



# Intro to Natural Language Processing (NLP)

FinTech  
Lesson 12.1



# 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?

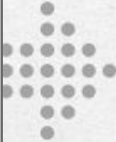
# Spell checkers



INTERNET MARKETING  
**NINJAS**  
FULL SERVICE INTERNET MARKETING & TOOLS


[Home](#) [Services](#) [Tools](#) [About](#) [Contact](#)

[Link Earning](#) [Brand Marketing](#) [Content Creation](#) [Consulting](#)



## Free Online Spell Check Tool

Spellcheck a page or an Entire Website



This tool does not check the following:

- Words that have a capital letter in them
- Words with numbers or special characters in them.

What would you like to spellcheck:

☒ Website ☐ Paste Text ☐ Document

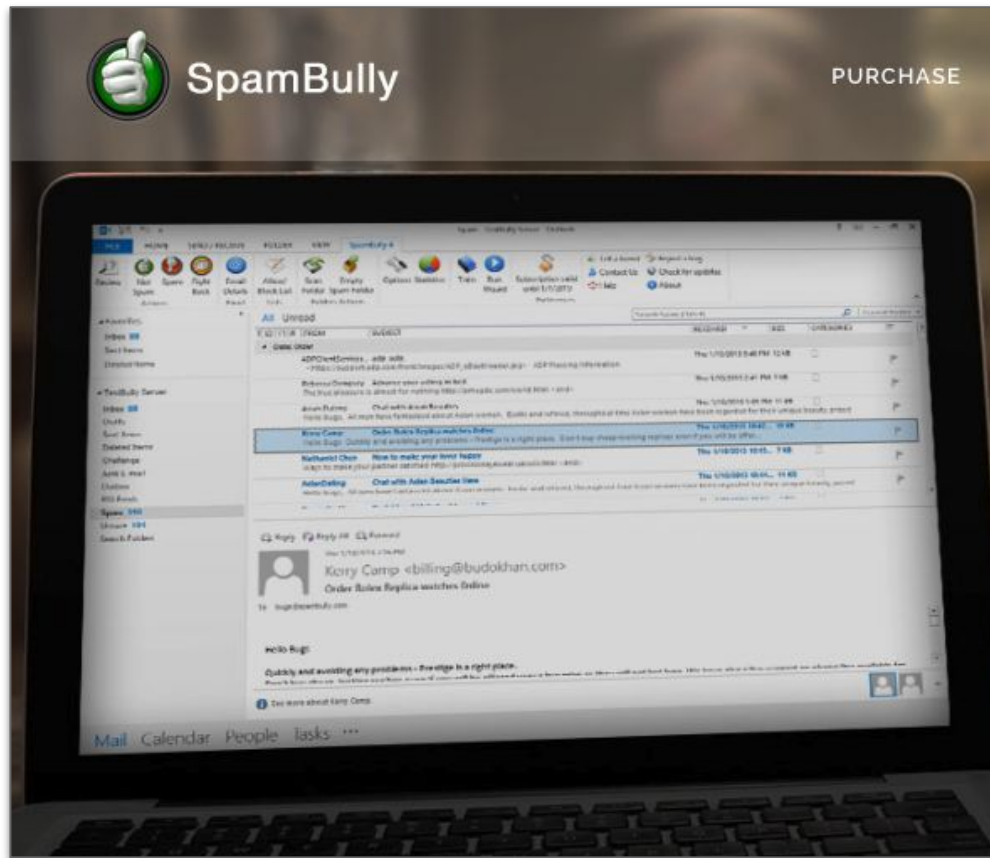
# Virtual Assistants

(Alexa, Google Home, Siri)





# Spam filters



The image shows a laptop screen displaying the SpamBully software interface. The interface includes a sidebar with navigation options like 'Inbox', 'Spam', and 'Trash'. The main window shows a list of emails, with one email from 'Kerry Camp' selected. The email content is visible, showing a message about 'Order Rules Replicas watches Online'. The background of the laptop screen is a blurred image of a coffee cup and a stack of papers.

**SpamBully**

PURCHASE DOWNLOAD WATCH DEMO HELP CONTACT

## Keep your inbox spam free.

**TRY IT NOW**

# Virtual translations (Google Translate)

The screenshot shows the Google Translate web interface. At the top, the Google logo is on the left, and a search bar contains the text 'google translate' with a microphone icon and a search icon on the right. Below the search bar, navigation links for 'All', 'Books', 'News', 'Shopping', 'Images', 'More', 'Settings', and 'Tools' are displayed. The search results show 'About 547,000,000 results (0.72 seconds)'. A translation box is open, showing 'English - detected' on the left and 'Scottish Gaelic' on the right. The text 'Virtual translations' is entered in the English box, and its translation 'Eadar-theangachaidhean brìgheil' is shown in the Scottish Gaelic box. Below the text boxes are icons for audio playback and voice input. At the bottom of the translation box, there is a link to 'Open in Google Translate' and a 'Feedback' link.

Google

google translate

All Books News Shopping Images More Settings Tools

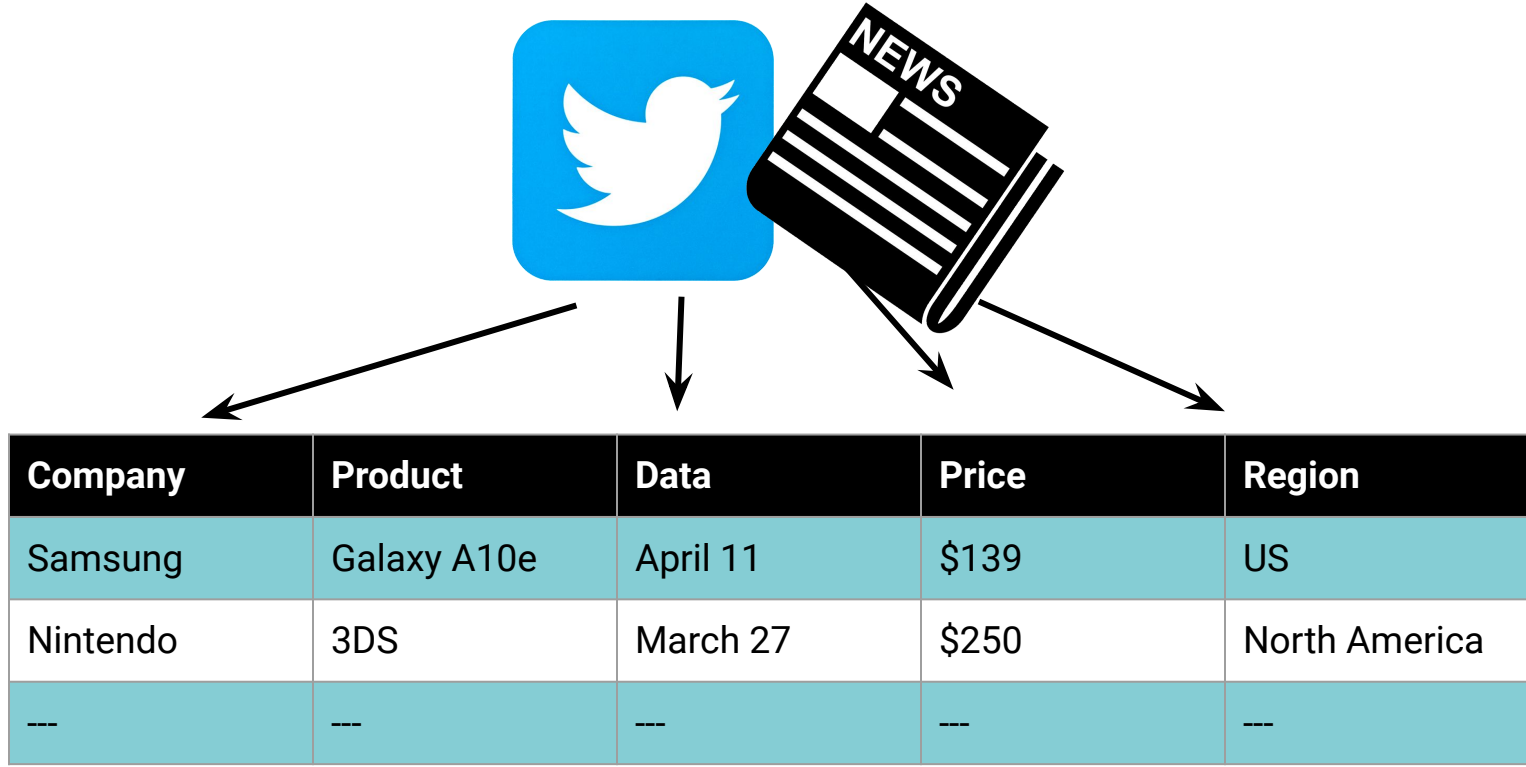
About 547,000,000 results (0.72 seconds)

English - detected ↔ Scottish Gaelic

Virtual translations × Eadar-theangachaidhean brìgheil

Open in Google Translate Feedback

# Handling unstructured data from tweets and Facebook posts



**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



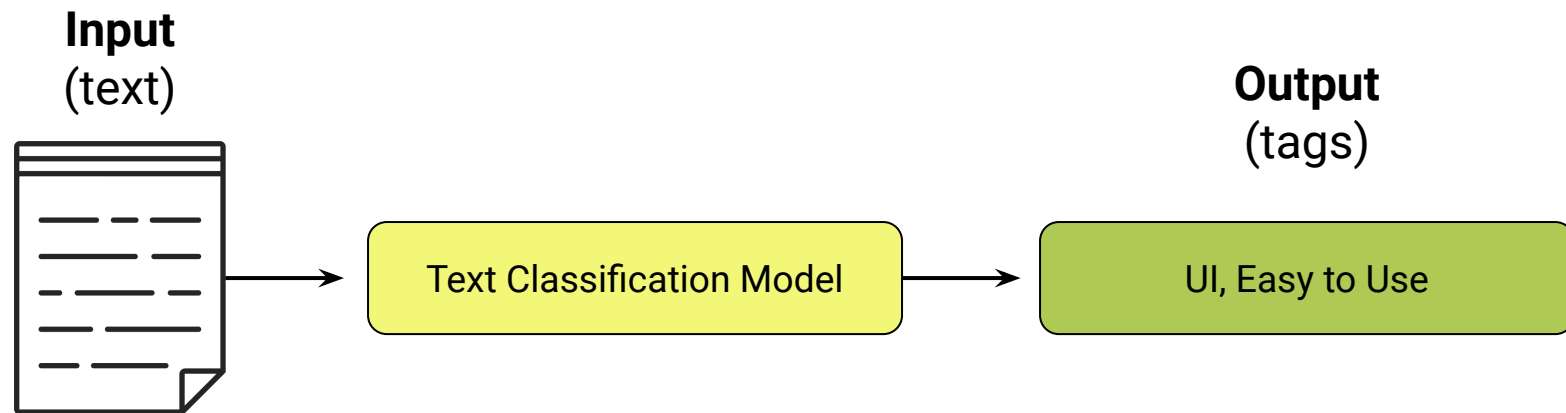
# **A Few NLP Applications**



# Text Classification

---

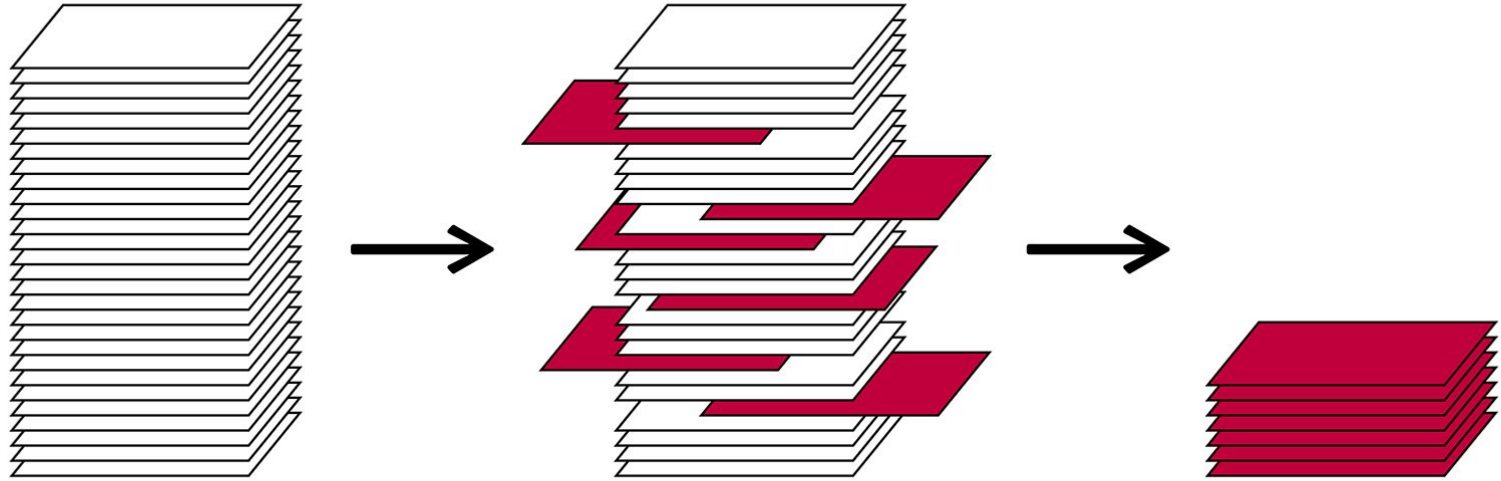
Classifying statements as subjective/objective, positive/negative; finding the reading level or genre of a text



# 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 r/dataisbeautiful

Search r/dataisbeautiful

LOG IN SIGN UP

28.5k I created a tool to automatically extract the most important sentences... OC

Posted by u/Bruce-M OC: 8 1 year ago

28.5k I created a tool to automatically extract the most important sentences from an article of text; it also has a physics-based network visualization of the underlying algorithm [OC]

OC

Enlarge / Dr. Dre performs onstage with Eminem during the 2018 Coachella Valley Music and Arts Festival Weekend 1 at the Empire Polo Field in Indio, California.

130

A federal trademark judge has ruled in favor of a Pennsylvania-based gynecologist who goes by the name Dr. Dre—finding that use of this name does not violate the trademark of Dr. Dre, the famed rapper.

FURTHER READING

Man ridicules Olive Garden's deplorable letter over trademark dispute

The case, which was filed in October 2015 to the United States Patent and Trademark Office's Trademark Trial and Appeal Board (TTAB), claimed that Dr. Dreton M. Burch's efforts to use the "Dr. Dre" moniker in a trademark were a "close approximation" of the stage name of Andre Young. Dre's lawyers wanted the Dre trademark, which was first filed in 2011, to be annulled.

\*Applicant has admitted that DR, DRAI sounds identical to DR, DRE (Burch Tr. at 154-20 155-1).

NEW EXHIBIT NOW OPEN  
FEELING CURIOUS?  
Buy Tickets  
GARDEN OF AQUARIUM CANADA

r/dataisbeautiful

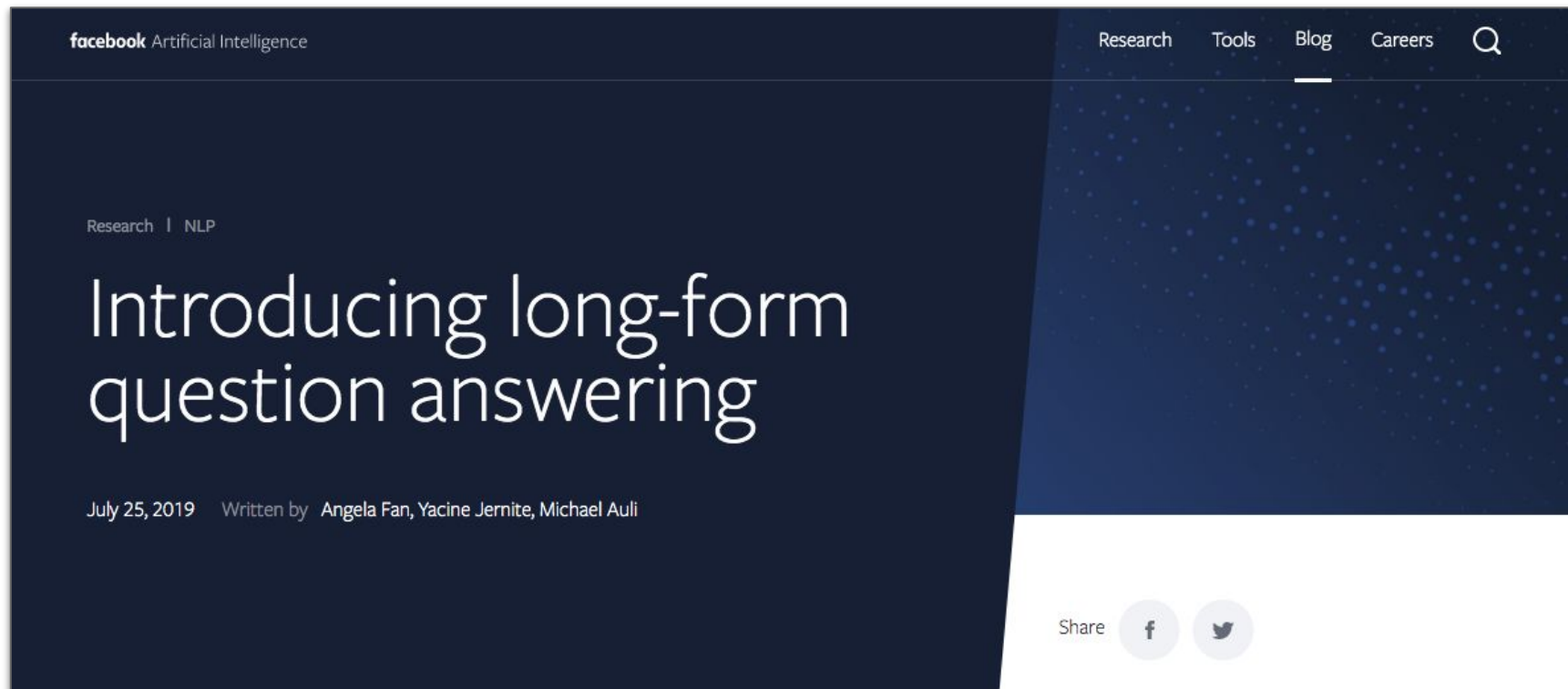
13.8m Members 10.1k Online Feb 14, 2012 Cake Day

A place for visual representations of data: Graphs, charts, maps, etc. DataIsBeautiful is for visualizations that effectively convey information. Aesthetics are an important part of information visualization, but pretty pictures are not the aim of this subreddit.

JOIN

# Complex Question Answering

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





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

### **Contextual:**

The meaning of text depends on situation, speaker, and listener.

### **Ambiguous:**

Words have multiple meanings and can mean different things in different contexts.

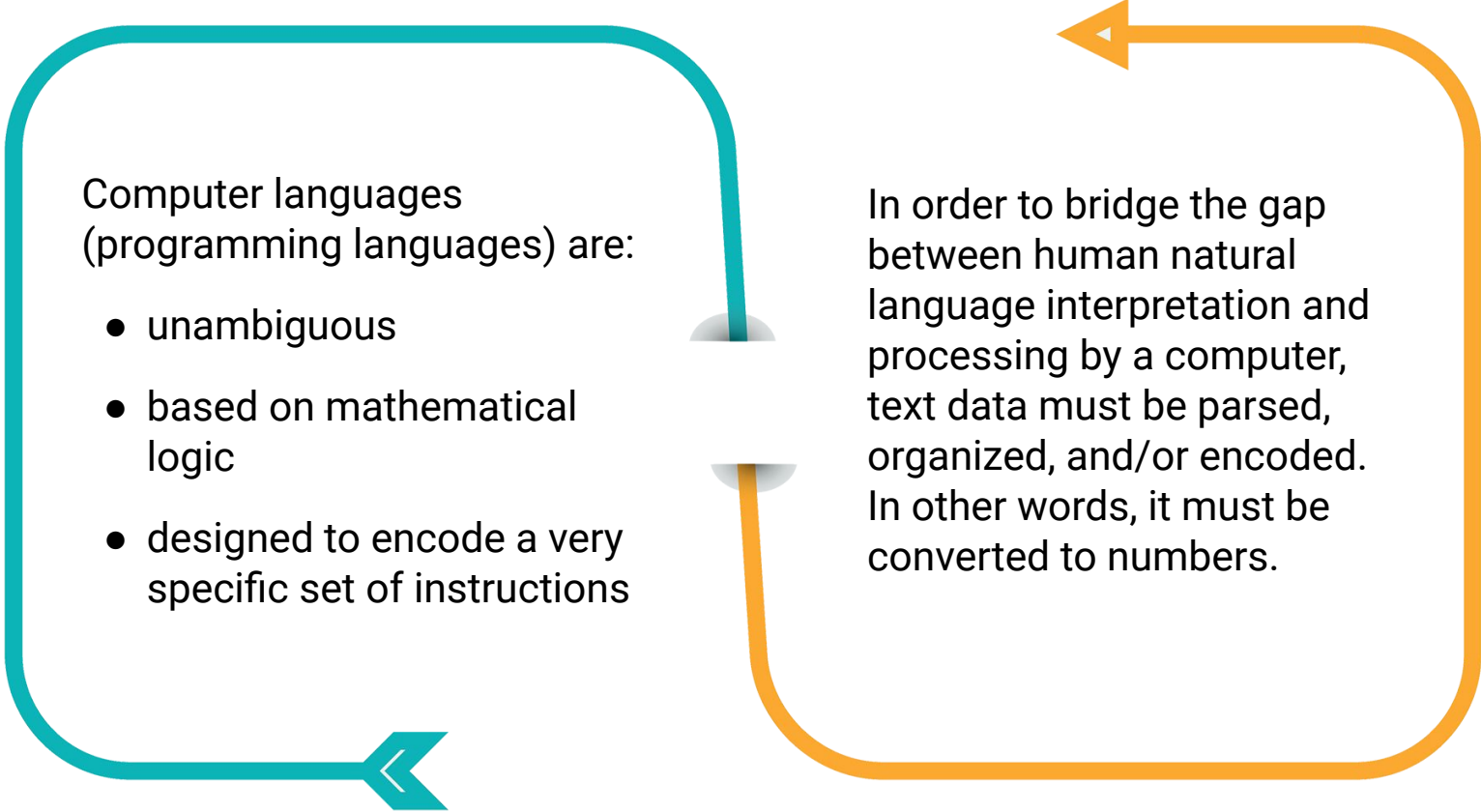
### **Nonstandard:**

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:** 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!

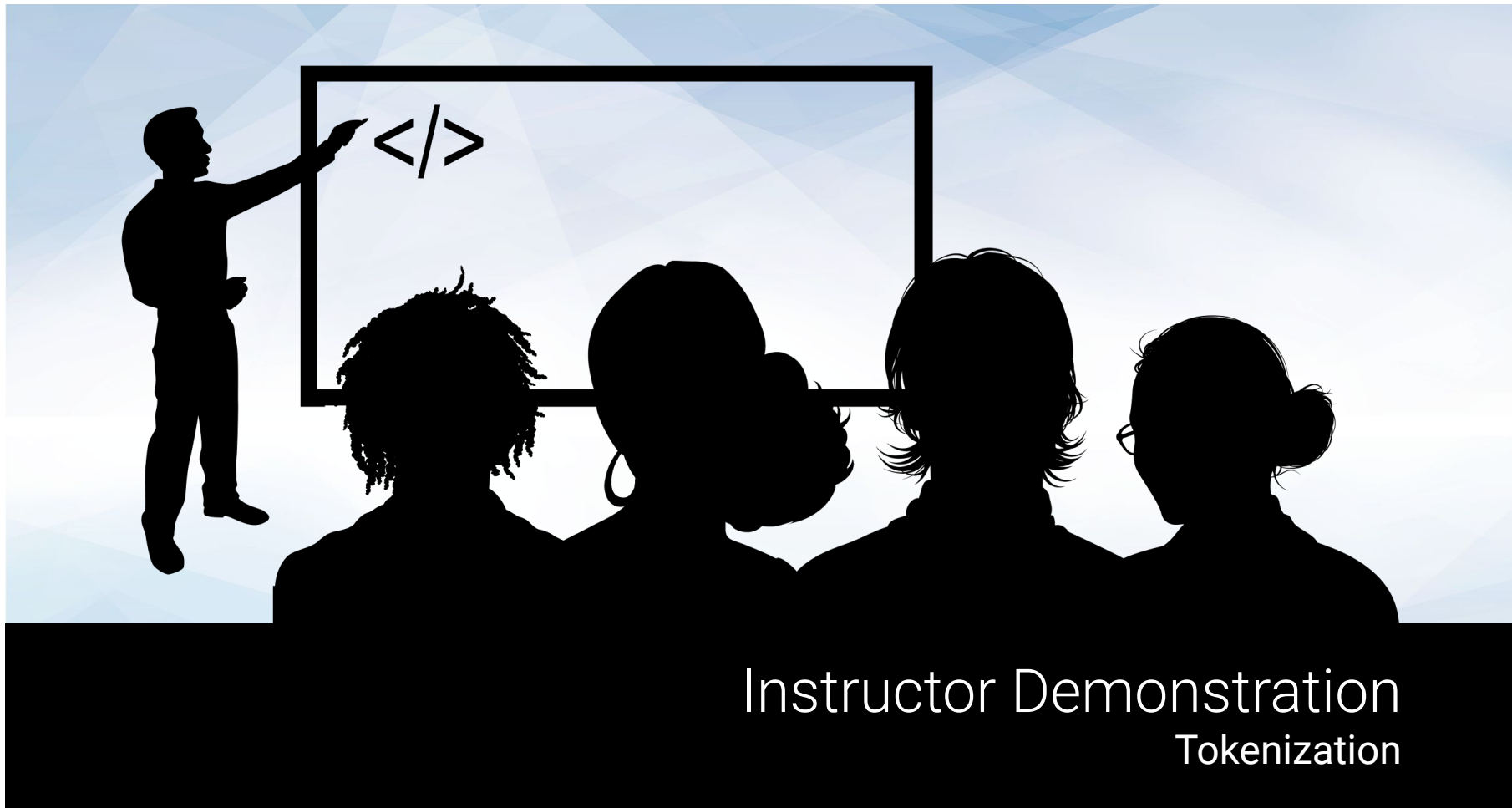


`["let's", "eat", "grandpa"]`





**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.

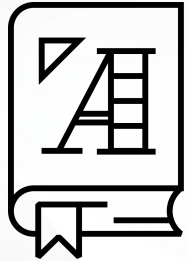
**Suggested Time:**  
15 minutes





**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





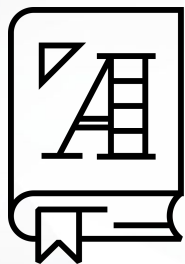
**Time's Up!** Let's Review.

# Take a Break!

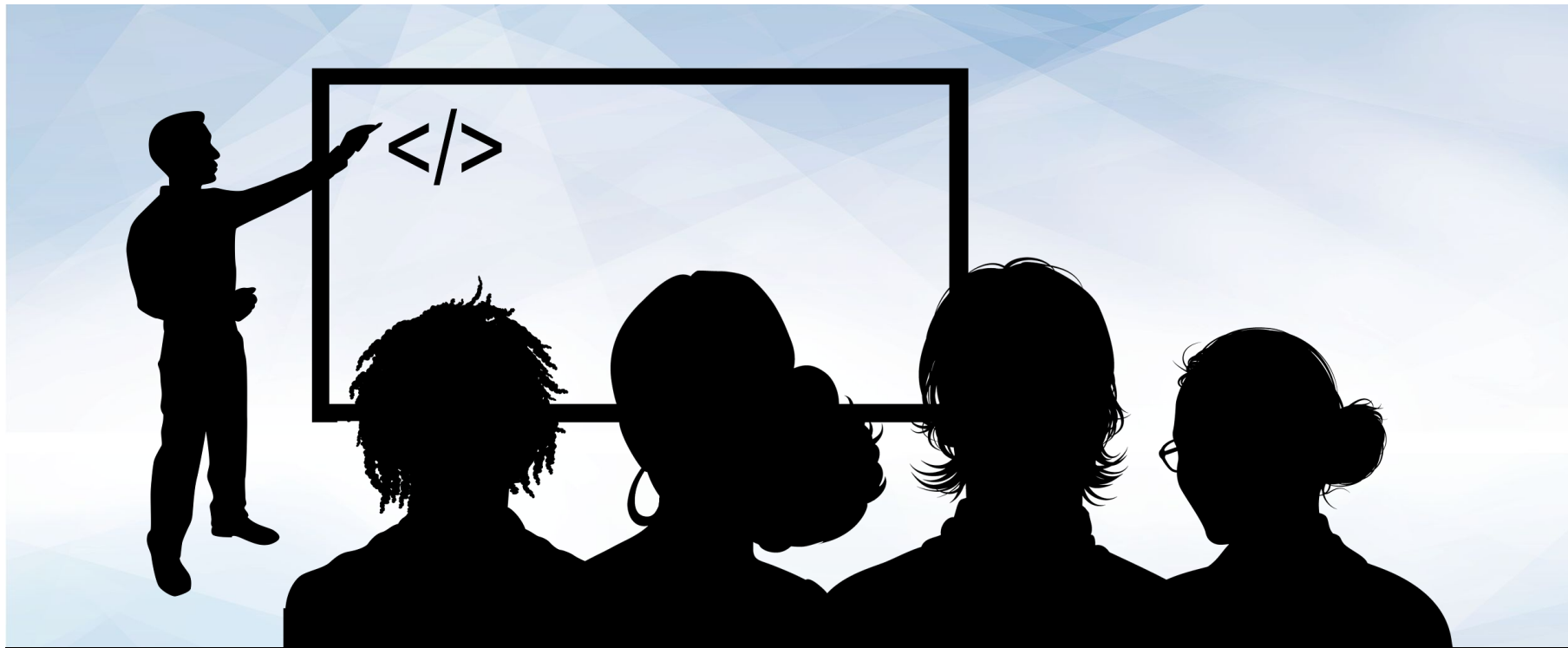
---



# Lemmatization



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



# Instructor Demonstration

## Lemmatization



## **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

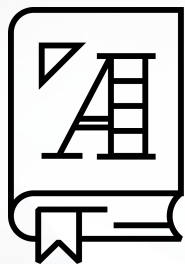




**Time's Up!** Let's Review.



# N-Grams

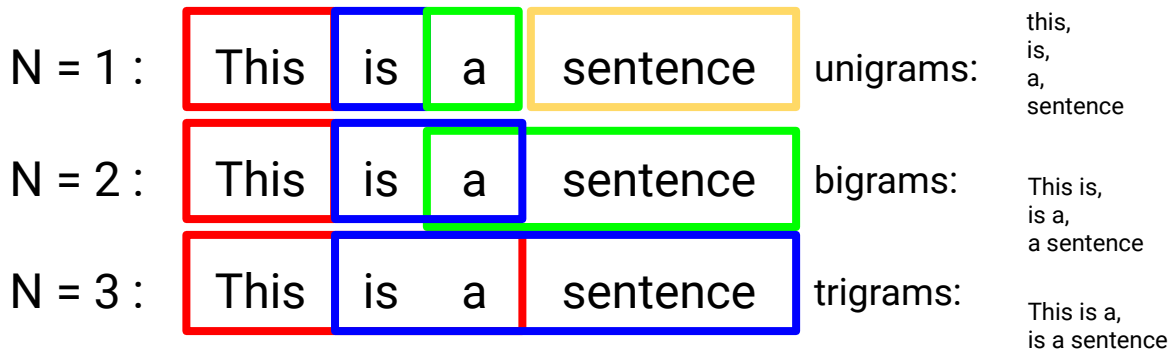


**N-Grams:** Tokens that include multi-word phrases. The  $n$  is the number of words—for example, bigrams are two-word combinations.

# 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.





# Instructor Demonstration

## N-Grams



## **Activity:** Counter

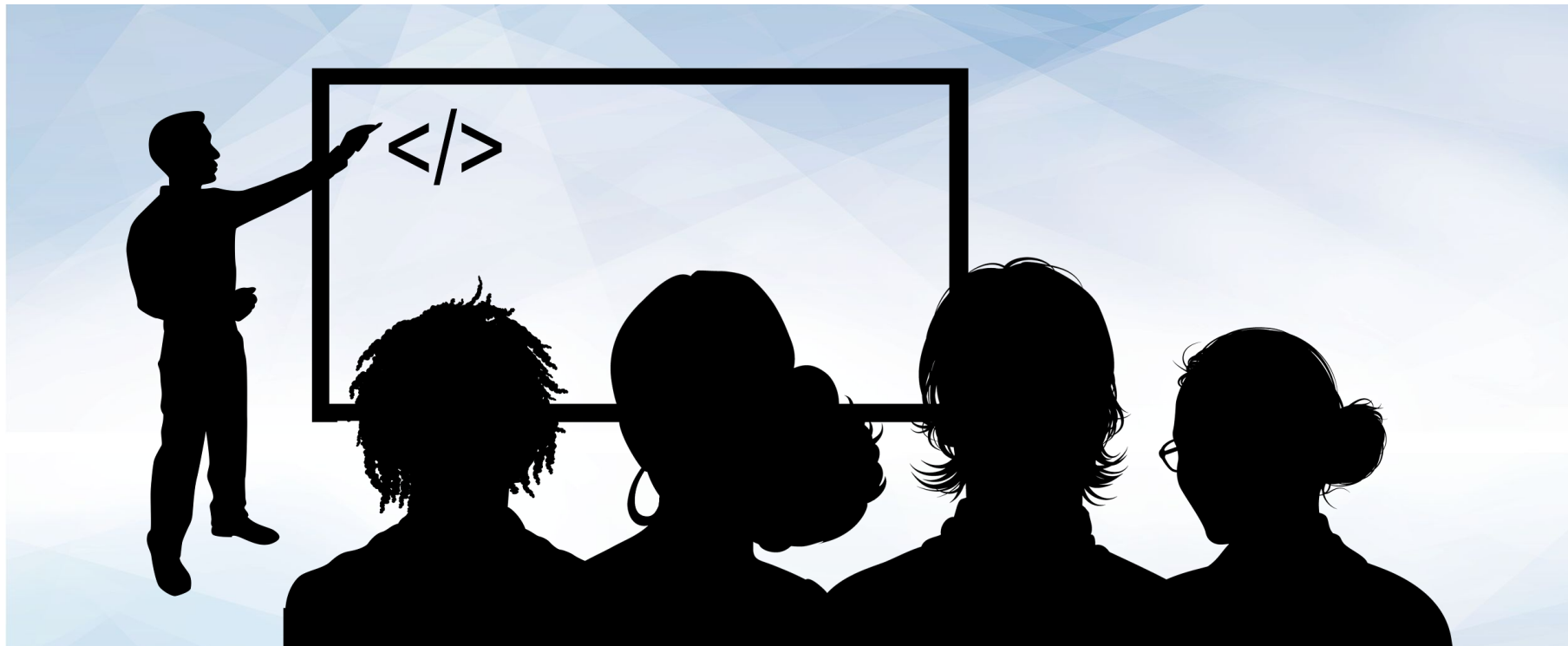
In this activity, you will create a function that pre-processes and outputs a list of the most common words in a corpus.

**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?