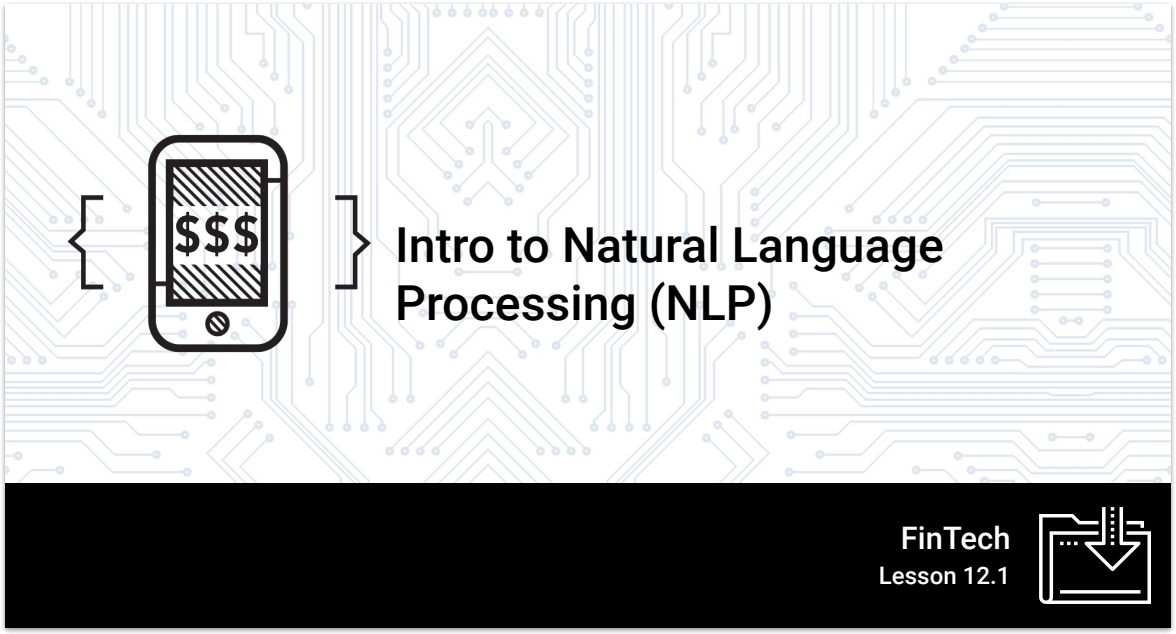
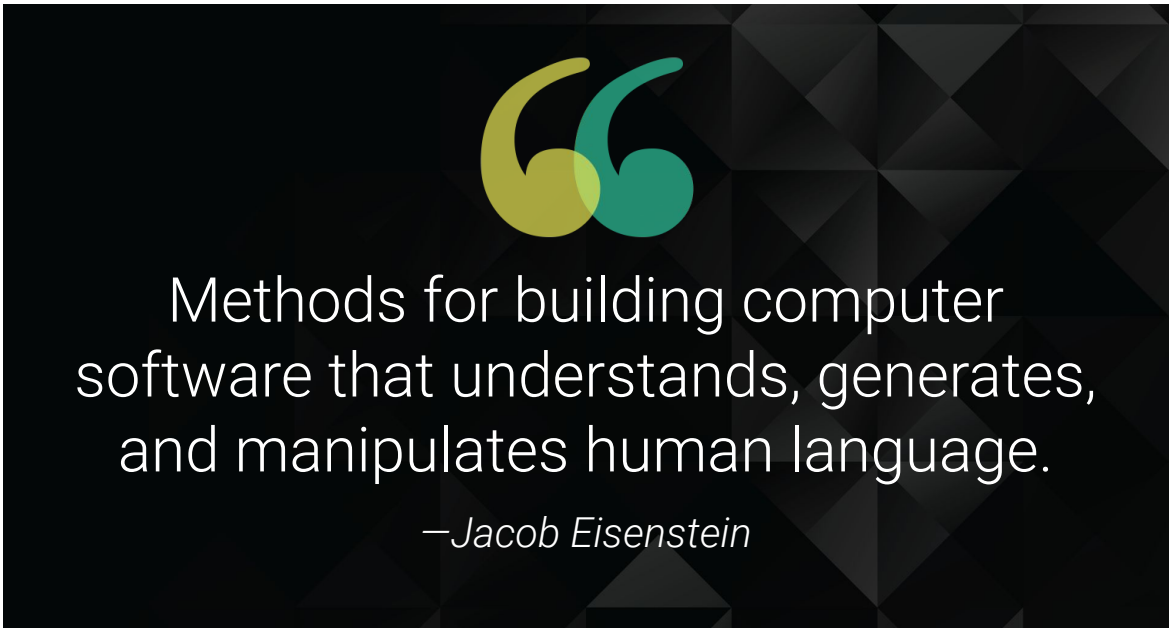


Class Objectives

- Understand what NLP is, why we use it
- Demonstrate ability to tokenize texts into sentences and words, including handling punctuations and non-alpha characters gracefully
- Implement lematization and stopwording with the understanding of pros and cons of various choices
- Experiment with a few ways of counting tokens and displaying the most frequent ones
- Define concept of ngrams and implement with scikit-learn
- Create wordcloud to show most frequent terms in a text

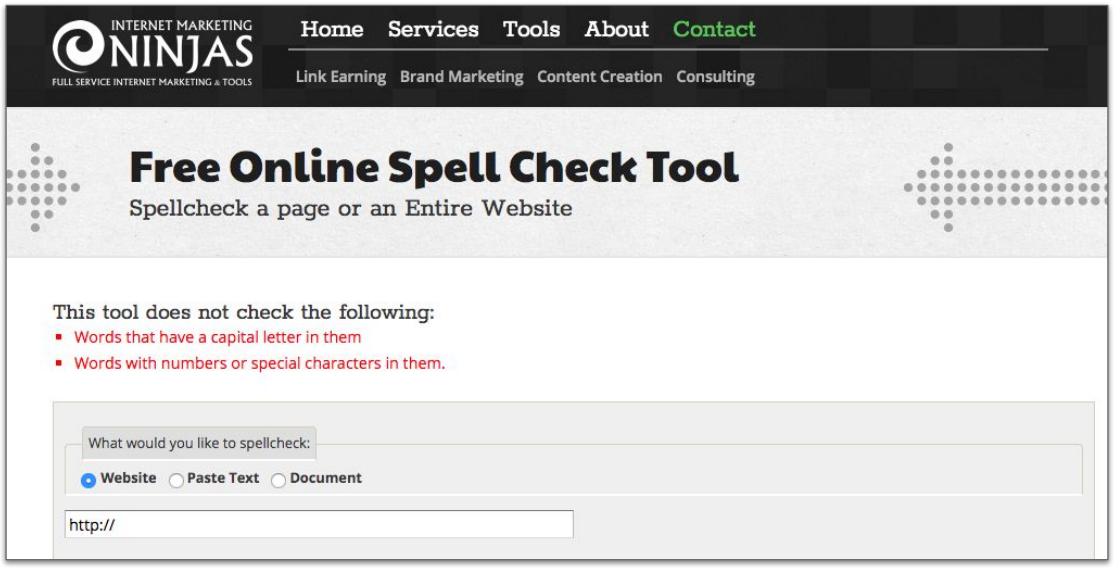
Disclaimer: The response and content of live data cannot be censored or predicted.

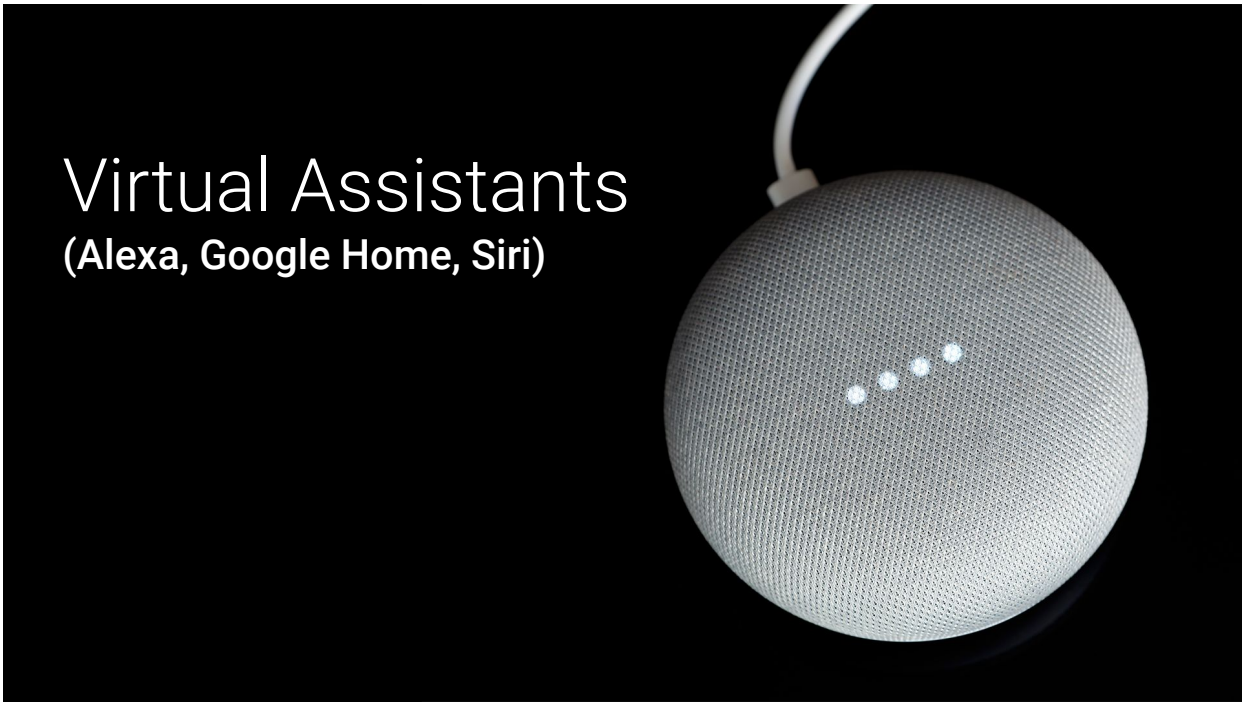
What is Natural Language Processing?


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

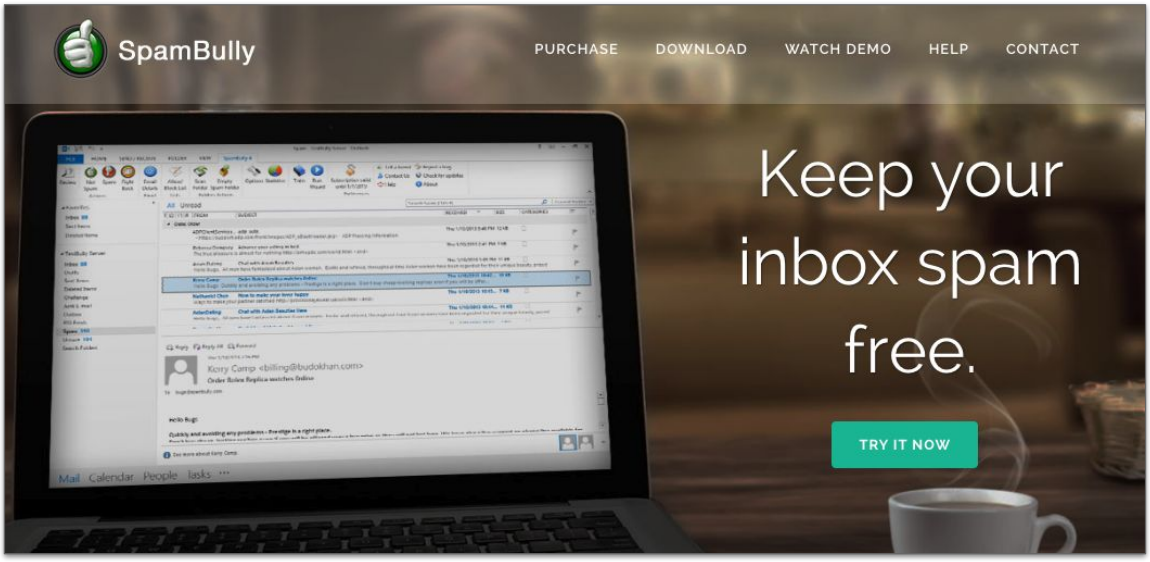

What Is NLP Used For?

Spell checkers



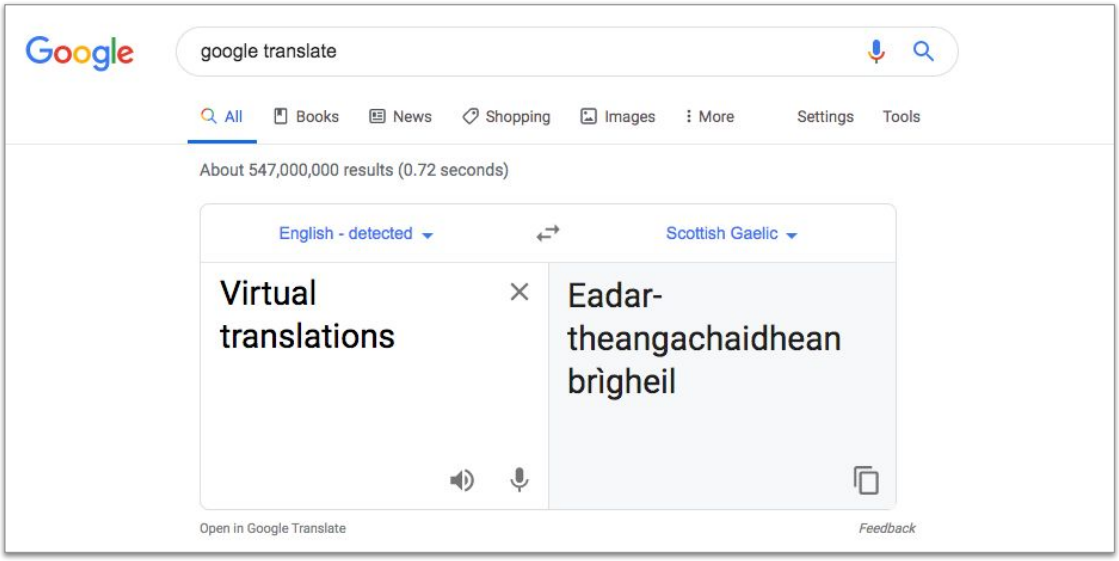


Spam filters



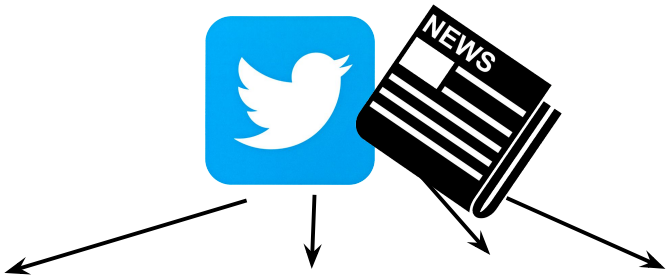
spambully.com

Virtual translations (Google Translate)



google.com

Handling unstructured data from tweets and Facebook posts



Company	Product	Data	Price	Region
Samsung	Galaxy A10e	April 11	$139	US
Nintendo	3DS	March 27	$250	North America
---	---	---	---	---

Product Release

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/history, clinical notes, symptoms

03

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

NLP In Finance

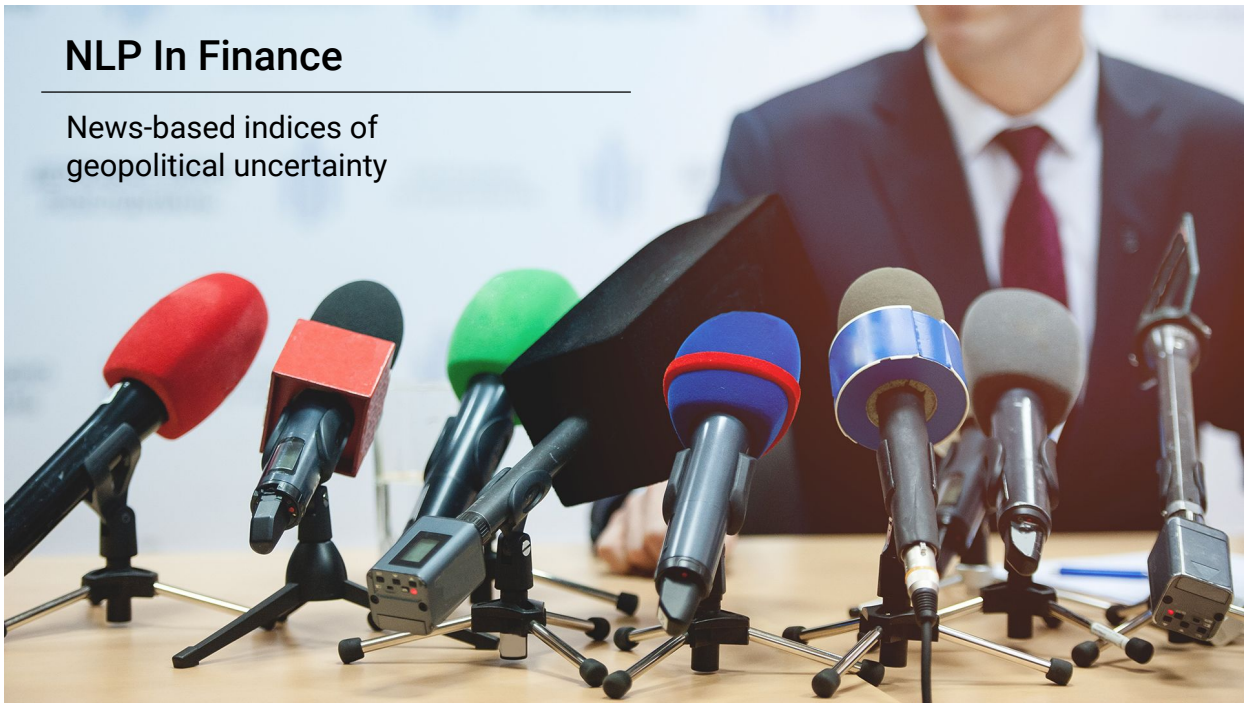
Automated sentiment analysis of earnings statements and investor calls

NLP In Finance

Predictions of interest rate from Federal Reserve testimony

NLP In Finance

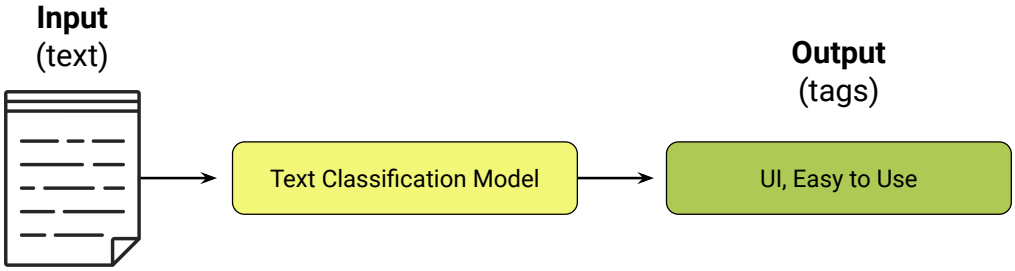
News-based indices of geopolitical uncertainty



A Few NLP Applications

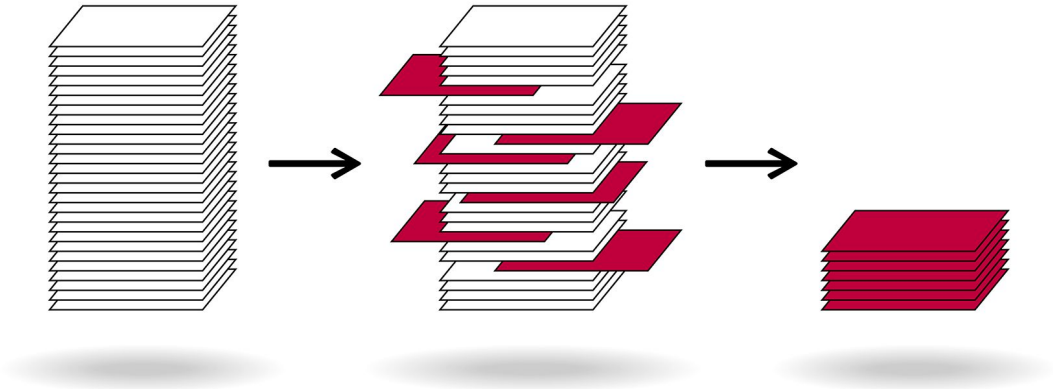
Text Classification

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



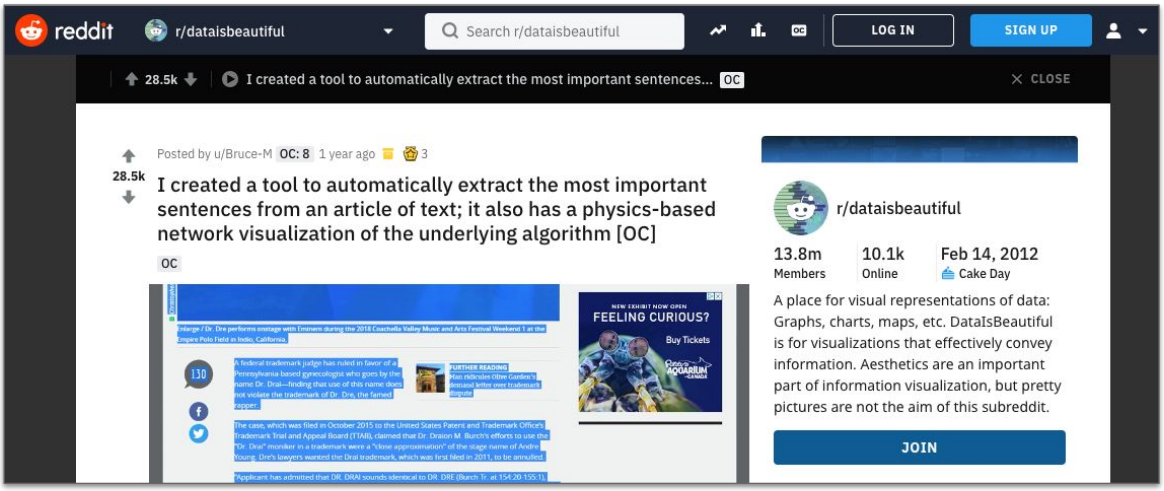
Information Extraction

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



Document Summarization

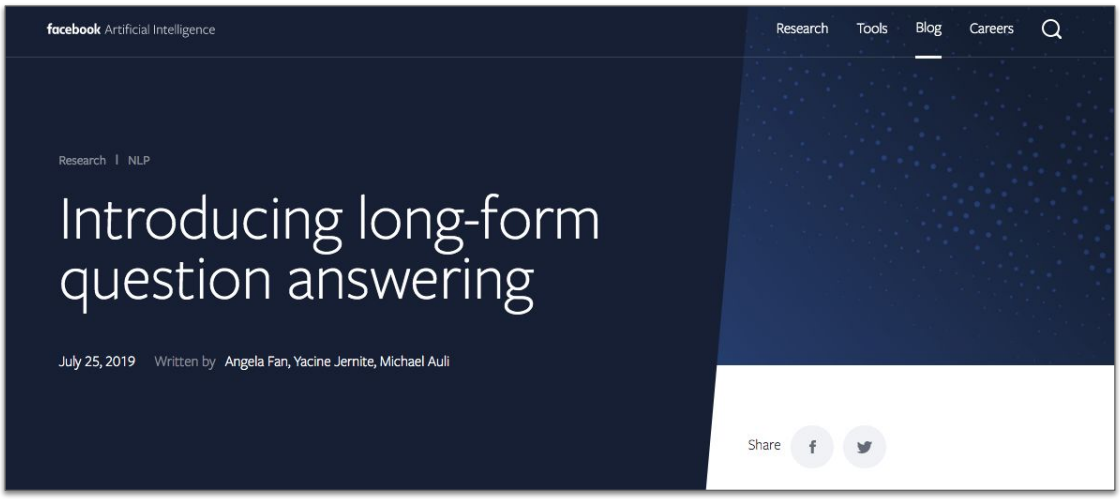
Generating a headline or abstract for a document



reddit.com

Complex Question Answering

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



ai.facebook.com

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:** get interesting features of the text

03 **Analysis:** summarize these features

04 **Representation:** visualize your analysis

Tokenization

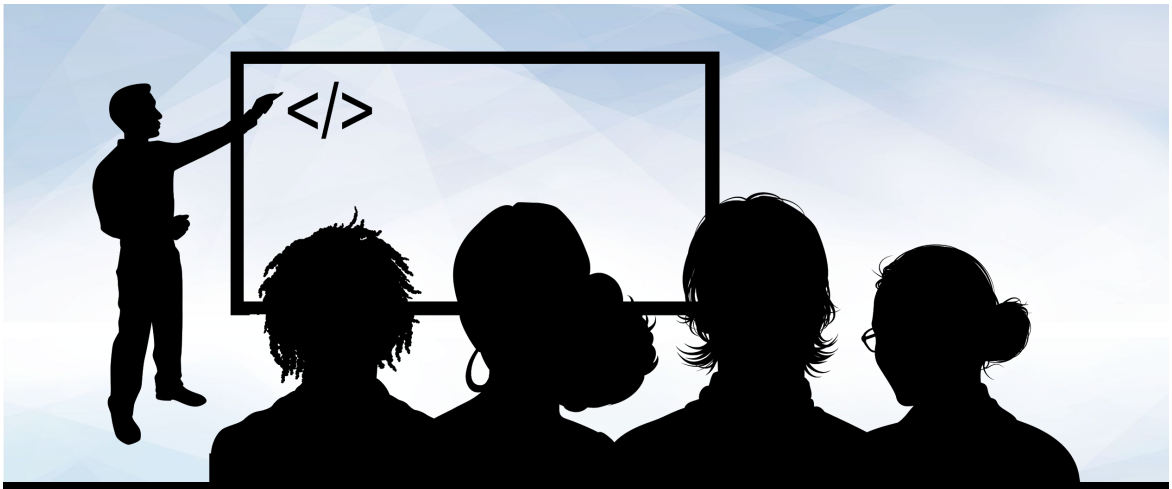
Tokenization

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.

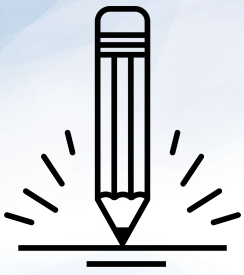
Let's eat, Grandpa! → `["let's", "eat", "grandpa"]`

Tokenization: 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 sentence and word tokenization on some articles from the Reuters corpus, and place the results in a pandas DataFrame.

Suggested Time:
15 minutes

Time's Up! Let's Review.

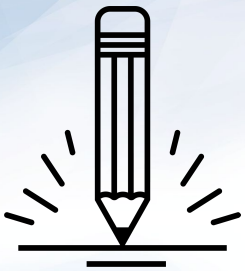
Stopwords

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 Stopwords

In this activity you will practice creating a function that strips non-letter characters from a document and then applies stopwording.

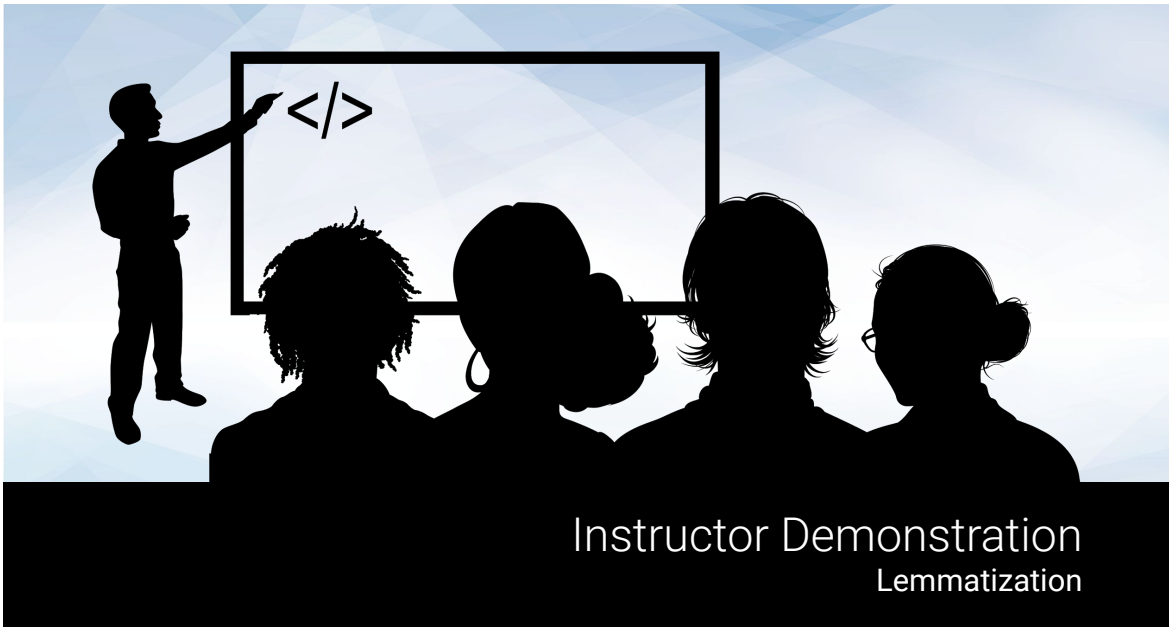
Suggested Time:
15 minutes

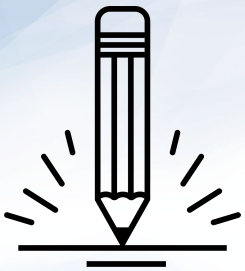
Time's Up! Let's Review.

Take a Break!

Lemmatization

Lemmatization: standardizing the "morphology" of words. For example, *walking*, *walked*, and *walks* will all become *walk*.





Activity: Lemmatize

In this activity, you will create a function that performs stopwording, regex cleaning of non-letter characters, word tokenizing, and lemmatization on each word in the article.

Suggested Time:
15 minutes

Time's Up! Let's Review.

N-Grams

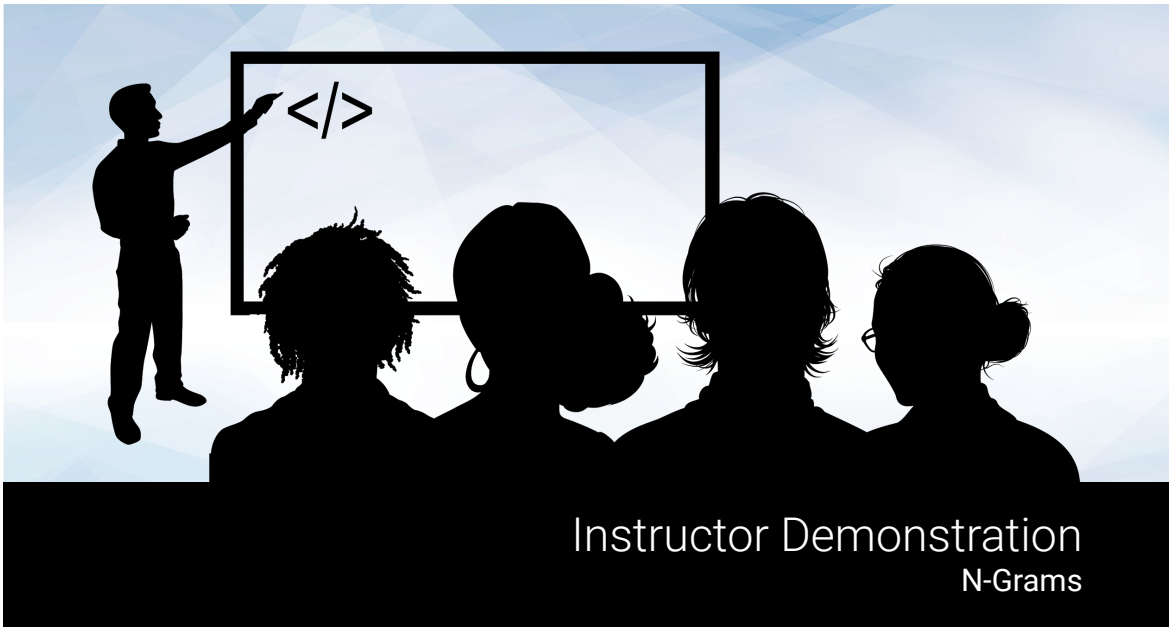
N-Grams: Tokens that include multi-word phrases. The n is the number of words—for example, bigrams are two-word combinations.

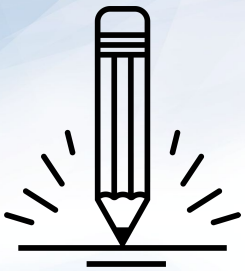
N-Grams

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 keyboard)
- N-grams are helpful in applications like sentiment analysis, where the ordering of the words is important to the context.

N = 1 :	This	is	a	sentence	unigrams:	this, is, a, sentence
N = 2 :	This	is	a	sentence	bigrams:	This is, is a, a sentence
N = 3 :	This	is	a	sentence	trigrams:	This is a, is a sentence



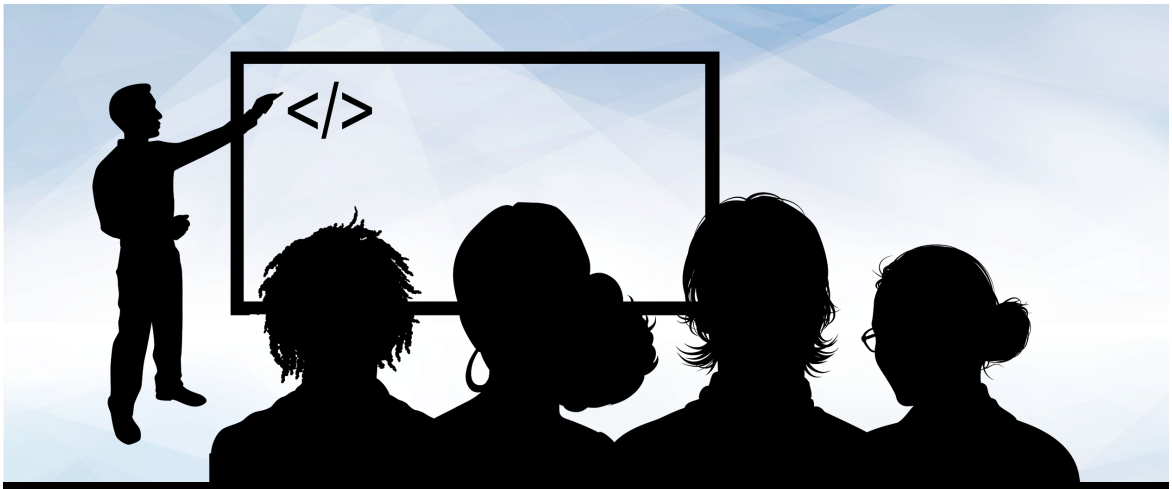


Activity: Counter

In this activity, you will create a function that pre-processes and outputs a list of the most common words in a corpus.

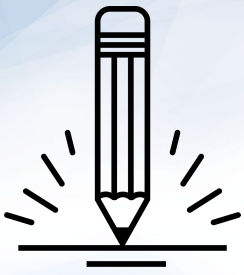
Suggested Time:
15 minutes

Time's Up! Let's Review.



Instructor Demonstration

Word Cloud



Activity: Gas Cloud

In this activity, you will practice creating a word cloud from a subset of the Reuters corpus.

.

Suggested Time:
15 minutes

Time's Up! Let's Review.

Questions?