NLP Python Basics

Natural Language Processing Bootcamp

- Section Goals
  - Set up Spacy and Language Library
  - Understand Basic NLP Topics
    - Tokenization
    - Stemming
    - Lemmatization
    - Stop Words
  - Spacy for Vocabulary Matching

Natural Language Processing Bootcamp

- We will also have a few introductory lectures to discuss common libraries such as NLTK and Spacy.
- As well as a general discussion of what Natural Language Processing is.

Spacy Setup

Natural Language Processing Bootcamp

- What is Spacy?
  - Open Source Natural Language Processing Library.
  - Designed to effectively handle NLP tasks with the most efficient implementation of common algorithms.

Natural Language Processing Bootcamp

- What is Spacy?
  - For many NLP tasks, Spacy only has one implemented method, choosing the most efficient algorithm currently available.
  - This means you often don’t have the option to choose other algorithms.

Natural Language Processing Bootcamp

- What is NLTK?
  - NLTK - Natural Language Toolkit is a very popular open source.
  - Initially released in 2001, it is much older than Spacy (released 2015).
  - It also provides many functionalities, but includes less efficient implementations.

Natural Language Processing Bootcamp

- NLTK vs Spacy
  - For many common NLP tasks, Spacy is much faster and more efficient, at the cost of the user not being able to choose algorithmic implementations.

Natural Language Processing Bootcamp

- NLTK vs Spacy
  - However, Spacy does not include pre-created models for some applications, such as sentiment analysis, which is typically easier to perform with NLTK.

Natural Language Processing Bootcamp

- NLTK vs Spacy
  - In this course, due to Spacy’s state of the art approach and efficiency, we will focus on Spacy, but use NLTK when it is easier to use.
  - By the end of the course, you should feel comfortable utilizing both libraries when they are best suited for a task.

Natural Language Processing Bootcamp

- NLTK vs Spacy Processing Tests

	ABSOLUTE (MS PER DOC)			RELATIVE (TO SPACY)		
SYSTEM	TOKENIZE	TAG	PARSE	TOKENIZE	TAG	PARSE
spaCy	0.2ms	1ms	19ms	1x	1x	1x
CoreNLP	0.18ms	10ms	49ms	0.9x	10x	2.6x
ZPar	1ms	8ms	850ms	5x	8x	44.7x
NLTK	4ms	443ms	*n/a*	20x	443x	*n/a*

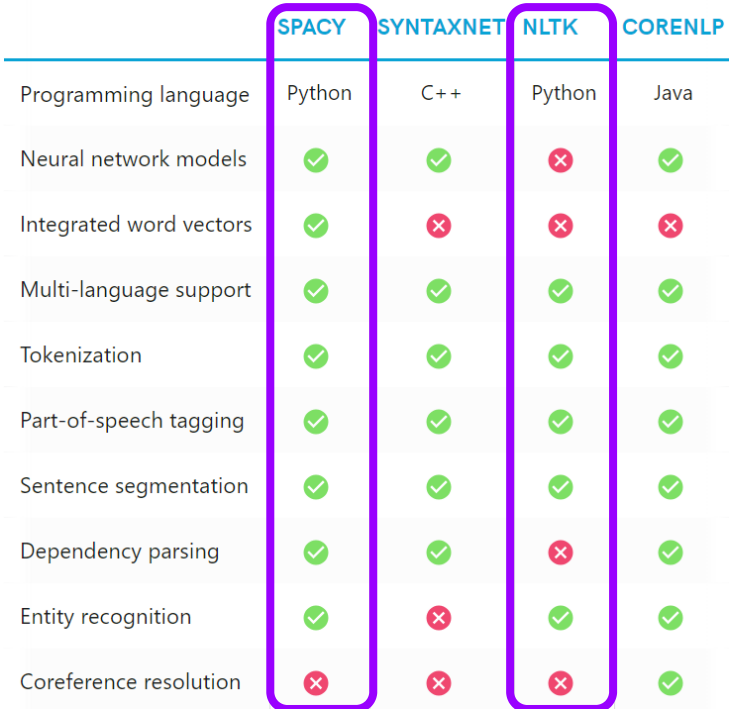
Natural Language Processing Bootcamp

https://spacy.io/usage/facts-figures

	SPACY	SYNTAXNET	NLTK	CORENLP
Programming language	Python	C++	Python	Java
Neural network models	✓	✓	✗	✓
Integrated word vectors	✓	✗	✗	✗
Multi-language support	✓	✓	✓	✓
Tokenization	✓	✓	✓	✓
Part-of-speech tagging	✓	✓	✓	✓
Sentence segmentation	✓	✓	✓	✓
Dependency parsing	✓	✓	✗	✓
Entity recognition	✓	✗	✓	✓
Coreference resolution	✗	✗	✗	✓

Natural Language Processing Bootcamp

https://spacy.io/usage/facts-figures



	SPACY	SYNTAXNET	NLTK	CORENLP
Programming language	Python	C++	Python	Java
Neural network models	✓	✓	✗	✓
Integrated word vectors	✓	✗	✗	✗
Multi-language support	✓	✓	✓	✓
Tokenization	✓	✓	✓	✓
Part-of-speech tagging	✓	✓	✓	✓
Sentence segmentation	✓	✓	✓	✓
Dependency parsing	✓	✓	✗	✓
Entity recognition	✓	✗	✓	✓
Coreference resolution	✗	✗	✗	✓

Natural Language Processing Bootcamp

What is NLP?

What is NLP?

- According to wikipedia, "**Natural language processing (NLP)** is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.

Source: https://en.wikipedia.org/wiki/Natural_language_processing

Natural Language Processing Bootcamp

- Often when performing analysis, lots of data is numerical, such as sales numbers, physical measurements, quantifiable categories.
- Computers are very good at handling direct numerical information.
- But what do we do about **text data?**

Natural Language Processing Bootcamp

- As humans we can tell there is a plethora of information inside of text documents.
- But a computer needs specialized processing techniques in order to “understand” raw text data.
- Text data is highly unstructured and can be in multiple languages!

Natural Language Processing Bootcamp

- Example Use Cases:
  - Classifying Emails as Spam vs Legitimate
  - Sentiment Analysis of Text Movie Reviews
  - Analyzing Trends from written customer feedback forms.
  - Understanding text commands, “Hey Google, play this song”.

Natural Language Processing Bootcamp

- Natural Language Processing is constantly evolving and great strides are made every month !
- We will focus on fundamental ideas that all state of the art techniques are based off.
- Let’s begin by learning about the basics of using the Spacy library.

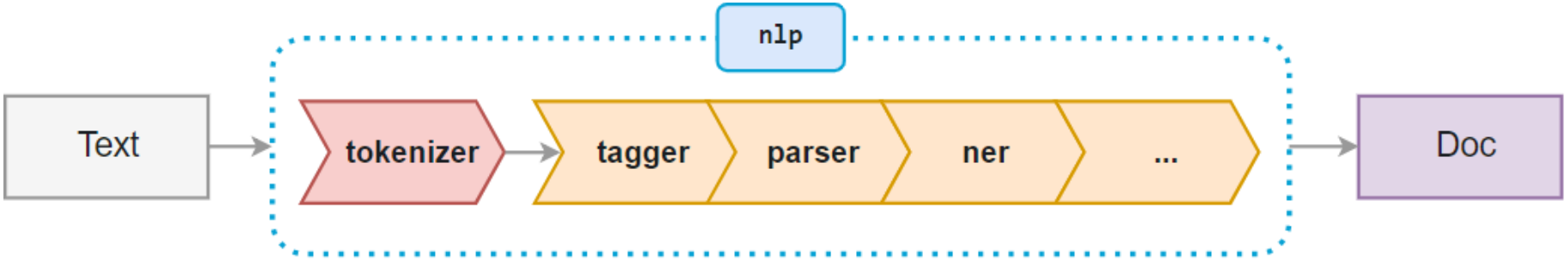
Spacy Basics

Natural Language Processing Bootcamp

- There are a few keys steps for working with Spacy that we will cover in this lecture:
  - Loading the Language Library
  - Building a Pipeline Object
  - Using Tokens
  - Parts-of-Speech Tagging
  - Understanding Token Attributes

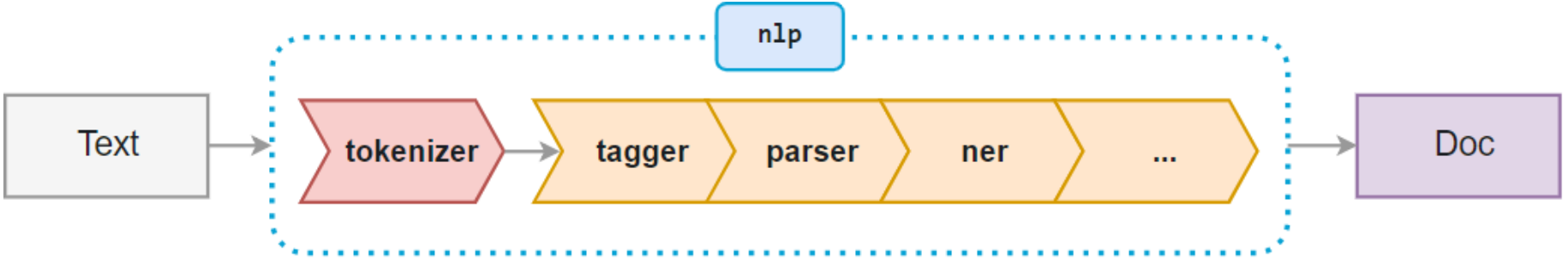
Natural Language Processing Bootcamp

- Spacy works with a Pipeline object



Natural Language Processing Bootcamp

- The **nlp()** function from Spacy automatically takes raw text and performs a series of operations to tag, parse, and describe the text data.



Natural Language Processing Bootcamp

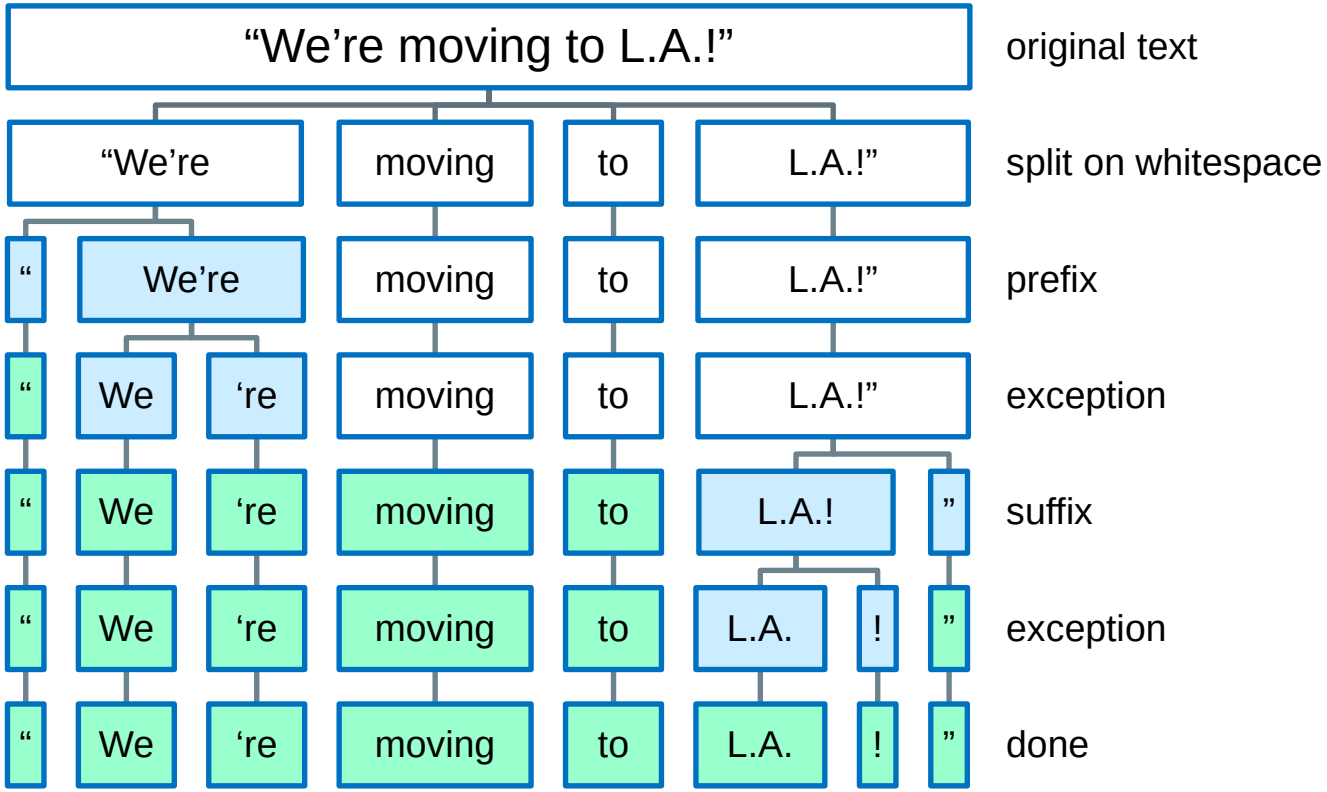
- Let’s discover the pipeline object and its series of operations.
- In subsequent lectures dive deeper into each of these aspects of NLP and Spacy (e.g. Tokenization, POS, Stemming, Lemmatization, etc…)

Tokenization

Natural Language Processing Bootcamp

- Tokenization is the process of breaking up the original text into component pieces (tokens).

Tokenization



Natural Language Processing Bootcamp

- Notice that tokens are pieces of the original text.
- We don't see any conversion to word stems or lemmas (base forms of words) and we haven't seen anything about organizations/places/money etc.

Natural Language Processing Bootcamp

- Tokens are the basic building blocks of a Doc object - everything that helps us understand the meaning of the text is derived from tokens and their relationship to one another.

Tokenization

- **Prefix:** Character(s) at the beginning
- **Suffix:** Character(s) at the end
- **Infix:** Character(s) in between
- **Exception:** Special-case rule to split a string into several tokens or prevent a token from being split when punctuation rules are applied

$ (“ ¿

km) , . !

”-- / …

let’s

U.S.

Natural Language Processing Bootcamp

- Tokens have a variety of useful attributes and methods.
- Let’s explore tokens with Spacy in further detail!

Stemming

Natural Language Processing Bootcamp

- Often when searching text for a certain keyword, it helps if the search returns variations of the word.
- For instance, searching for "boat" might also return "boats" and "boating". Here, "boat" would be the stem for [boat, boater, boating, boats].

Natural Language Processing Bootcamp

- Stemming is a somewhat crude method for cataloging related words; it essentially chops off letters from the end until the stem is reached.
- This works fairly well in most cases, but unfortunately English has many exceptions where a more sophisticated process is required.

Natural Language Processing Bootcamp

- In fact, spaCy doesn't include a stemmer, opting instead to rely entirely on lemmatization.
- There is a link in the notebook to a discussion on the maintainers of Spacy deciding on not including a Stemmer (in favor of Lemmatization)

Natural Language Processing Bootcamp

- Because of this decision to not include Stemming in Spacy, we will jump over to using NLTK and learn about various Stemmers.
- We’ll show both the Porter Stemmer and the Snowball Stemmer.

Natural Language Processing Bootcamp

- One of the most common - and effective - stemming tools is Porter's Algorithm developed by Martin Porter in 1980.
- The algorithm employs five phases of word reduction, each with its own set of mapping rules.

Natural Language Processing Bootcamp

- In the first phase, simple suffix mapping rules are defined, such as:

S1		S2	word		stem
SSES	→	SS	caresses	→	caress
IES	→	I	ponies	→	poni
			ties	→	ti
SS	→	SS	caress	→	caress
S	→		cats	→	cat

Natural Language Processing Bootcamp

- From a given set of stemming rules only one rule is applied, based on the longest suffix S1. Thus, caresses reduces to caress but not cares.

S1		S2	word		stem
SSES	→	SS	caresses	→	caress
IES	→	I	ponies ties	→ →	poni ti
SS	→	SS	caress	→	caress
S	→		cats	→	cat

Natural Language Processing Bootcamp

- More sophisticated phases consider the length/complexity of the word before applying a rule. For example:

S1		S2	word		stem
(m>0) ATIONAL	→	ATE	relational	→	relate
			national	→	national
(m>0) EED	→	EE	agreed	→	agree
			feed	→	feed

Natural Language Processing Bootcamp

- Snowball is the name of a stemming language also developed by Martin Porter.
- The algorithm used here is more accurately called the "English Stemmer" or "Porter2 Stemmer".
- It offers a slight improvement over the original Porter stemmer, both in logic and speed.

Natural Language Processing Bootcamp

- Let’s use Python and NLTK to show how to use these stemmers!

Lemmatization

Natural Language Processing Bootcamp

- In contrast to stemming, lemmatization looks beyond word reduction, and considers a language's full vocabulary to apply a morphological analysis to words.
- The lemma of 'was' is 'be' and the lemma of 'mice' is 'mouse'. Further, the lemma of 'meeting' might be 'meet' or 'meeting' depending on its use in a sentence.

Natural Language Processing Bootcamp

- Lemmatization is typically seen as much more informative than simple stemming, which is why Spacy has opted to only have Lemmatization available instead of Stemming.

Natural Language Processing Bootcamp

- Lemmatization looks at surrounding text to determine a given word's part of speech, it does not categorize phrases.
- In an upcoming lecture we'll investigate word vectors and similarity.
- For now, let’s learn how to perform lemmatization with Spacy.

Stop Words

Natural Language Processing Bootcamp

- Words like "a" and "the" appear so frequently that they don't require tagging as thoroughly as nouns, verbs and modifiers.
- We call these stop words, and they can be filtered from the text to be processed.
- Spacy holds a built-in list of some 326 English stop words.

Vocabulary and Matching

Natural Language Processing Bootcamp

- So far we've seen how a body of text is divided into tokens, and how individual tokens are parsed and tagged with parts of speech, dependencies and lemmas.
- In this lecture we will identify and label specific phrases that match patterns we can define ourselves.

Natural Language Processing Bootcamp

- We can think of this as a powerful version of Regular Expression where we actually take parts of speech into account for our pattern search.
- Let’s explore this with Spacy!

Vocabulary and Matching

Part Two

NLP Basics Assessment Overview