



Washington University in St. Louis

COLLEGE OF ARTS & SCIENCES

## Day 1 - Introduction

---

Introduction to Text Analysis in Python  
TRIADS Training Series

Ishita Gopal

February 8, 2024

# Today

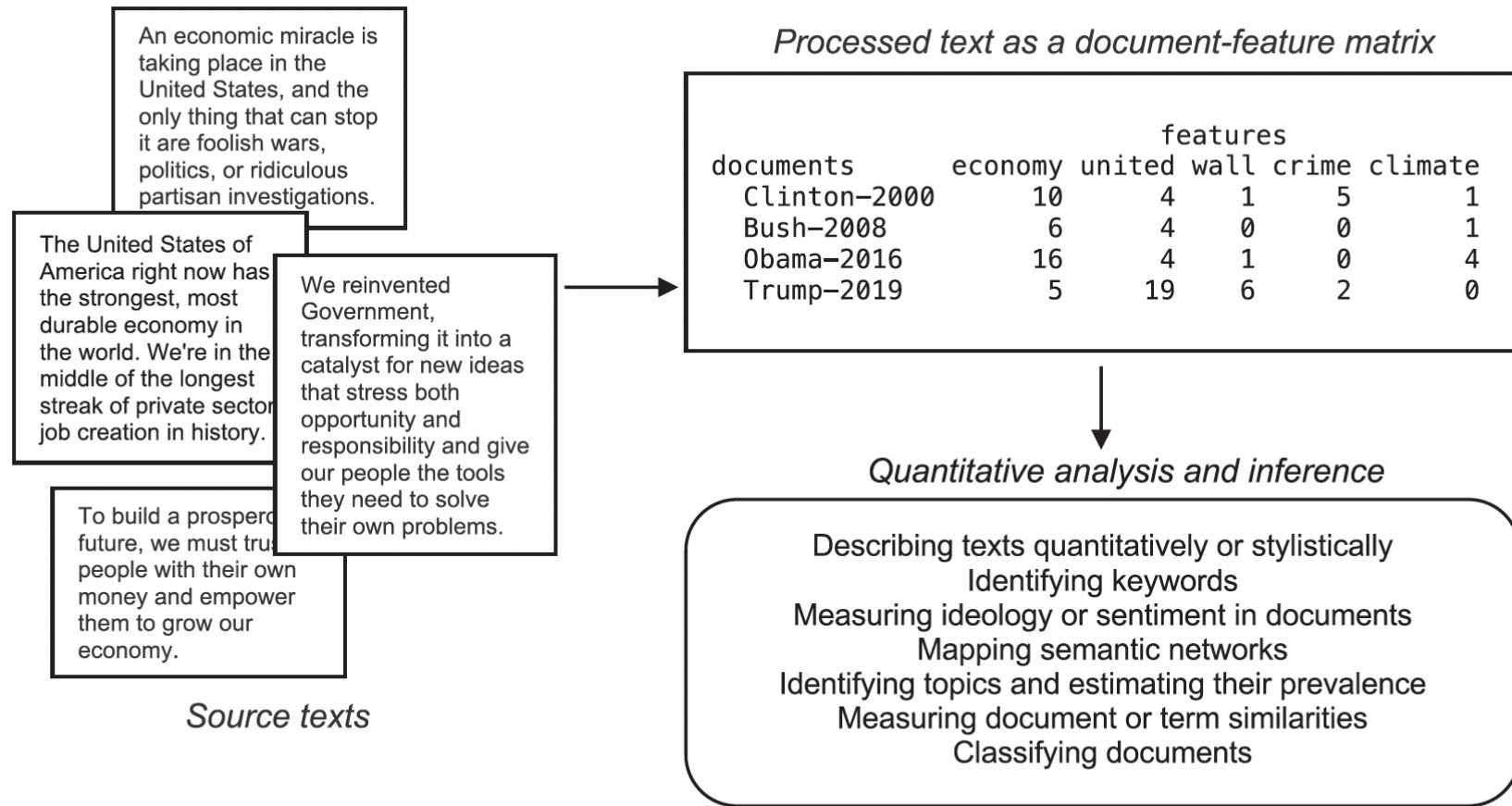
- Introduction
- Course Logistics
- Overview - Text Analysis
- Lab - String Manipulation

# Software

- We will be using Python.
- Python materials will be in notebooks and is provided through Google Colab.
- Google Colab allows you to modify and run copies of the notebooks without worrying about installations on your own machines.
- FYI, Google Colab provides free access to GPU and TPU computing, although it is overkill for this course, and we will not be using it.

# What we will cover

- Day 1: Overview of the goals and methods in text analysis, introduction to string manipulation.
- Day 2: NLP “pipelines”. General concepts and practical application of NLP labeling tasks like part-of-speech tagging, named entity recognition, Basics of bag-of-words (with emphasis on CountVectorizer).
- Day 3: Dictionary-based analysis (using tools like VADER), introduction to topic modeling methods (such as Latent Dirichlet Allocation or LDA)
- Day 4: Text classification (using Naïve Bayes)



**Figure 26.1 From text to data to data analysis**

Source: Benoit. 2020

## Typical text-as-data pipeline

(1) divides the text into the units we might care about,

(2) throws out some we're pretty sure we don't care about, and

(3) combines units that we're pretty sure should count as the same thing.

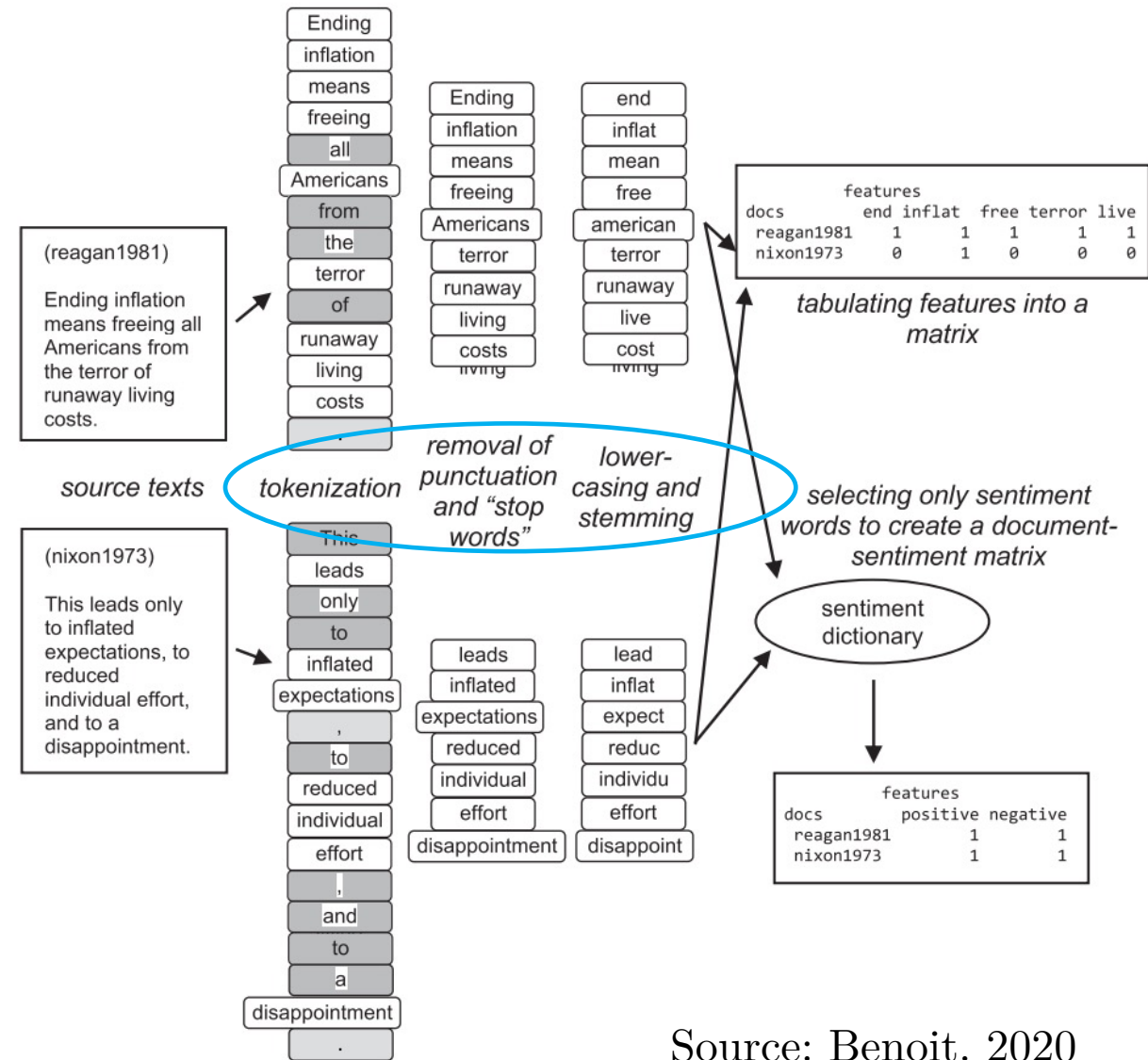
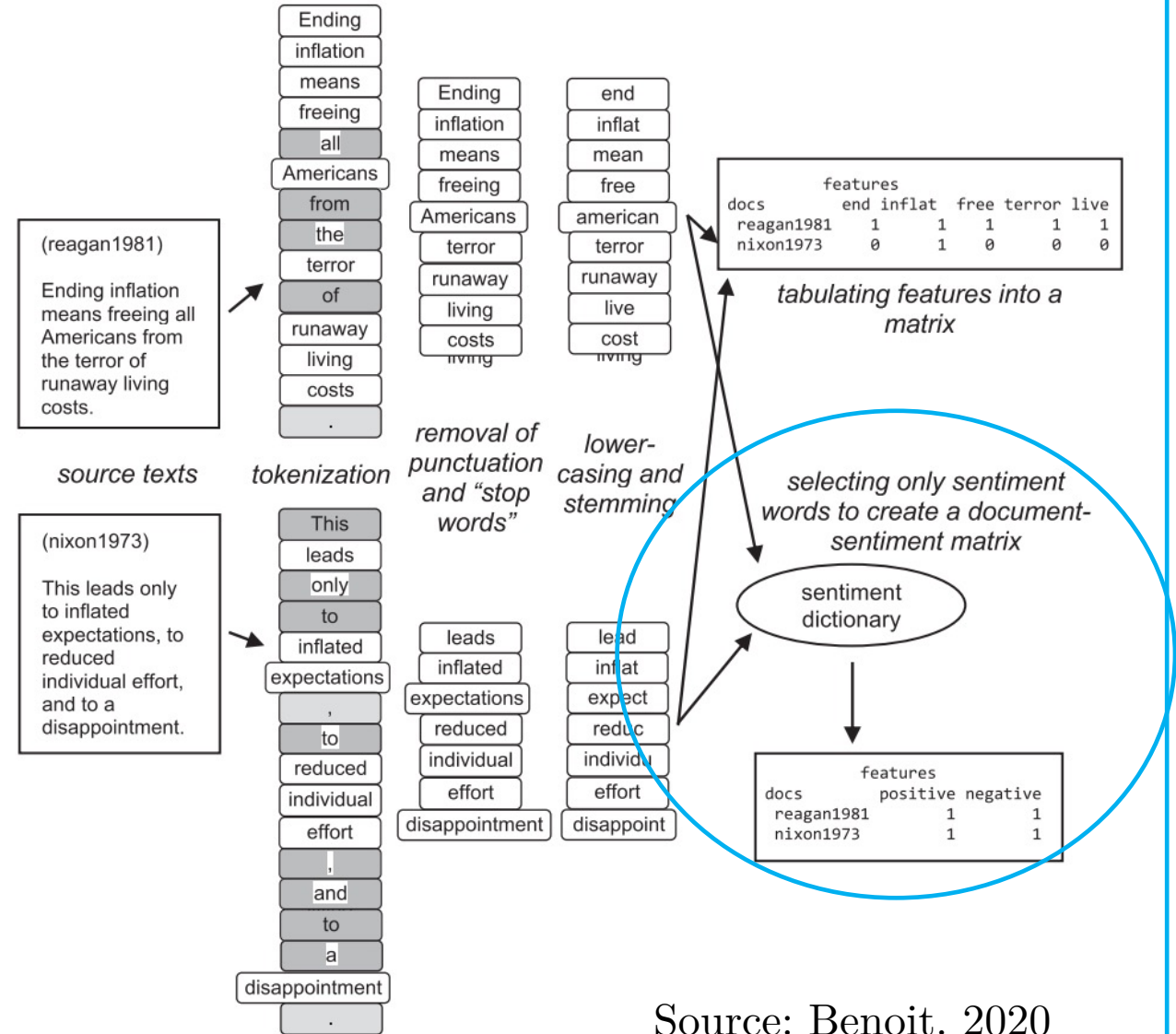


Figure 26.2 From text to tokens to matrix

# Dictionary Methods

- Most basic - you have a list of words → have a dictionary that maps a predetermined category/score for each word → you have the computer count/add them up.
- Very popular approach to sentiment analysis.
  - VADER
  - LIWC
  - Lexicoder



### Figure 26.2 From text to tokens to matrix

## **LIWC category**

## **Example words**

Positive emotions

Optimism

Happy, pretty, good

Ease, trust, hope

Negative emotions

Anxiety

Hate, worthless, enemy

Nervous, afraid, tense

Anger

Hate, kill, pissed

Sadness

Grief, cry, sad

Source: Settanni and Marengo. 2015



# Supervised Learning - Classification

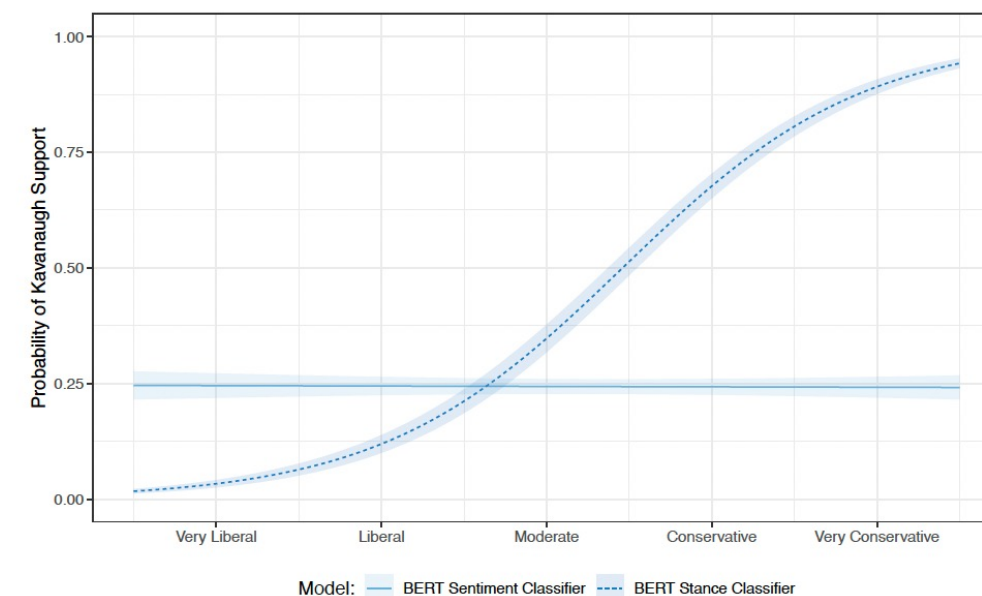
- We have labeled examples (documents) - training data.
- Input text (which can be a speech, tweet, email etc.) is paired with corresponding output labels (negative sentiment, mentions of COVID, identification as spam email).
- The task is to assign predefined categories or labels to input data.
- Much more flexible than dictionaries.

**Table 6.** Classifier Performance: Kavanaugh Tweets

Classifier	F1 Score (Predicting Sentiment)	F1 Score (Predicting Stance)
Lexicoder	0.788 (0.005)	0.572 (0.014)
VADER	0.754 (0.005)	0.514 (0.011)
SVM (sentiment-trained)	0.943 (0.003)	0.514 (0.012)
BERT (sentiment-trained)	0.954 (0.002)	0.582 (0.005)
SVM (stance-trained)		0.935 (0.006)
BERT (stance-trained)		0.938 (0.002)

Reported figures are the average F1 score over 5-fold cross validation  
Standard Errors in parentheses

**Figure 5.** Predicting Kavanaugh Support from Ideology



Source: Bestvater and Monroe, 2020.

*Sentiment* refers to the emotional tone in a piece of text.

*Stance* relates to the author's position a topic.

In the article above, the authors classify 1) if the tweet was positive or negative 2) if the tweet supports or opposes Kavanaugh's confirmation.

	Positive sentiment	Negative sentiment
Approving stance	<ul style="list-style-type: none"> <li>• #ConfirmKavanaugh this is a great man he will make a great Supreme Court judge</li> <li>• POWERFUL TESTIMONY of women that stand behind Brett Kavanaugh. Calling him Honorable with High Integrity, Family man, great friend and boss!</li> </ul>	<ul style="list-style-type: none"> <li>• The Republicans can certainly kiss the midterms goodbye if they blow this Kavanaugh confirmation</li> <li>• The sinister left will do absolutely anything to maintain power and attack conservatives . . . .Lie, cheat, delay. Whatever it takes. We absolutely cannot allow them to win. Confirm Kavanaugh!</li> </ul>
	<i>N</i> = 521	<i>N</i> = 1,467
Opposing stance	<ul style="list-style-type: none"> <li>• I hope she feels the love and support and the heartache that women feel in standing in solidarity with her</li> <li>• DR. FORD IS AN AMERICAN HERO</li> </ul>	<ul style="list-style-type: none"> <li>• Kavanaugh and the GOP have no idea of the power and anger they are unleashing</li> <li>• @SenateGOP withdraw Kavanaugh, you are just dragging yourselves, this country and Dr. Ford through the mud</li> </ul>
	<i>N</i> = 391	<i>N</i> = 1,281
Total	<i>N</i> = 3,660	<i>r</i> = 0.03

Source: Bestvater and Monroe. 2020

# Unsupervised Learning

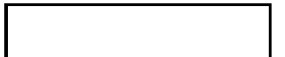
- We don't have labeled examples (documents) - no training data.
- This is often used as an exploratory / discovery exercise or for theory testing.
- Some of the main variants:
  - Clustering
  - Topic modeling
  - Word embeddings

# Unsupervised Learning - Topic Modeling

- Unlike a classification problem, you don't have pre-defined categories.
- But one main output of a topic model is a measure of how words are associated with the topic.
- When a topic model really works well, the appropriate label is obvious.
- A model of speeches in the US Senate contained the topic to the right ... what label would you give it?

school  
teacher  
educ  
student  
children  
test  
local  
learn  
district  
class  
account  
classroom  
achiev  
teach  
better

Table 4: *Key Words on*



# Today

## **Basics of String Manipulation in Python:**

1. Define a string
2. Strings and lists
3. String Manipulation
  - Concatenate
  - Format
  - Split
4. Substring Matching
5. Regular Expressions (if there is time!)

- Google Colab notebook linked below:  
<https://colab.research.google.com/drive/1RxcNLqLsAxl25CBdaRcdBtTFUKorCF1T?usp=sharing>
- Workshop folder is here:  
[https://github.com/IshitaGopal/TRIADS\\_workshops](https://github.com/IshitaGopal/TRIADS_workshops)
- *Slides and code are mostly from Burt L. Monroe's Text as Data Course.*