

AI Bootcamp

Advanced NLP Techniques– Text Extraction and Classification

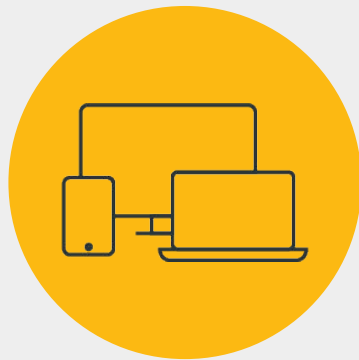
Module 20 Day 2



Class Objectives

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

- 1 Understand and apply TF-IDF to assess the importance of terms in text documents.
- 2 Implement a linear SVC model on text data and assess its performance as a binary classifier.
- 3 Implement an ML pipeline to vectorize and transform data.
- 4 Understand spaCy capabilities and where to find documentation.
- 5 Be able to use POS-tagged text to extract specific words.
- 6 Use dependency-parsed text to extract descriptors.
- 7 Extract specific types of entities from text.

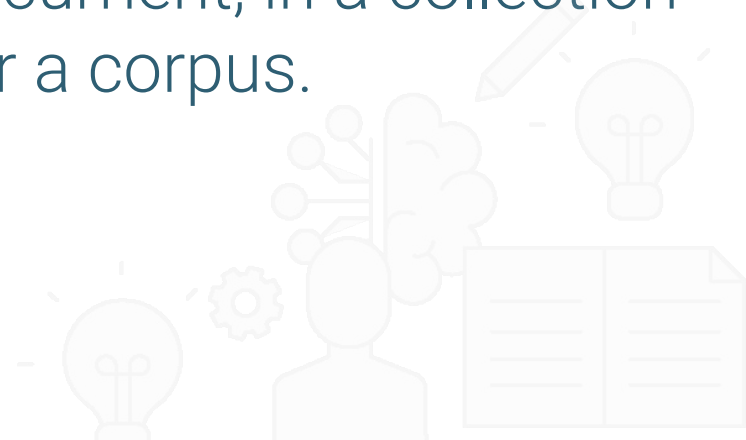


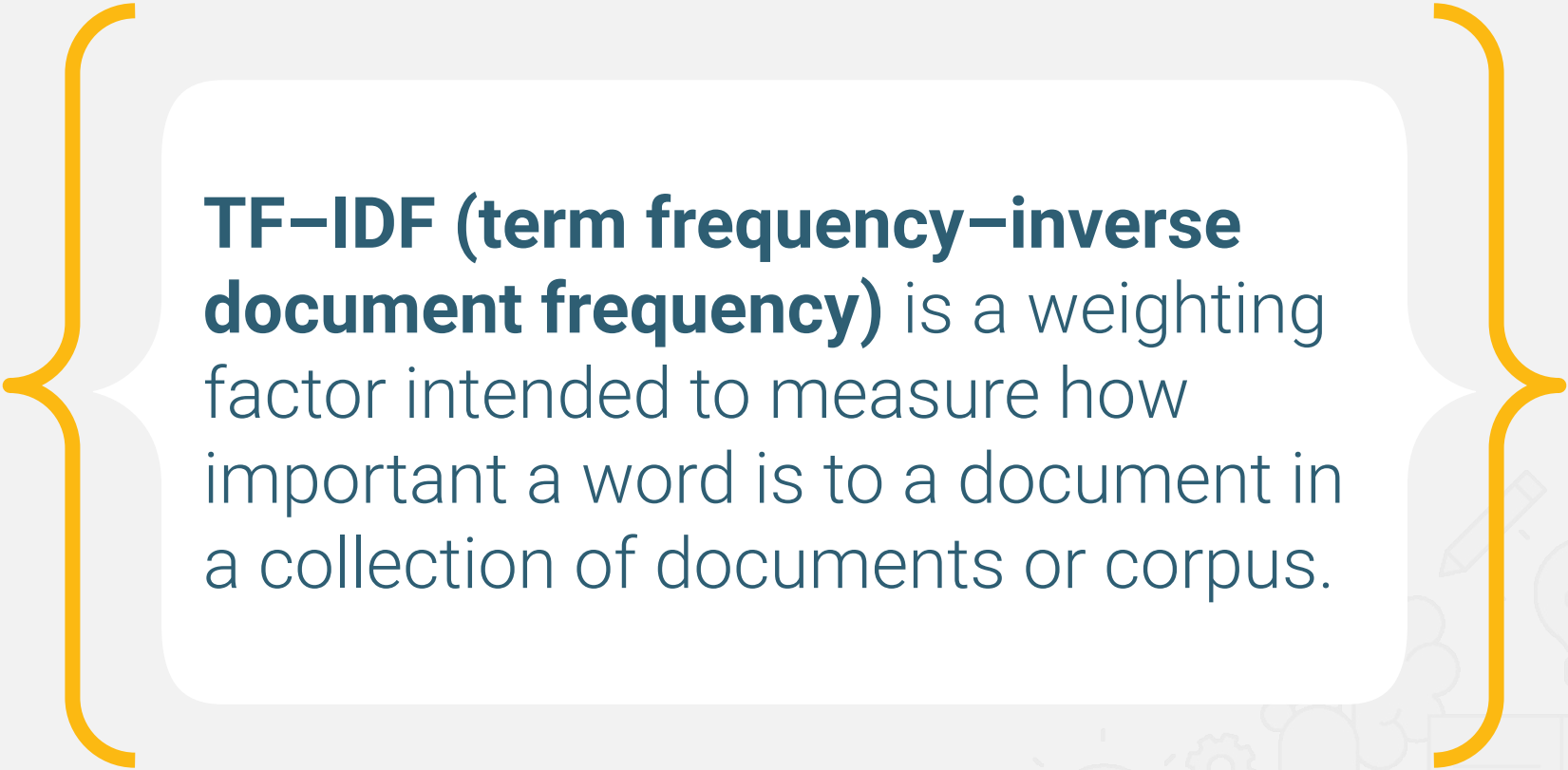
Instructor **Demonstration**

Understand Terms Relevance (TF-IDF)

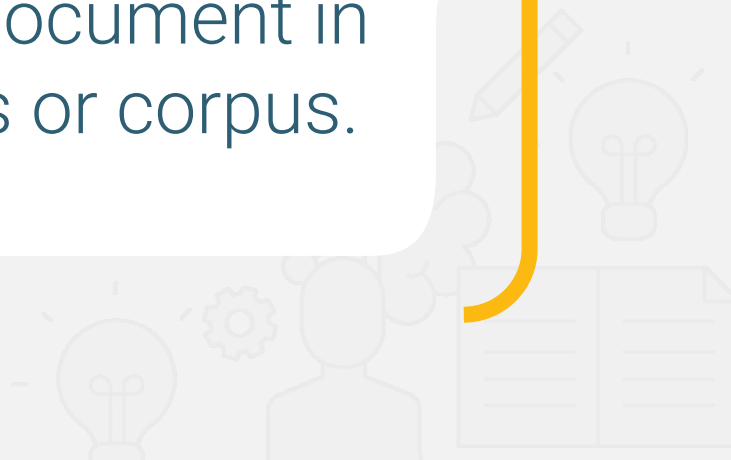


Terms relevance is quite important for sentiment analysis, since it offers a way to understand how important a word is to a document, in a collection of documents, or a corpus.





TF-IDF (term frequency-inverse document frequency) is a weighting factor intended to measure how important a word is to a document in a collection of documents or corpus.



TD-IDF

A measure intended to reflect the importance of a word in a text.

1 TF (term frequency): A count of a word in a document.

2 IDF (inverse document frequency):

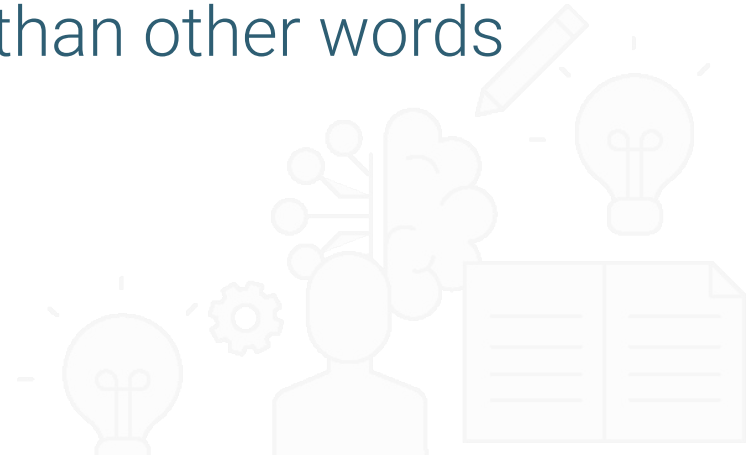
$$\log \left(\frac{\text{Total number of documents}}{\text{Number of documents containing target word}} \right)$$

3 IDF: The more documents that include the term, the lower the IDF score.

4 TF-IDF is the product of the two. TF drives up the score, but IDF will bring it down if the word occurs in all or many documents.

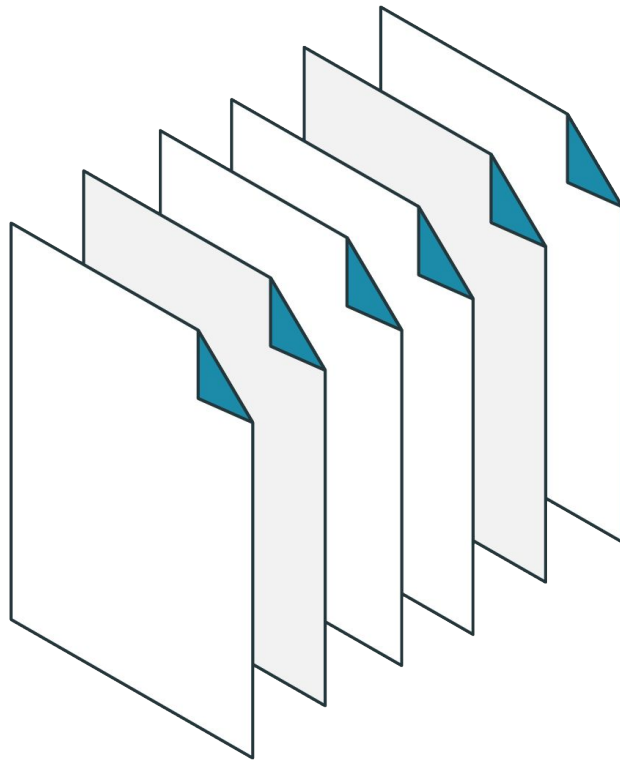


The rationale behind TF: it indicates how frequently a word appears in a document. A word that appears multiple times is likely more relevant and meaningful than other words in the same text.



The Rationale Behind TF-IDF

IDF comes into action when you're analyzing multiple documents. If a word also appears many times among a collection of documents, it may be a frequent word and not a relevant one.

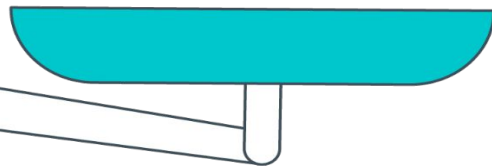


The Weights of TF-IDF on a Corpus

Terms with a **low TF** but a high document frequency in the corpus; normally this reflects that a term is commonly used across the corpus and that it could be less meaningful or valuable to an analysis.



Terms with a **high TF** but a low document frequency in the corpus; normally this reflects that a term is more meaningful or valuable to an analysis.



TF: Bag of Words

The methods for determining TF use a “bag-of-words” approach. Each document is represented by a “bag of words”, where grammar and word order are disregarded but multiplicity is kept.

The Bag-of-Words Representation

I love this movie! It's sweet but with satirical humor. The dialogue is great and the adventure scenes are fun...It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



It	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1



Activity:

Money Words

In this activity, you will use TF-IDF to find the most relevant words from a collection of Reuters news articles about money.

Suggested Time:

15 Minutes





Time's up!
Let's review



Questions?



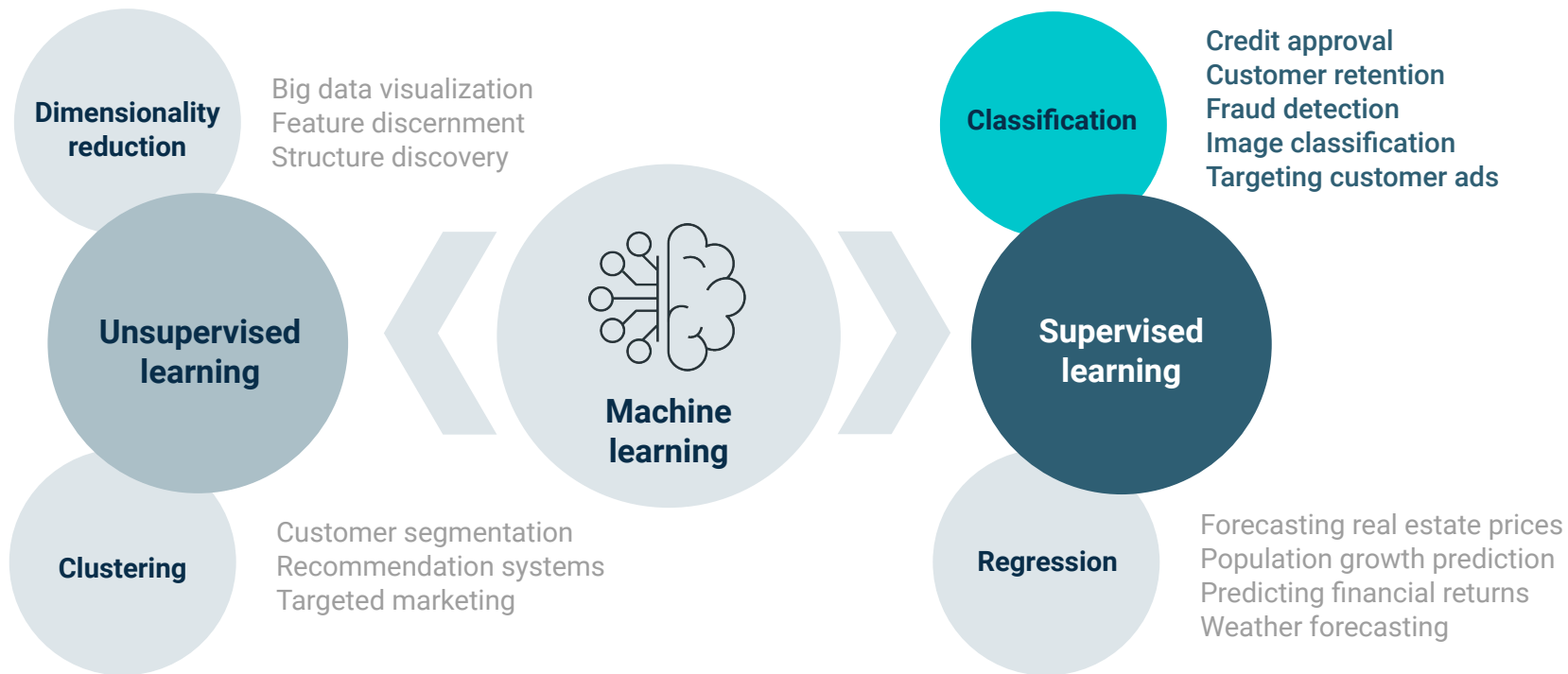


Instructor **Demonstration**

Text Classification

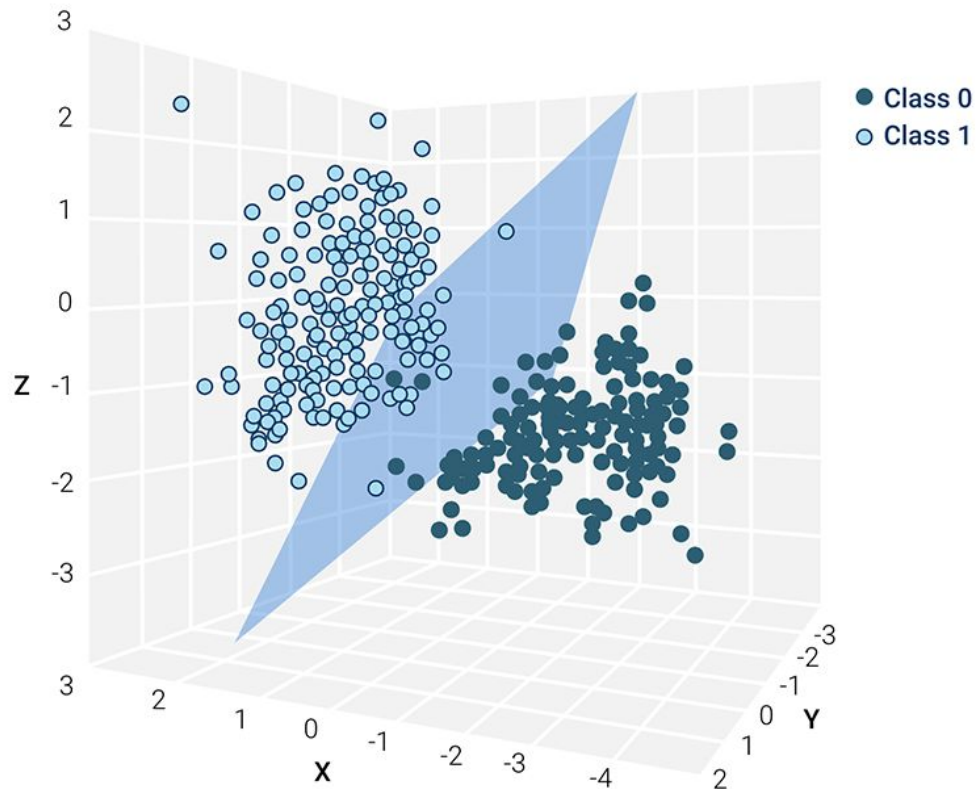
Types of ML

Automatic sentiment analysis is normally done using supervised machine learning (ML) techniques using classification algorithms such as decision trees, support vector machines (SVMs), naive Bayes, or neural networks.



SVM benefits

An SVM is able to arrange data points in multidimensional space. For example, consider the following visualization of a dataset with three features:





Activity:

Movie Review Classification

In this activity, you will determine if stopwords can affect the ability of a linear SVC model to predict the classification of a movie review.

Suggested Time:

25 Minutes





Time's up!
Let's review



Questions?





Break

15 mins



Instructor **Demonstration**

Introduction to spaCy

spaCy

spaCy

- Core functions depend on language models learned from tagged text
- Fast and flexible
- Designed specifically for production use

NLTK

- Core functions depend on language models learned from programmed rules
- Accurate
- Intended for educational and prototyping purposes

spaCy

We will be using spaCy for:



Part-of-speech (POS) tagging



Named entity recognition (NER)



Dependency parsing



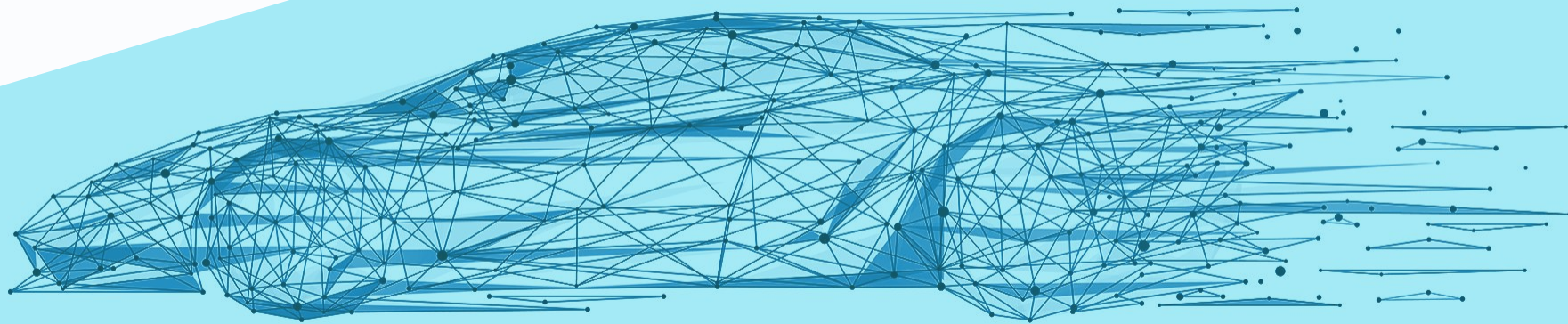
These tasks are more suitable for **model-based solutions**, because they are complex and depend highly on context.

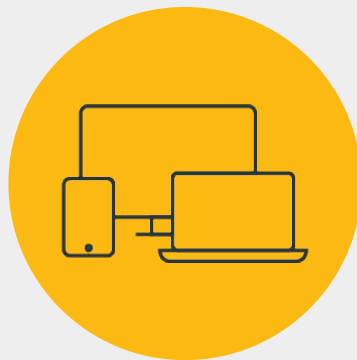


spaCy also provides tools for tasks like **tokenization** and **lemmatization**, which we've already learned about with NLTK, and creating word vectors.

spaCy

In comparison to NLTK, spaCy's language models trade accuracy for speed, so if the corpus is large, you may prefer a simpler, rule-based solution.



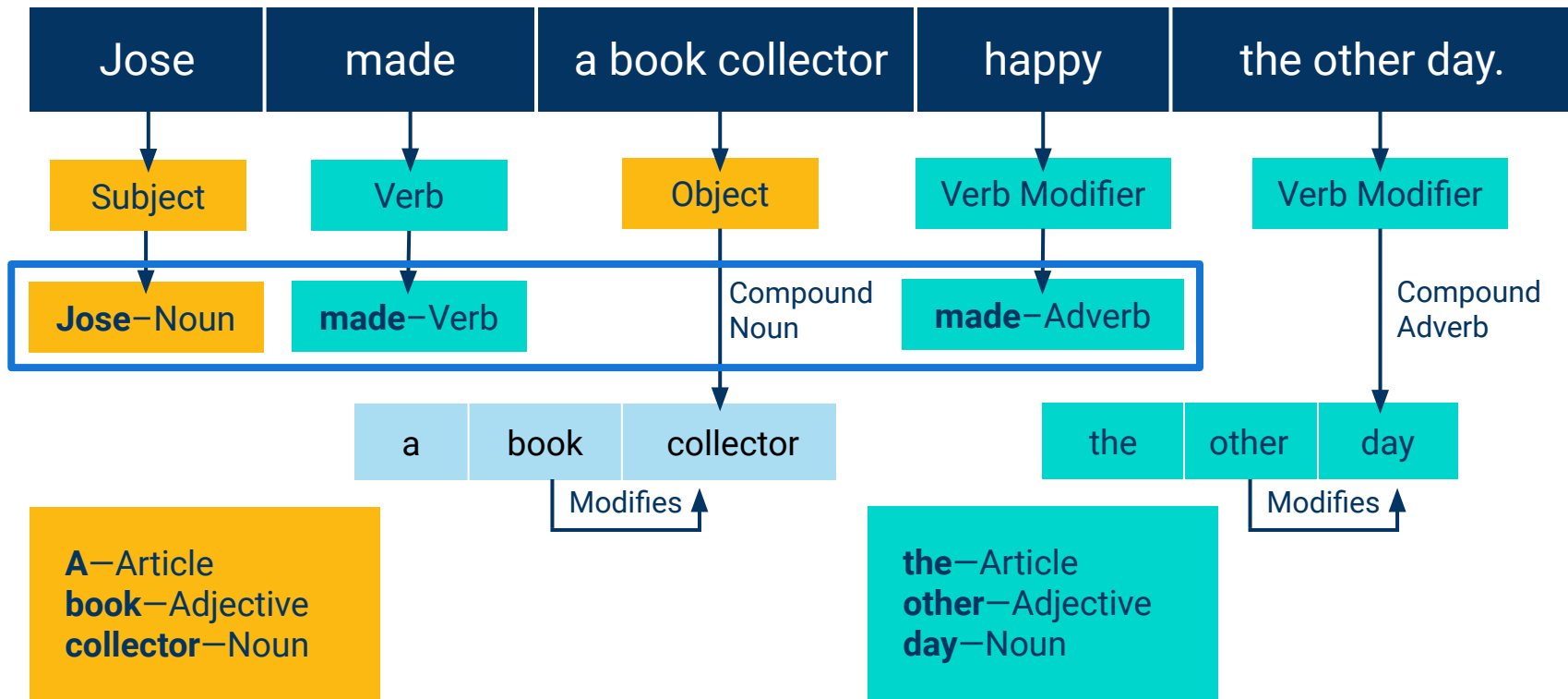


Instructor **Demonstration**

POS Tagging and Dependency Parsing

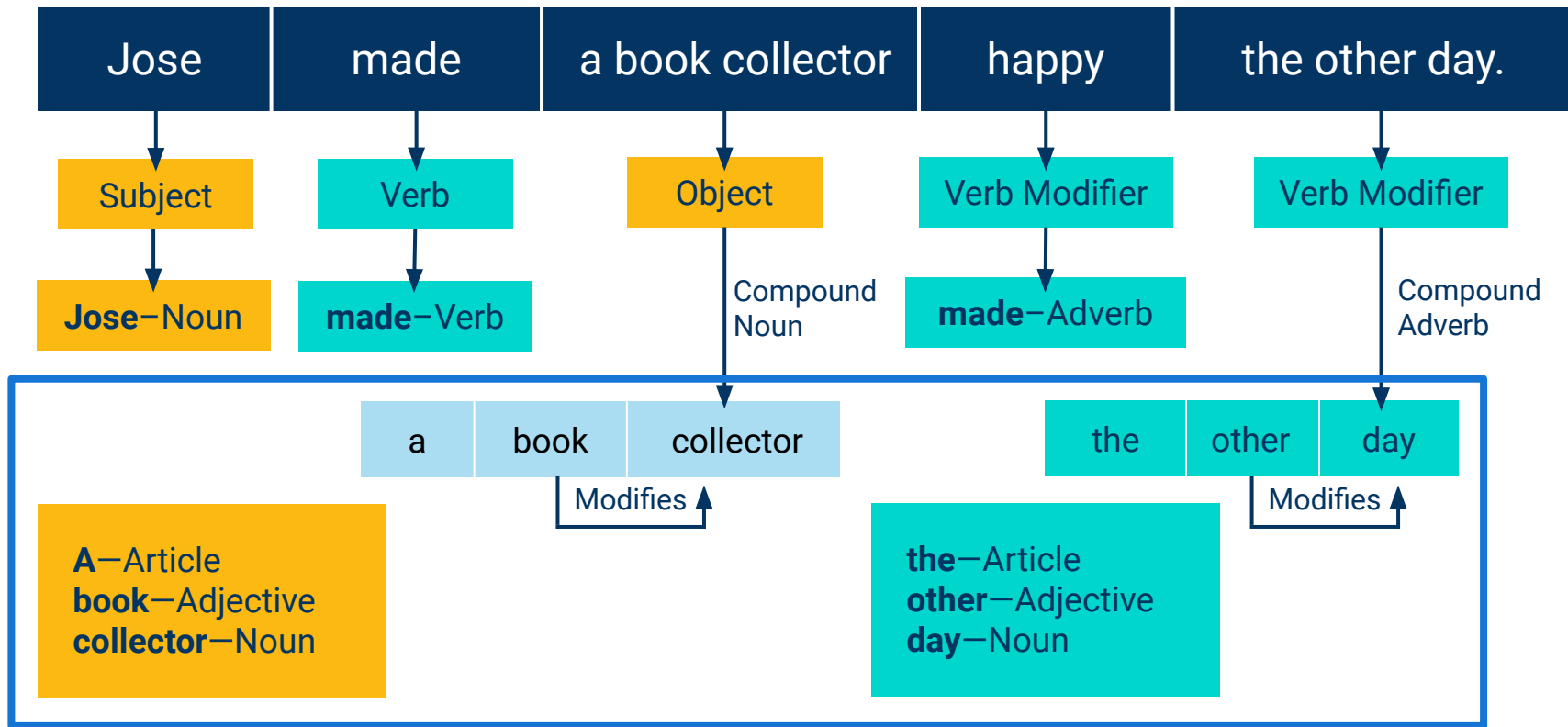
Part-of-Speech (POS) Tagging

POS tagging categorizes each word in a sentence by its grammatical role.



Part-of-Speech (POS) Tagging

POS tagging categorizes each word in a sentence by its grammatical role.





Activity:

Describing America

In this activity, you'll pair off and use NLTK and spaCy to identify the most common adjectives used by the Presidents' inaugural addresses since 1798. Then, you'll identify the most common adjectives used in the inaugural addresses to describe America.

Suggested Time:

20 Minutes





Time's up!
Let's review



Questions?





Named Entity Recognition (**NER**)



Named Entity Recognition (NER)

Extracting named entities, which include proper nouns and other specific types of nouns such as currencies, from a text.

COLUMBIA COFFEE REGISTRATIONS REMAIN OPEN

Columbia GPE's coffee export registrations remain open and there are no plans to close them since a new marketing policy means an unlimited amount can be registered, Gilberto Arango PERSON, president of the private exporters' association said.

"The philosophy of the new policy is not to close registrations. Nobody so far said may would be closed," he told Reuters ORG.

On March 13 DATE, Columbia GPE opened registrations for April DATE and May DATE for an unlimited amount.

Without giving breakdowns, Arango GPE said private exporters had registered 1,322,804 CARDINAL bags This calendar year DATE up to April 6 DATE, or Roughly 440,000 CARDINAL bags per month, slightly lower than the average in recent years DATE.



Activity:

Named Entity Recognition on Coffee

In this activity, you will code along with your instructor, who will prompt you on coding activities. These probing questions, combined with the whole group's participation in the demonstration, are meant to reinforce the concept of named entity recognition.

Suggested Time:

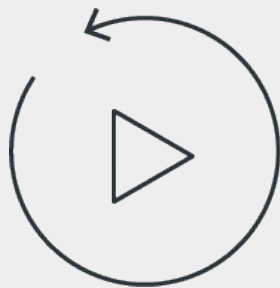
15 Minutes





Questions?





Let's **recap**



Recap

After today's lesson you are able to:

- 1 Understand and apply TF-IDF to assess the importance of terms in text documents.
- 2 Implement a linear SVC model on text data and assess its performance as a binary classifier.
- 3 Implement an ML pipeline to vectorize and transform data.
- 4 Understand spaCy capabilities and where to find documentation.
- 5 Be able to use POS-tagged text to extract specific words.
- 6 Use dependency-parsed text to extract descriptors.
- 7 Extract specific types of entities from text.



Next

In the next lesson, you'll will apply your NLP preprocessing skills to classify news headlines into categories using unsupervised learning for topic modeling. You will also train a deep learning LSTM RNN model on a large corpus of text to be used for text generation.



Questions?





The End