AI Bootcamp

Introduction to Natural Language Processing (NLP)

Module 20 Day 1

© 2023 edX Boot Camps LLC. Confidential and Proprietary. All Rights Reserved.

Class Objectives

By the end of class, you will be able to:

1. Define NLP and implement its workflow.
2. Demonstrate how to tokenize text.
3. Proficiently preprocess text, including tokenization and punctuation handling, for analysis.
4. Manage and process punctuation marks and other non-alphabetic characters.
5. Differentiate between stemming and lemmatization.
6. Understand the importance of removing stopwords.
7. Understand and demonstrate how to count tokens and n-grams.

Welcome

Instructor **Demonstration**

Introduction to NLP

What is **Natural Language Processing (NLP)**?

Methods for building computer software that understands, generates, and manipulates human language.

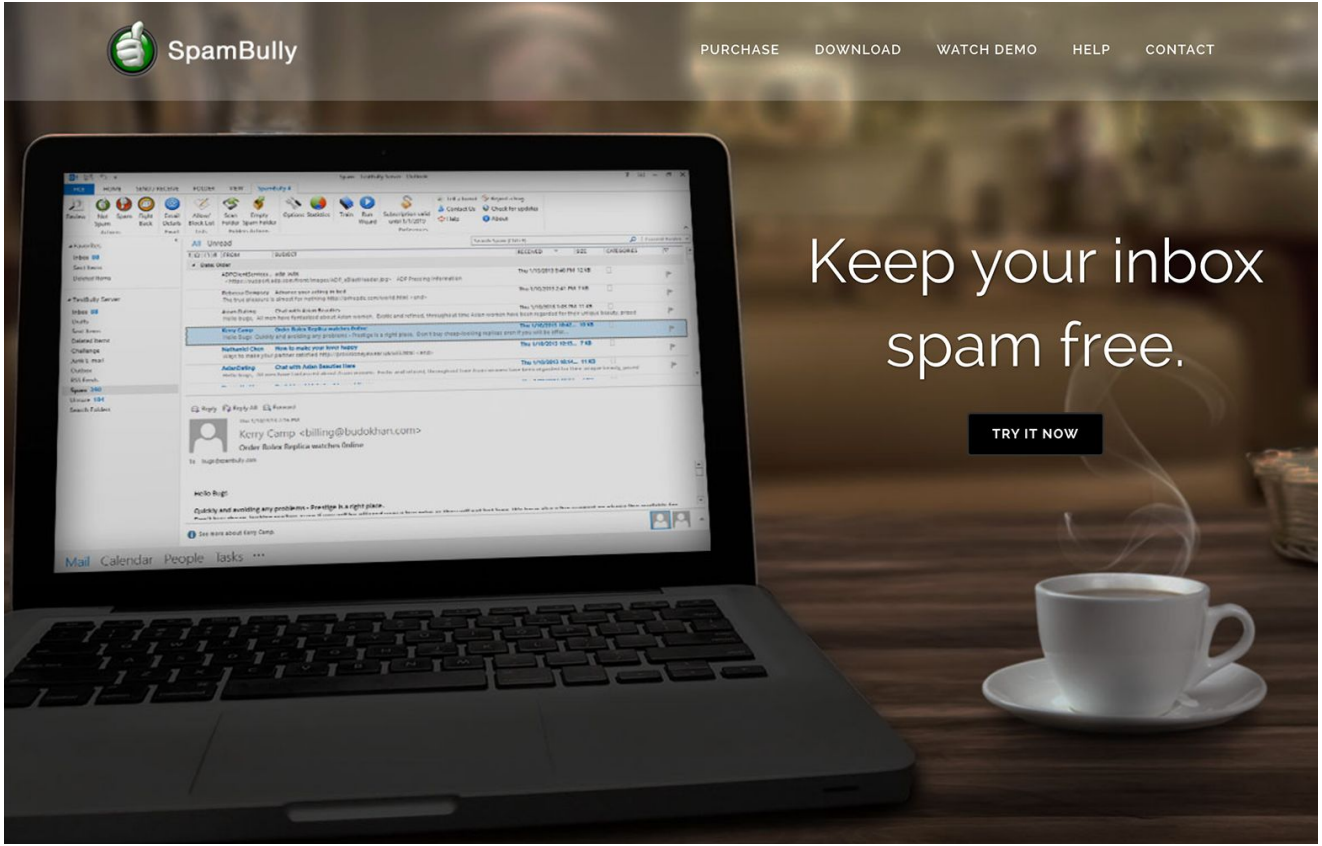
—Jacob **Eisenstein**

What is **NLP** used for?

Virtual Assistants

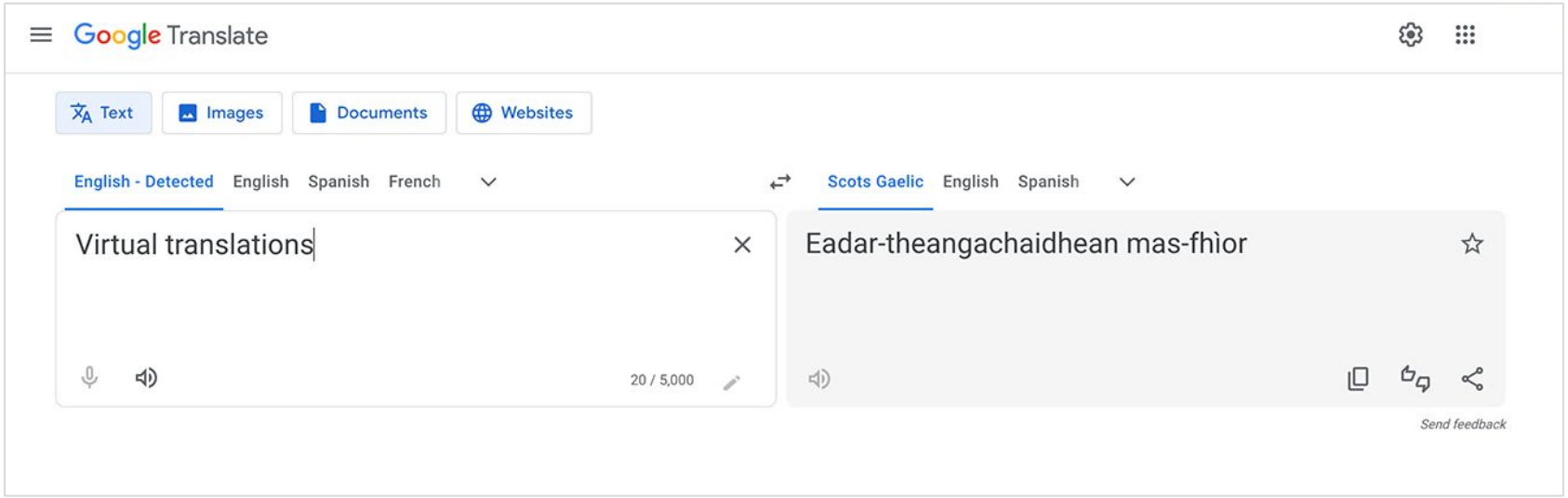
Alexa, Google Home, Siri

SPAM Filters



spambully.com

Virtual Translations (Google Translate)



google.com

NLP

Most industries have large quantities of textual data that can't be efficiently processed manually.

01 **Law:** Research, notes, documents, records of legal transactions, governmental information

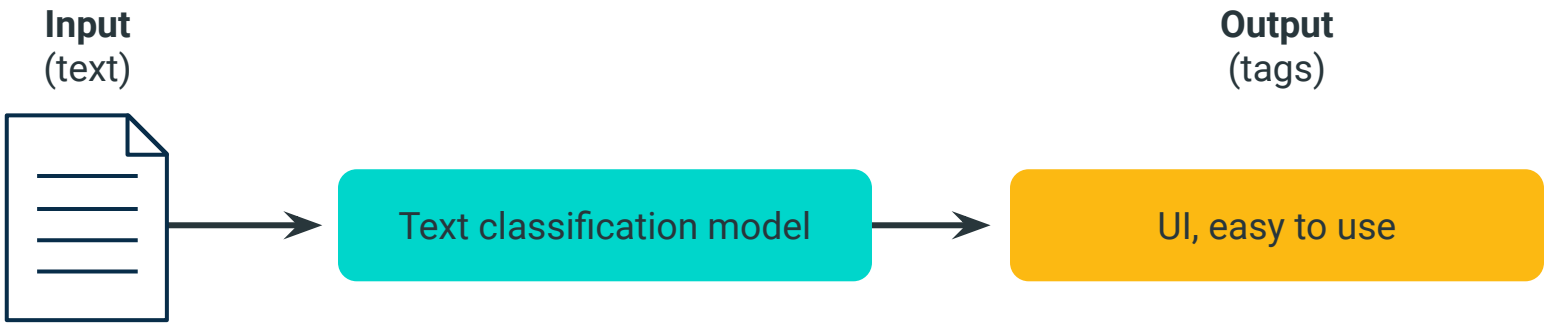
02 **Medical Research:** Patient information and history, clinical notes, symptoms

03 **Stock Market Analysis:** Company disclosures, news articles, report narratives

A Few **NLP Applications**

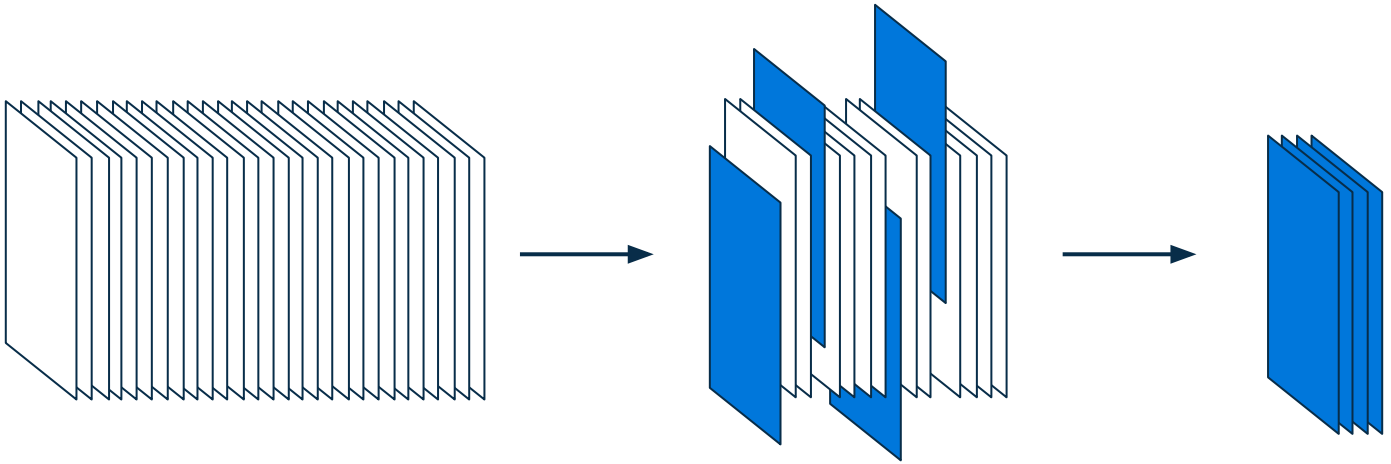
Text Classification

Classifying statements as subjective/objective, positive/negative; finding the reading level or genre of a text



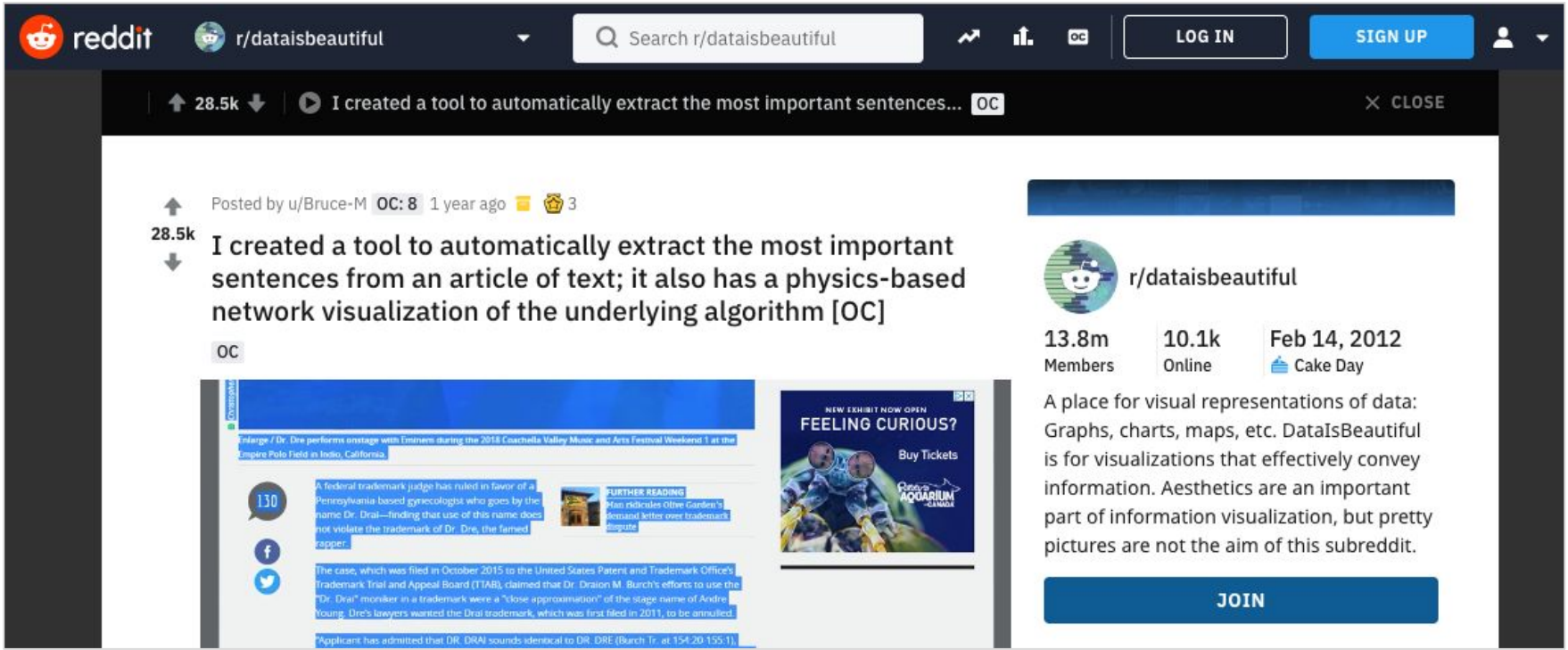
Information Extraction

Finding the diagnosis from a doctor’s notes; identifying names of individuals from a witness statement



Document Summarization

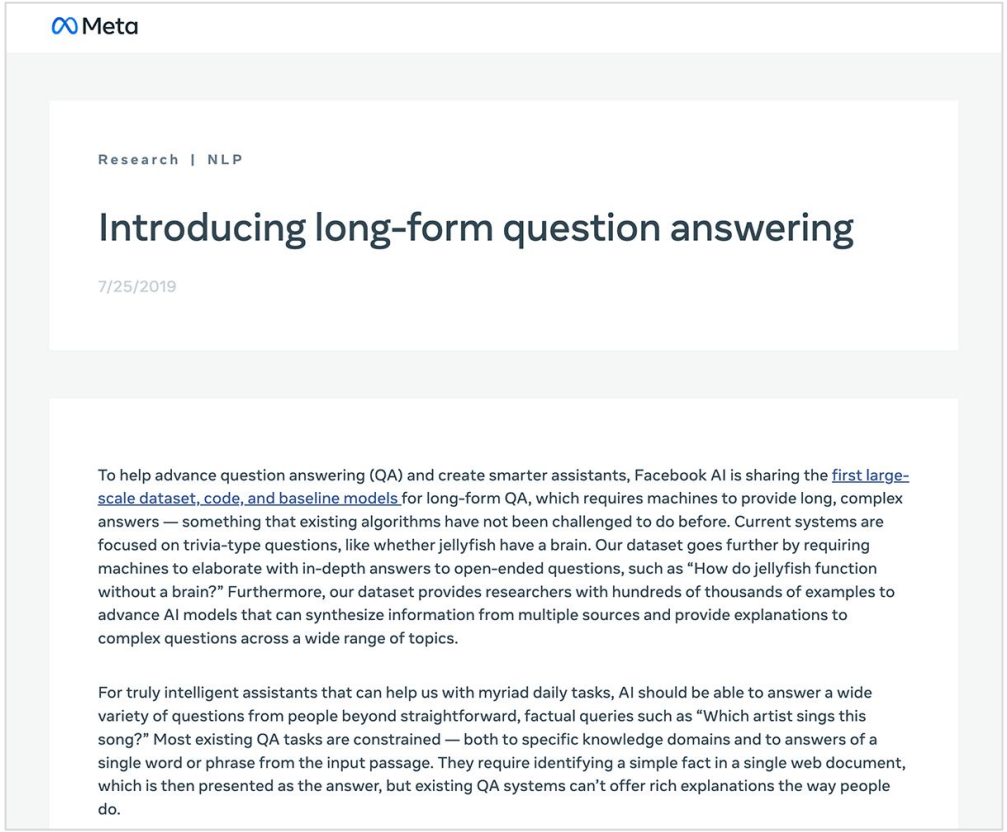
Generating a headline or abstract for a document



reddit.com

Complex Question Answering

Answering a question about a subject, given resources or a document on that subject



Meta

Research | NLP

Introducing long-form question answering

7/25/2019

To help advance question answering (QA) and create smarter assistants, Facebook AI is sharing the first large-scale dataset, code, and baseline models for long-form QA, which requires machines to provide long, complex answers — something that existing algorithms have not been challenged to do before. Current systems are focused on trivia-type questions, like whether jellyfish have a brain. Our dataset goes further by requiring machines to elaborate with in-depth answers to open-ended questions, such as "How do jellyfish function without a brain?" Furthermore, our dataset provides researchers with hundreds of thousands of examples to advance AI models that can synthesize information from multiple sources and provide explanations to complex questions across a wide range of topics.

For truly intelligent assistants that can help us with myriad daily tasks, AI should be able to answer a wide variety of questions from people beyond straightforward, factual queries such as "Which artist sings this song?" Most existing QA tasks are constrained — both to specific knowledge domains and to answers of a single word or phrase from the input passage. They require identifying a simple fact in a single web document, which is then presented as the answer, but existing QA systems can't offer rich explanations the way people do.

ai.facebook.com

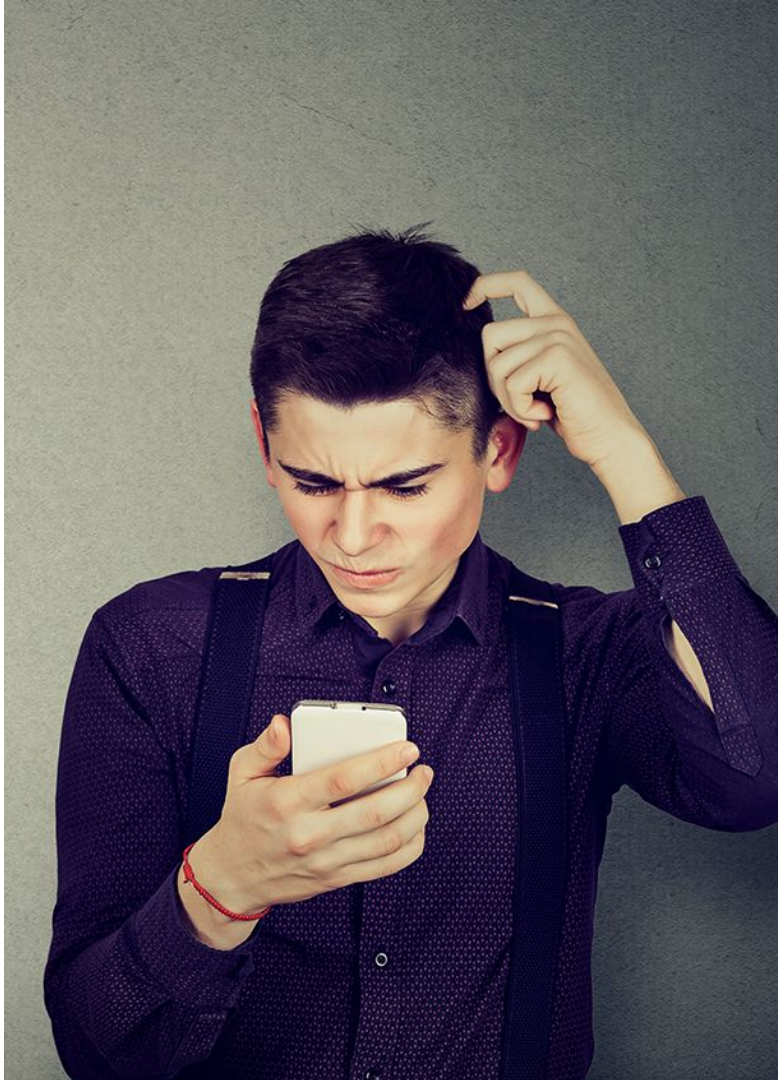
NLP is Hard

Humans intuitively interpret natural language, but even we aren't great at it all the time. Natural language is:

1. **Contextual:** The meaning of text depends on situation, speaker, and listener.

2. **Ambiguous:** Words have multiple meanings and can mean different things in different contexts

3. **Nonstandard:** There is no general set of rules, especially across dialects, groups, etc.



Natural Languages vs. Computer Languages

Computer languages (programming languages) are:

- Unambiguous
- Based on mathematical logic
- Designed to encode a very specific set of instructions

In order to bridge the gap between human natural language interpretation and processing by a computer, text data must be parsed, organized, and/or encoded. In other words, it must be converted to numbers.

NLP Workflow

01 **Preprocessing:** Preparing the text, including ingestion.

02 **Extraction:** Getting interesting features of the text.

03 **Analysis:** Summarizing these features.

04 **Representation:** Visualizing your analysis.

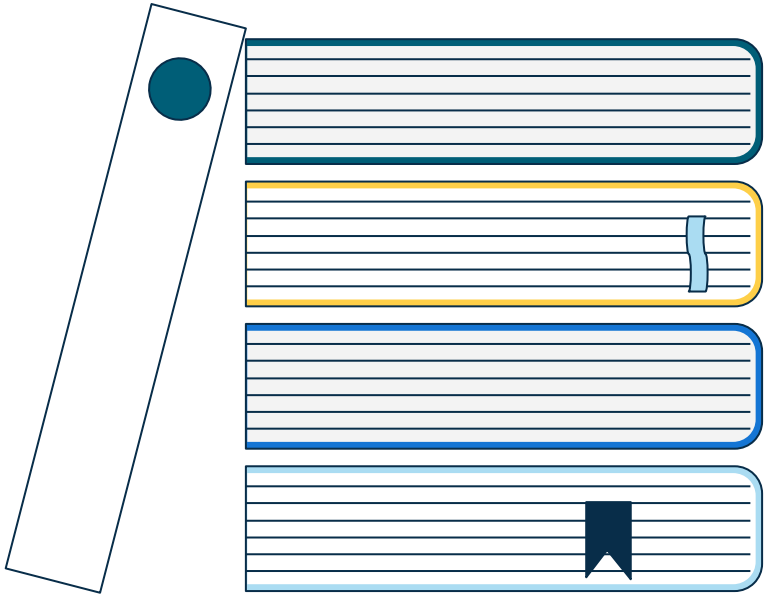
Tokenization

What is a **Corpus**?

Corpus

A corpus (plural, corpora) is a large, structured, and organized collection of text documents that usually focus on a specific subject.

A corpus may contain texts in a single language (monolingual corpus) or text data in multiple languages (multilingual corpus).



Tokenization

The process of segmenting running text into words, sentences, or phrases.

- Text needs to be segmented into units in order for any processing to be done.
- A token is a group of characters that have meaning. It can be words, sentences, or phrases.
- Sometimes characters such as punctuation are discarded.
- Tokenization is similar to using `split()` in Python.
- Sentence segmentation and tokenization are often the first steps in an NLP pipeline.

Tokenization is the process of splitting up a text document into units, most often sentences or words.

Instructor **Demonstration**

Tokenization

Activity:

Tokenizing Reuters

In this activity, you will practice both sentence and word tokenization on some articles from the Reuters Corpus.

Suggested Time:
15 Minutes

Time’s up!

Let’s review

Questions?

Instructor **Demonstration**

Stopwords

Stopwords: Words that, for analysis purposes, do not have informational content. Words like “the,” “there,” and “in.”

Stopwords

Stopwords are words that are useful for grammar and syntax, but they don't contain any important content.

- Generally, stopwords are the most commonly used words in the document.
- Examples: *this*, *to*, *the*, *a*, *there*, *an*
- Stopwords are often removed because they don't distinguish between relevant and irrelevant content.

Activity:

Crude Oil Stopwords

In this activity, you will practice creating a function that performs the preprocessing steps on a news article about crude oil.

Suggested Time:
15 Minutes

Time’s up!

Let’s review

Questions?

Break

15 mins

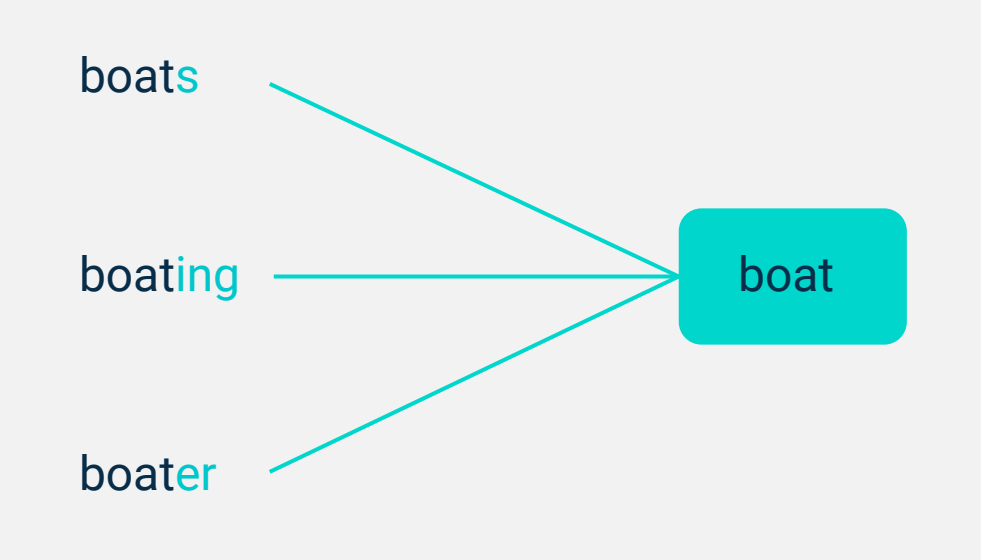
Instructor **Demonstration**

Stemming and Lemmatization

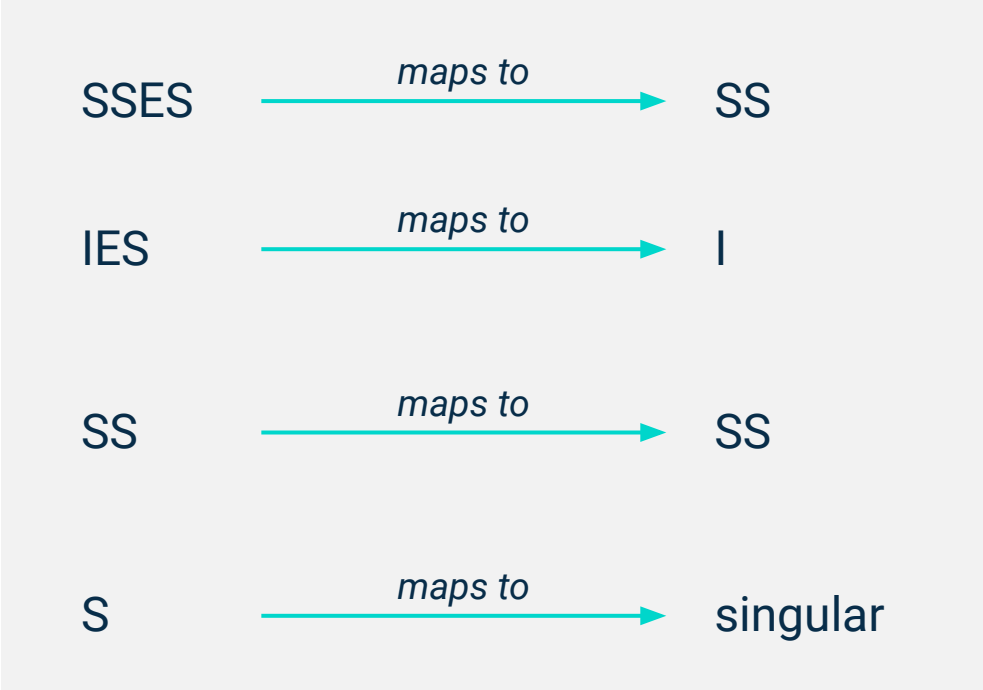
Stemming: reducing words to their “stem” or root form, typically by removing suffixes. For example, “jumps,” “jumping,” and “jumped” will all be stemmed to “jump.”

Stemming

If you were to search the word “boat,” the search results would include results containing the words “boats,” “boater,” and “boating” as well. The word “boat” was the stem for all these words.



Stemming rules for plural words



Lemmatization: reducing words to their “base” or root form. For example, the base form of “walking,” “walked,” and “walks” is “walk.”

Activity:

Stemming and Lemmatization of *Moby Dick*

In this activity, you will write a function that performs the preprocessing steps of removing stopwords and filtering out non-letter characters using regular expressions, as well as applying tokenizing, stemming and lemmatization to the American classic novel *Moby Dick*, written by Herman Melville.

Suggested Time:
15 Minutes

Time’s up!

Let’s review

Questions?

Instructor **Demonstration**

N-gram Counter

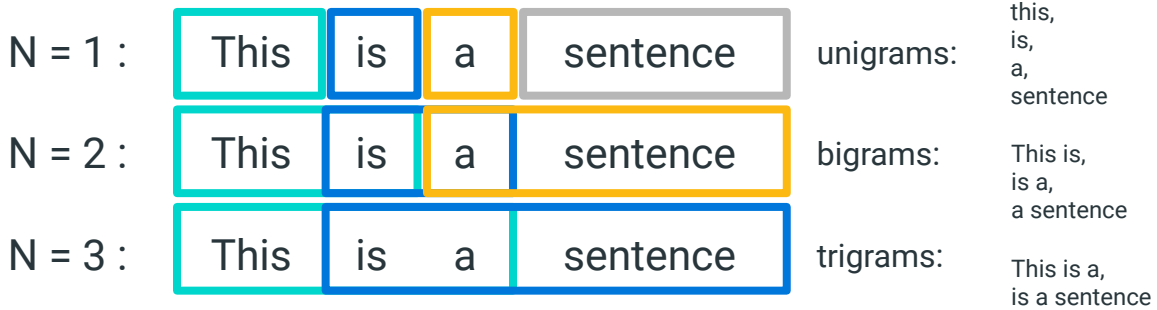
Frequency analysis involves counting words and phrases to uncover patterns and themes within a text or corpus. By prioritizing frequently occurring terms while excluding stopwords, you will have a good idea of what the document is about.

N-grams: Tokens that include multi-word phrases. The “n” is the number of words—for example, bigrams are two-word combinations.

N-grams

A group of n words appearing in sequence from a text.

- Splitting on single words can result in a model where syntax and order are ignored.
- Using an n-gram can be helpful in identifying the multi-word expressions or phrases.
- N-grams can be used to calculate how often words follow one another and are applied in generating text (predictive keyboards).
- N-grams are helpful in applications like sentiment analysis, where the ordering of the words is important to the context.



Activity:

Word and Bigram Corpus Counter

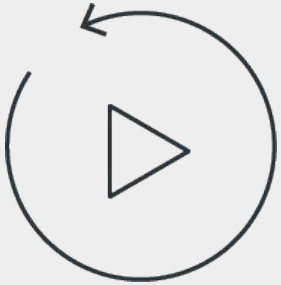
In this activity, you will create two DataFrames, one that has the top 10 most common words and another that has the top 10 most common bigrams from Reuters articles on grain.

Suggested Time:
15 Minutes

Time’s up!

Let’s review

Questions?



Let’s **recap**

Recap

After today's lesson you are able to:

1. Define NLP and implement its workflow.
2. Demonstrate how to tokenize text.
3. Proficiently preprocess text, including tokenization and punctuation handling, for analysis.
4. Manage and process punctuation marks and other non-alphabetic characters.
5. Differentiate between stemming and lemmatization.
6. Understand the importance of removing stopwords.
7. Understand and demonstrate how to count tokens and n-grams.

Next

In the next lesson, you'll learn advanced NLP techniques such as determining the importance of a word or words in a document, use supervised learning to classify the sentiment of text, and be introduced to the NLP tool spaCy, which has efficient and fast capabilities for tasks like tokenization, part-of-speech tagging, and more.

Questions?

The End