

© 2020 Trilogy Education Services, a 2U, Inc. brand. All Rights Reserved.

## 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.

3



## What is Natural Language Processing?

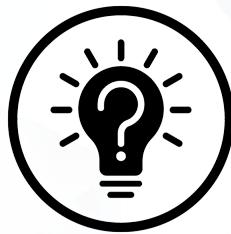
4



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

*—Jacob Eisenstein*

5



## What Is NLP Used For?

6

## Spell checkers

The screenshot shows the homepage of Internet Marketing Ninjas. At the top, there's a navigation bar with links for Home, Services, Tools, About, Contact, Link Earning, Brand Marketing, Content Creation, and Consulting. Below the navigation is a large banner for a "Free Online Spell Check Tool". The banner features a dotted arrow graphic and the text "Free Online Spell Check Tool" and "Spellcheck a page or an Entire Website". A note below states: "This tool does not check the following:" followed by two bullet points: "Words that have a capital letter in them" and "Words with numbers or special characters in them". There's a form field asking "What would you like to spellcheck?" with options for "Website" (selected), "Paste Text", and "Document". Below the form is a URL input field containing "http://".

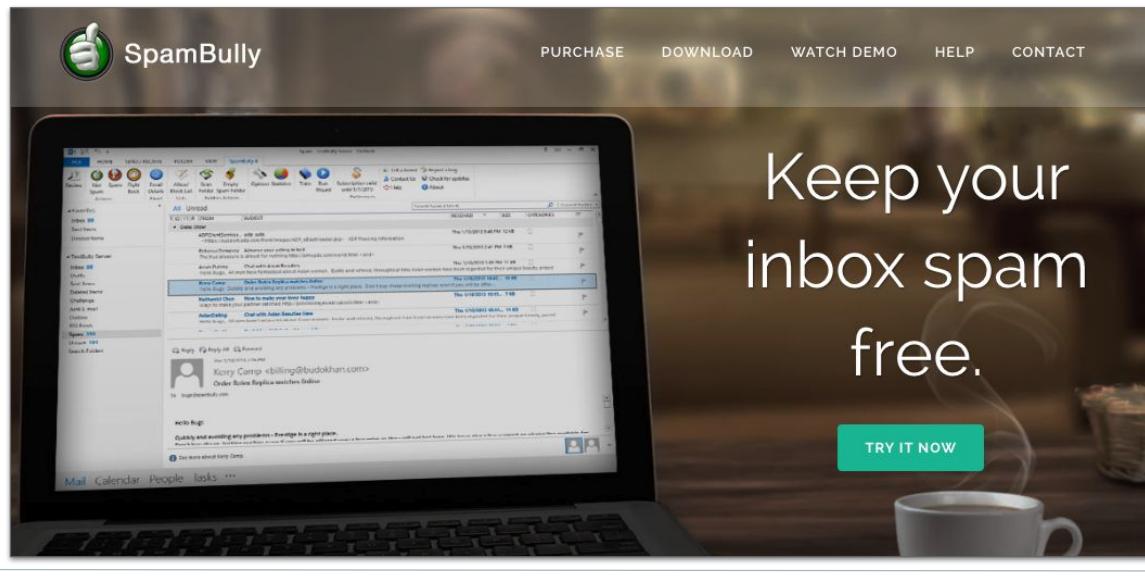
[internetmarketingninjas.com](http://internetmarketingninjas.com)

7

## Virtual Assistants (Alexa, Google Home, Siri)



## Spam filters

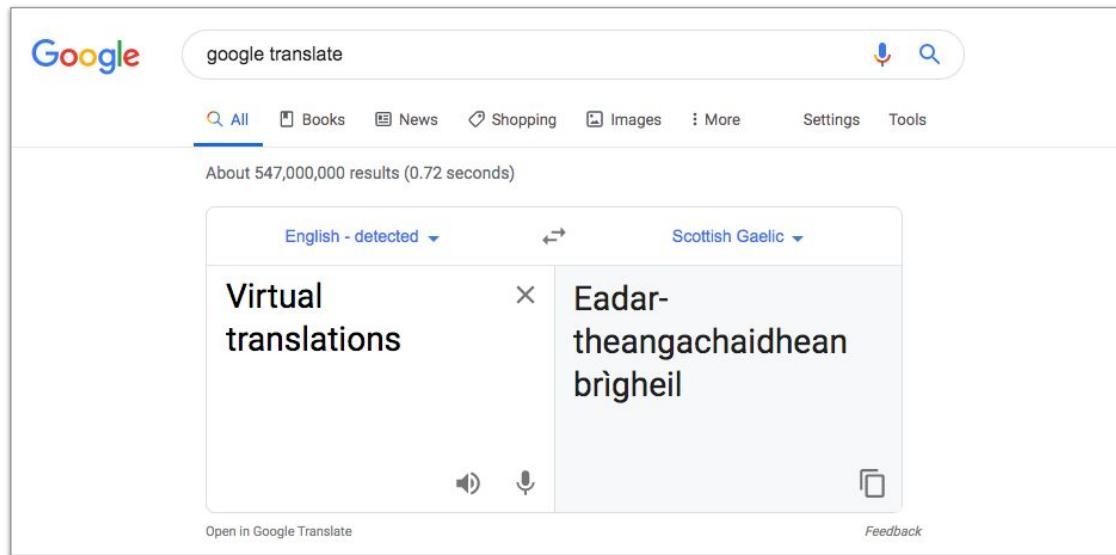


The screenshot shows a laptop screen displaying an email client interface with a sidebar containing various icons and sections like 'Inbox', 'Sent', 'Drafts', etc. A central pane shows several email messages, one of which is highlighted with a blue border. To the right of the laptop, there is a large, semi-transparent text overlay that reads 'Keep your inbox spam free.' Below this text is a green button with the white text 'TRY IT NOW'.

[spambully.com](http://spambully.com)

9

## Virtual translations (Google Translate)

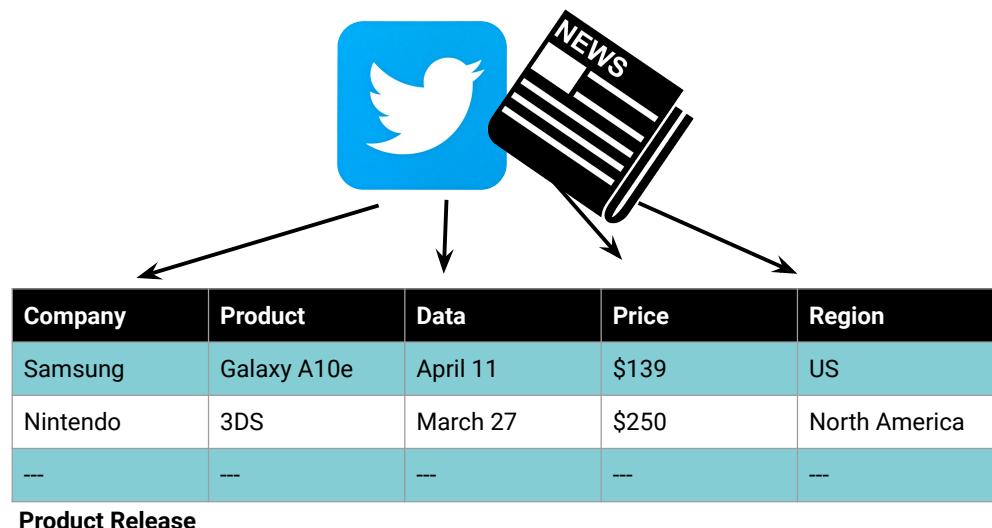


The screenshot shows a Google search results page for 'google translate'. Below the search bar, there are navigation links for 'All', 'Books', 'News', 'Shopping', 'Images', 'More', 'Settings', and 'Tools'. The main content area displays the search results with the text 'About 547,000,000 results (0.72 seconds)'. A prominent feature is a translation box at the bottom. On the left, it says 'English - detected' and on the right, it says 'Scottish Gaelic'. Between them is a double-headed arrow icon. Inside the box, the English phrase 'Virtual translations' is on the left and its translation 'Eadar-theangachaidhean brigheil' is on the right. Below the box are two small audio icons and a 'Feedback' link.

[google.com](http://google.com)

10

## Handling unstructured data from tweets and Facebook posts



11

## 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

12

## NLP In Finance

Automated sentiment analysis of earnings statements and investor calls



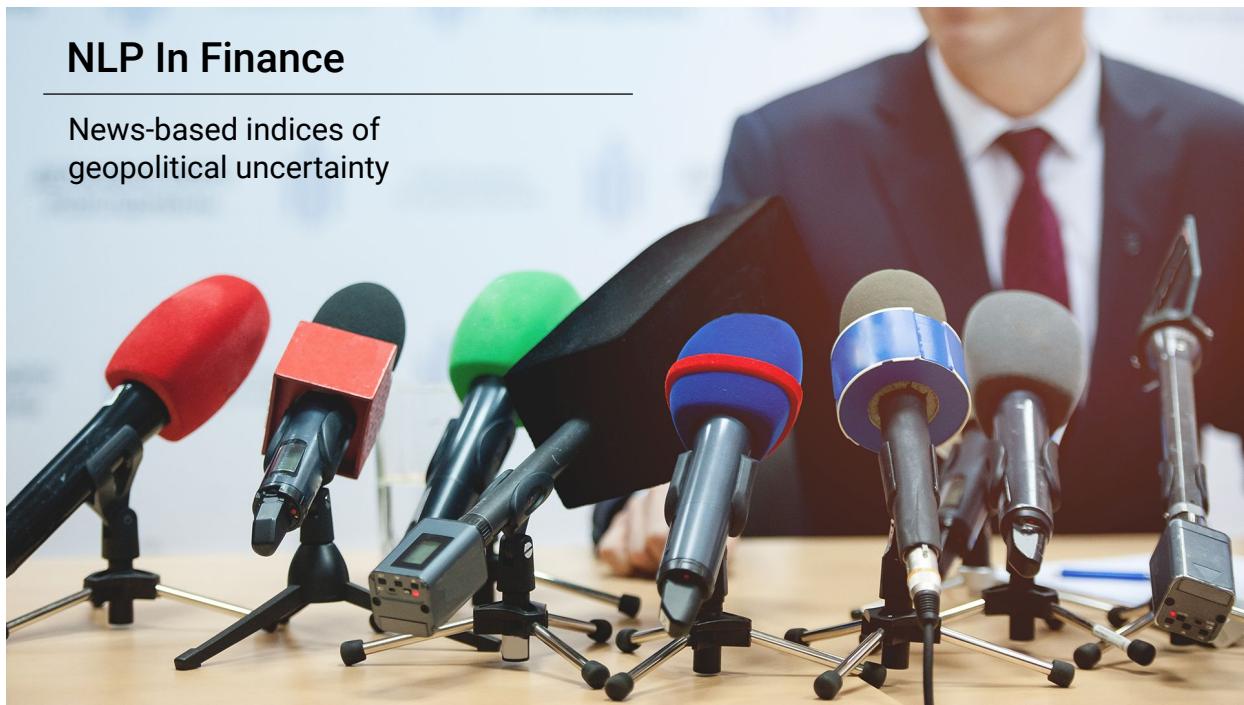
## NLP In Finance

Predictions of interest rate from Federal Reserve testimony



## NLP In Finance

News-based indices of geopolitical uncertainty

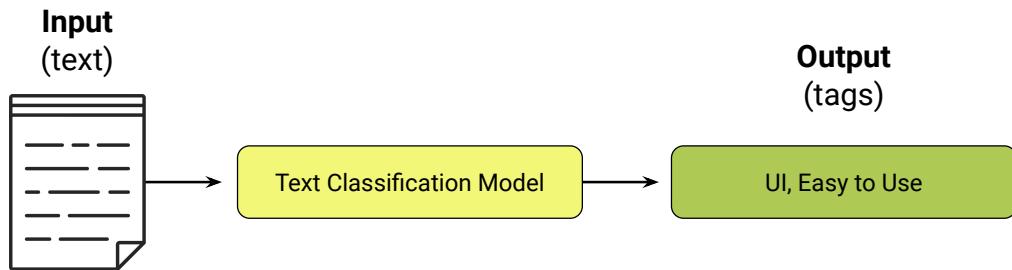


## A Few NLP Applications

## Text Classification

---

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



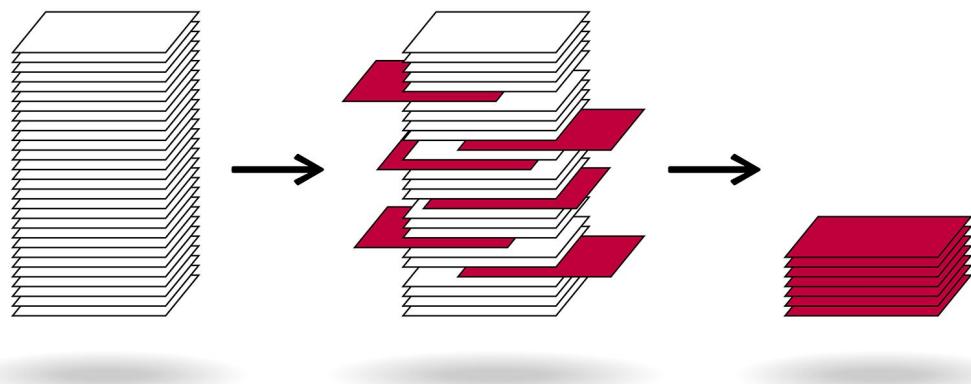
---

17

## Information Extraction

---

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



---

18

# Document Summarization

Generating a headline or abstract for a document

↑ 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]

Posted by u/Bruce-M OC: 8 1 year ago 3

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

A federal trademark judge has ruled in favor of a Pennsylvania-based genealogical who goes by the name Dr. Dna...Finding that use of this name does not violate the trademark of Dr. Dna, the federal court has rejected the case.

FURTHER READING Was granted other Gardner's license after over trademarks used

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

Applicant has admitted that DR. DRAK sounds identical to DR. DRE (Draugh Tr. at 154:20-155:1).

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

[reddit.com](https://www.reddit.com)

19

# Complex Question Answering

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

facebook Artificial Intelligence

Research | NLP

## Introducing long-form question answering

July 25, 2019 Written by Angela Fan, Yacine Jernite, Michael Auli

Share

[ai.facebook.com](https://ai.facebook.com)

20

**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

---

23



Tokenization

24

## 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!



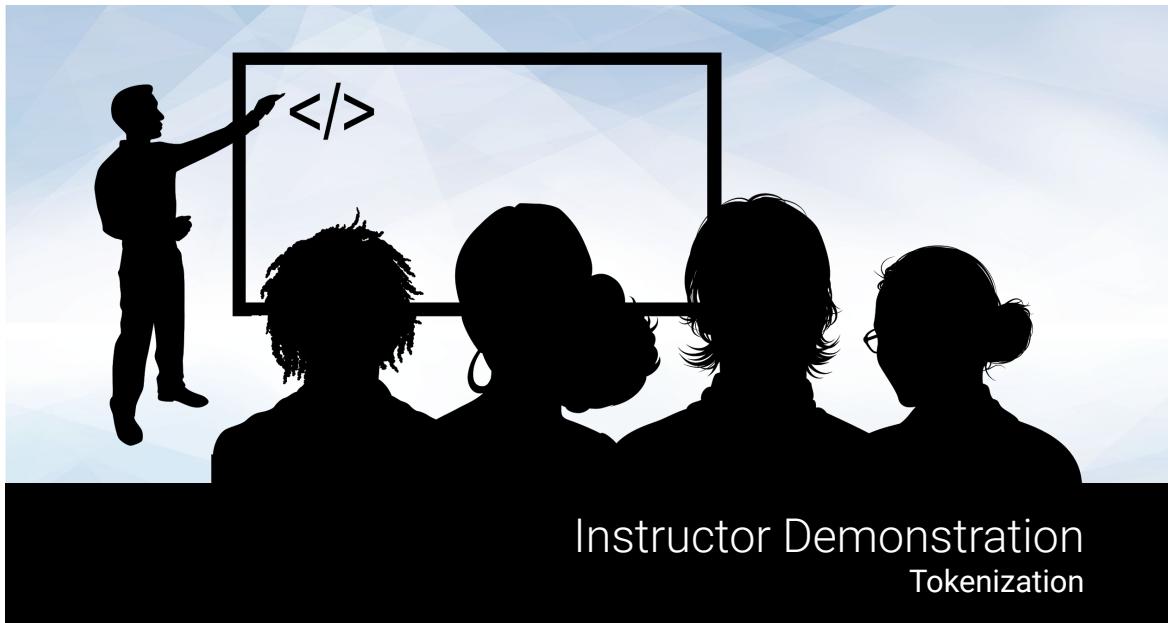
[“let’s”, “eat”, “grandpa”]



25



**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.



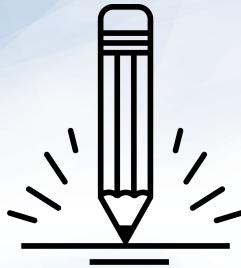
Generally, stopwords are the most commonly used words in the document.



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!

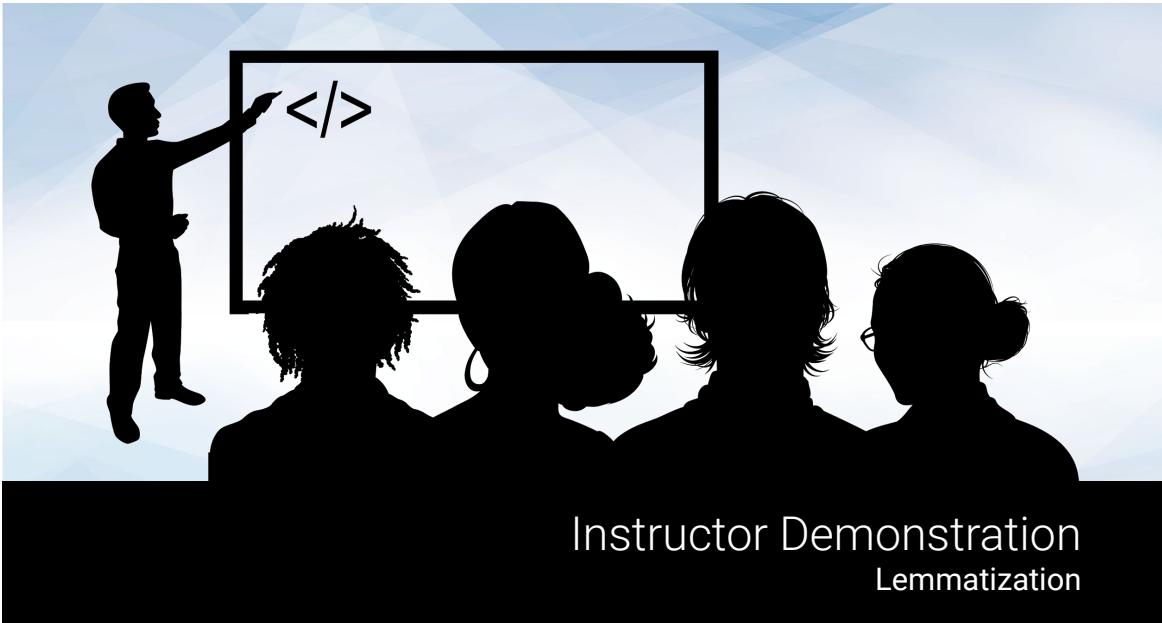
---

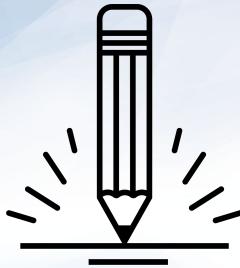


## Lemmatization



**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

41

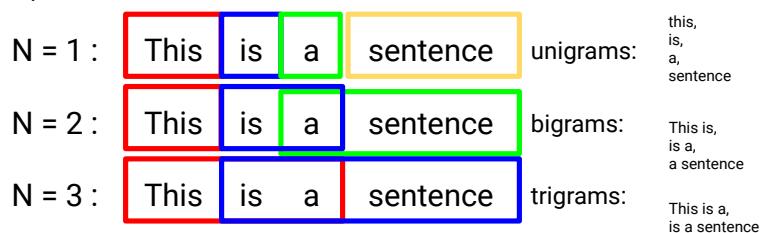


**N-Grams:** Tokens that include multi-word phrases. The n is the number of words—for example, bigrams are two-word combinations.

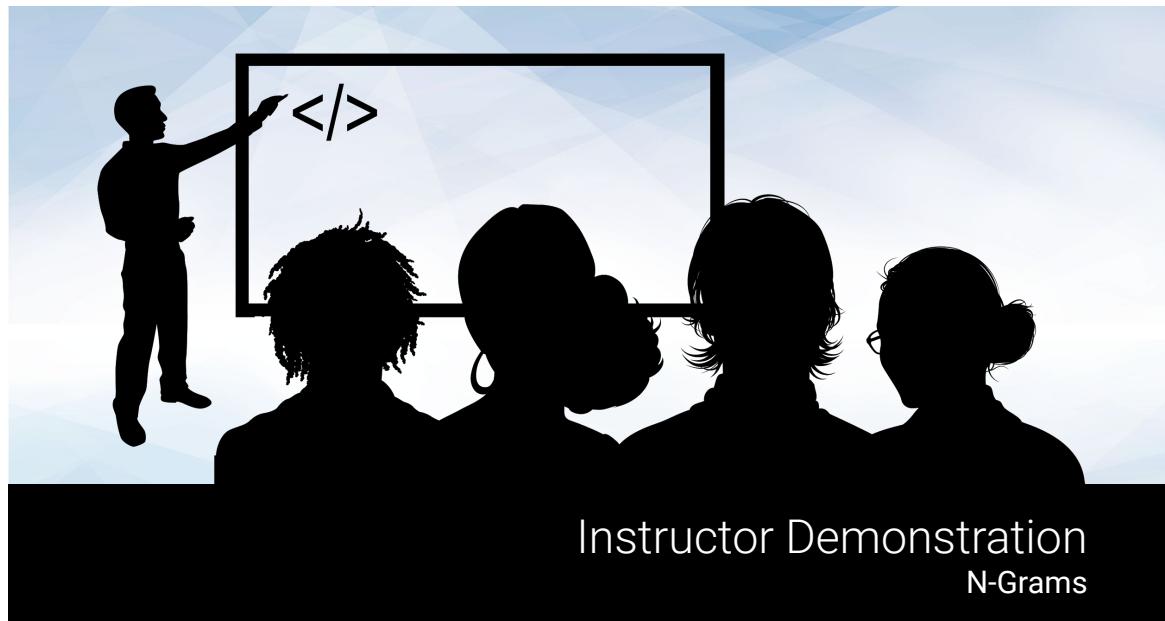
## N-Grams

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.



43





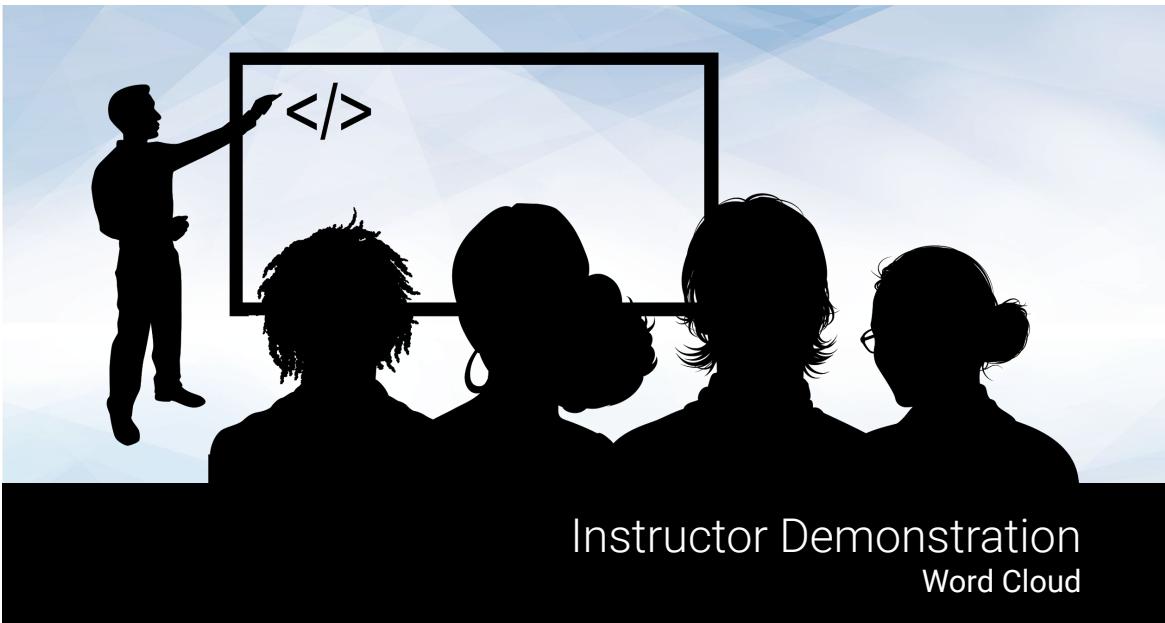
## Activity: Counter

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

A graphic icon of a pencil with three short lines radiating from its tip, positioned on the left side of the slide. To the right of the icon, the title 'Activity: Gas Cloud' is displayed in bold text. Below the title, a descriptive paragraph explains the activity. At the bottom right, there is a black bar containing text and a clock icon.

## 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?