ Text preprocessing: tasks
- ▶ Text preprocessing: tokenizing
- ▶ Text preprocessing: sentence splitting
- ▶ Writing your own tokenizer and sentence splitter
- ▶ NLTK tokenizers and Sentence splitters.

Tuesday's exercise solution discussion

Regular Expressions in Real-world

- ▶ Look at some regular expressions in Eliza program
- ▶ Chris Manning's talk about the use of regular expressions in Stanford Tokenizer

Different Text Pre-processing tasks

Several pre-processing tasks are performed on corpora, depending on what you want.

1. lower casing
2. **tokenization, sentence splitting**
3. removing most frequent, or most rare words
4. normalizing contractions, abbreviations, words with social media like spellings (happyyyyyyyyyy!) etc.
5. spelling normalization and correction

Text preprocessing: Tokenizing

What does tokenizing mean?

- ▶ Splitting a text into tokens (words, punctuation markers, etc.)
- ▶ Example: After tokenizing, "I have a sentence!!" becomes a list of tokens: I, have, a, sentence, !, !
- ▶ Almost all NLP tasks require this as a pre-processing task.
- ▶ While tokenization looks like a simple task, there are several issues in designing one.

What is the big deal about tokenizing?

Issues to consider in a tokenizer

- ▶ How many tokens are there in this sentence: "Hmm, I worry uh a lot about next week."

What is the big deal about tokenizing?

Issues to consider in a tokenizer

- ▶ How many tokens are there in this sentence: "Hmm, I worry uh a lot about next week."
- ▶ Should C.N.N be one token as CNN, or three word tokens?

What is the big deal about tokenizing?

Issues to consider in a tokenizer

- ▶ How many tokens are there in this sentence: "Hmm, I worry uh a lot about next week."
- ▶ Should C.N.N be one token as CNN, or three word tokens?
- ▶ Should phone numbers: 555-333-222 be split into three tokens or one (don't forget it is not written like this all over the world)

What is the big deal about tokenizing?

Issues to consider in a tokenizer

- ▶ How many tokens are there in this sentence: "Hmm, I worry uh a lot about next week."
- ▶ Should C.N.N be one token as CNN, or three word tokens?
- ▶ Should phone numbers: 555-333-222 be split into three tokens or one (don't forget it is not written like this all over the world)
- ▶ Mr. Anderson - one token or two?

What is the big deal about tokenizing?

Issues to consider in a tokenizer

- ▶ How many tokens are there in this sentence: "Hmm, I worry uh a lot about next week."
- ▶ Should C.N.N be one token as CNN, or three word tokens?
- ▶ Should phone numbers: 555-333-222 be split into three tokens or one (don't forget it is not written like this all over the world)
- ▶ Mr. Anderson - one token or two?
- ▶ Doesn't - one or two?

What is the big deal about tokenizing?

Issues to consider in a tokenizer

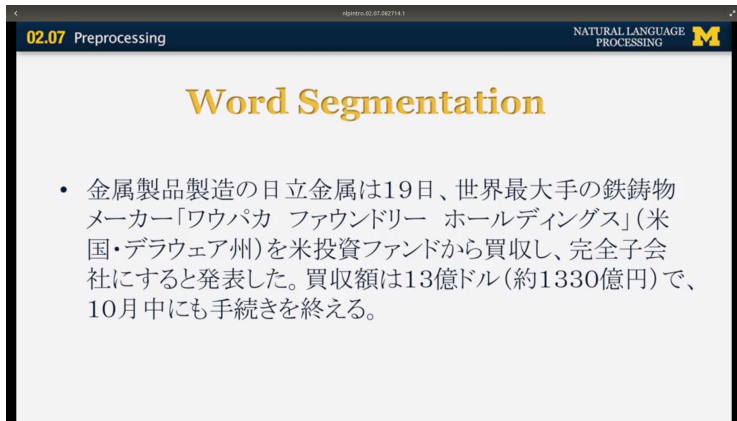
- ▶ How many tokens are there in this sentence: "Hmm, I worry uh a lot about next week."
- ▶ Should C.N.N be one token as CNN, or three word tokens?
- ▶ Should phone numbers: 555-333-222 be split into three tokens or one (don't forget it is not written like this all over the world)
- ▶ Mr. Anderson - one token or two?
- ▶ Doesn't - one or two?
- ▶ Agent Smith's Matrix - how many tokens are there in Agent Smith's?

What is the big deal about tokenizing?

Issues to consider in a tokenizer

- ▶ URLs: Should they be considered single token? or split at every underscores, slash etc?
- ▶ Chicago-Des Moines flight: If we split this on space, Chicago-Des Moines is one token.
- ▶ But splitting on - separates part-time which is one token.
- ▶ Some words are compound words (like with some long German nouns). What will we use to split such words?

Tokenizing for non-English languages: Japanese



The screenshot shows a presentation slide with a dark blue header. On the left, it says '02.07 Preprocessing'. On the right, it says 'NATURAL LANGUAGE PROCESSING' next to a yellow 'M' logo. The main title 'Word Segmentation' is in large yellow font. Below it is a bulleted list in Japanese. At the bottom right of the slide, there are navigation icons.

02.07 Preprocessing NATURAL LANGUAGE PROCESSING M

Word Segmentation

- 金属製品製造の日立金属は19日、世界最大手の鉄鋳物メーカー「ワウパカ ファウンドリー ホールディングス」(米国・デラウェア州)を米投資ファンドから買収し、完全子会社にすると発表した。買収額は13億ドル(約1330億円)で、10月中にも手続きを終える。

(Source: Radev's coursera course, Week 2, lecture on pre-processing)

Tokenizing: Conclusion

Tokenizing is not as easy as it seems. But, a lot of English tokenizing issues can be handled by current day tokenizers (with heavy use of regular expressions). Customized tokenizers for different types of data too exist. For languages such as Japanese and Chinese, there are existing tools (for relatively Standard form of language). Here is an example tokenizer code:

<http://goo.gl/vJQTAz>

Text preprocessing: Sentence Splitting

What does sentence splitting do?

- ▶ Splits the text into sentences.
- ▶ Sentence splitting is essential to do any higher order NLP task such as parsing, discourse and semantic analysis.
- ▶ Typically done by writing regular expressions, constructing decision trees of rules, and using machine learning methods (more on this last one later)
- ▶ Issues: Not all sentences end in a full stop. Some have other punctuation markers. Some punctuation can be ambiguous.
- ▶ One code that can learn sentence splitting by itself from large amounts of data: http://www.nltk.org/_modules/nltk/tokenize/punkt.html

What is the big deal about sentence splitting?

Just splitting in full-stop or ? or ! will not do.

- ▶ This sentence: "There are several methods such as A, B, C etc., but there is no best method yet. It is a work in progress."
- ▶ People don't follow conventions or grammar sometimes.
Missing capitalization at the start of a sentence, not leaving a space after sentence breaker etc.
- ▶ Spoken language, tweets etc - do not follow same conventions as news articles. This diversity may affect the accuracy of our sentence splitting rules.

Sentence splitting conclusion

While sentence splitting for English in its standard usage is a good, there are some issues with learner texts and other non-canonical forms. Good sentence splitters exist for English like languages. About others: Figure out.

Question: If English did not have capitalization, how easy or difficult would this task be?

Practice Exercises

Practice with general Python

- ▶ Write a tokenizer that splits a sentence into tokens, based on your definition of tokens, and using regular expressions.
- ▶ Write a sentence splitter, that splits a text into sentences, using regular expressions.
- ▶ Test both your tokenizer and sentence splitter by giving some noisy text (tweets, or non-native language, speech transcripts etc.)

Practice with NLTK

- ▶ There are a couple of word tokenizer implementations in NLTK. Explore the differences between the following tokenizers you can access from `nltk.tokenize` package: `TreebankWordTokenizer`, `StringTokenizer`, `TweetTokenizer`, `MWETokenizer`, `RegExpTokenizer` (Take a collection of common sentences and compare how they perform).
- ▶ What is the difference between `WordPunktTokenizer` and `PunktWordTokenizer`?
- ▶ Figure out how to do sentence splitting in Python using NLTK, and how many options exist.
- ▶ Finally: If you are interested, explore tokenization/sentence splitting options in NLTK for a non-English language (from documentation)

Next Week

- ▶ Topics: Spelling Errors, Normalization, Morphological analysis
- ▶ Readings: Chapter 2 in J &M, Chapter 3 in NLTK Book
- ▶ Video lectures: Week 2, and first 2 lectures in Week 3 from Radev's coursera course.
- ▶ Assignment 1 - submit before midnight on saturday
- ▶ I will try to grade before tuesday's class as much as possible.