DATA SOCIETY:

Intro to text mining - Intro to NLP - 1

One should look for what is and not what he thinks should be. (Albert Einstein)

Intro to NLP: Topic introduction

In this part of the course, we will cover the following concepts:

- NLP definition and use cases
- Text analysis terminology

meIdR

Chat question

- Here are some reviews of the movie *Doctor Strange in the Multiverse of Madness* that a data analyst wants to explore
- When dealing with **text data** like reviews, what **variables** should the analyst consider?
- How **long** might it take to analyze the data?

Source: **rotten_tomatoes**

meldR

Chat question

- Here is a graph built using a text-mining algorithm
- What **categories** define the text?
- What **insights** does this visualization allow you to see?
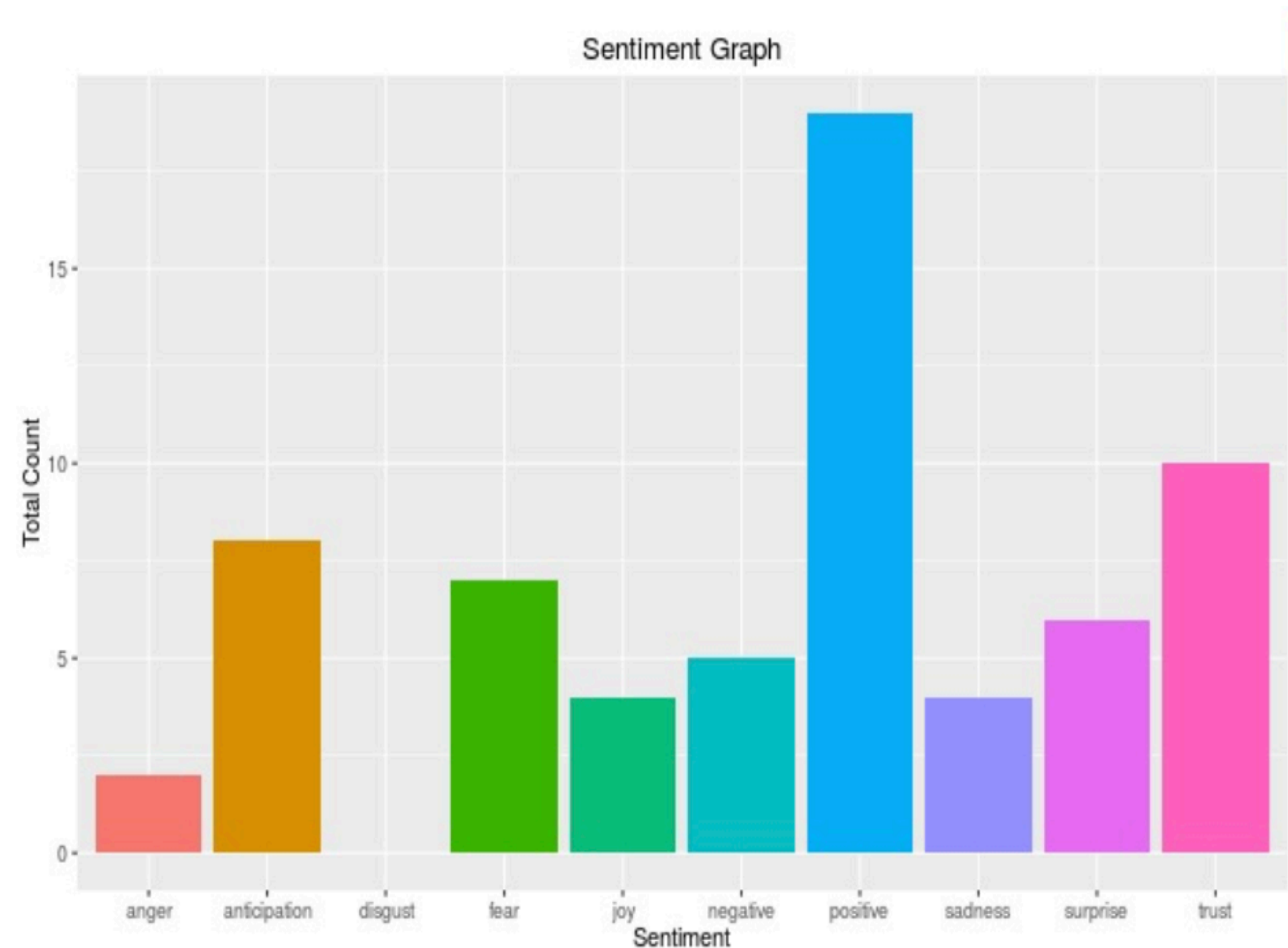
Source: *blog.journeyofanalytics*

meIdR

Module completion checklist

Objective	Complete
Explain the concept of NLP and how it is used in industry	
Discuss what a corpus is based on use cases	

meIdR

What are NLP and text analysis?

- **Natural Language Processing (NLP)** is a host of methods and disciplines, at the crossroad of :
  - data science
  - computational linguistics
  - artificial intelligence

- NLP's main focus is *processing all forms of human language (spoken or written) and getting insightful information out of it*
- **Text analysis is a part of NLP** and we can use it for written text

meldR

What is NLP used for?

- NLP is used for:
  - Generating human speech
  - Automating tasks like machine translation
  - Producing predictive models or detecting patterns in human language

Check out this **_article_** on an interesting application of NLP

meldR

Why text analysis?

- Most of the information and knowledge in the world is stored as text

Text analysis : What it is ?

- Text Analysis (TA) is an application of machine learning designed to derive meaningful information from unstructured textual data
- Text analysis solutions are deployed to efficiently process and analyze vast quantities of online data and documents, converting them into practical insights
- Text analysis can be employed to identify particular elements such as keywords, names, or corporate details within a massive volume of emails, or to classify feedback from surveys according to sentiment and subject matter
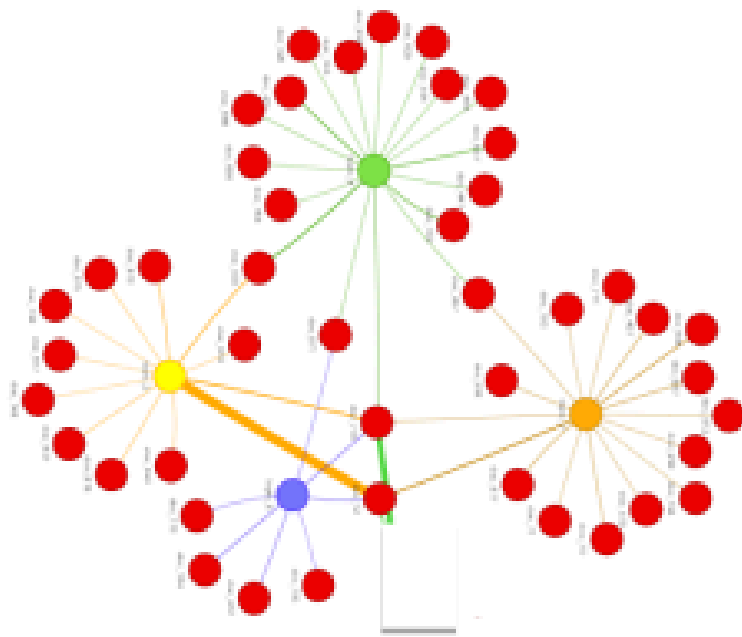
meldR

Text analysis: flow and basic methods

- Text analysis tools range from the simplest to the most advanced
- Here's an example of a flow of text analysis pipeline:

meIdR

Text analysis: more advanced methods

- Once you have scraped, cleaned, organized, visualized, and summarized data, more complex methods can be used

Classify documents

Extract entities

Analyze sentiment

meIdR

Text analysis: use cases

Field	Use case
Natural language processing	Discover patterns in speech; understand how human language works
Business	Help increase profits through analyzing product reviews; summarize and group reports, catalog documents
Psychology & Psychiatry	Analyze sentiments to detect and prevent dangerous social behavior
Network analysis	Augment network analysis to better understand how people interact within a group or react to a certain event

meIdR

Module completion checklist

Objective	Complete
Explain the concept of NLP and how it is used in industry	✔
Discuss what a corpus is based on use cases	

meldR

Units of text data

Words Sentences Documents

d r
w o

sentence
word is
This a not

To make a whole document.

This word is not a sentence.

I will type it anyway.

Nonsensical Document

This word is not a
sentence. I will type
it anyway. To make a
whole document.

Text = Document ?

meldR

Units of text data: corpus

Corpus

Doc 5
Doc 4
Doc 3
Doc 2
Doc 1

≈

Country as a whole has common characteristics:

-ing/to... -ing/to...
like infinitive
remember Gerund
be
see continue
stop would like

meIdR

Units of text data: corpora

Corpus = collection of documents
Latin word for "body"

Corpora = multiple document collections

We are corpora!

meldR
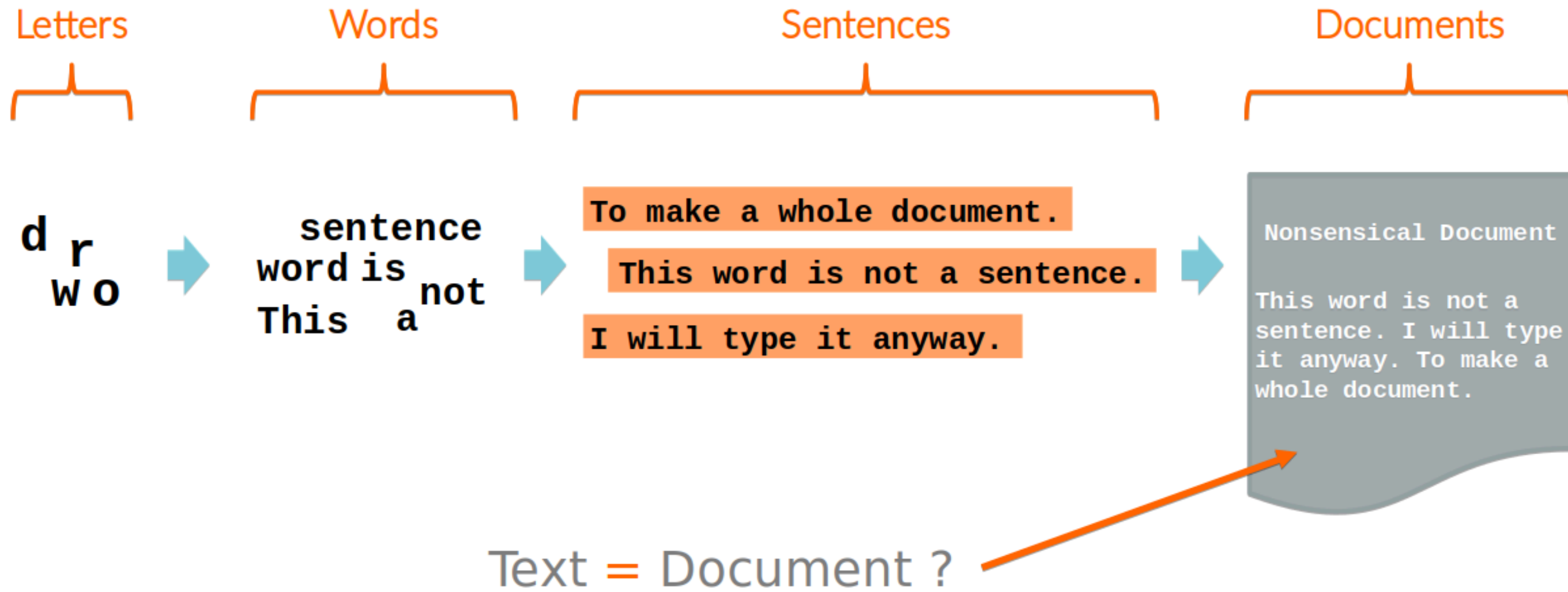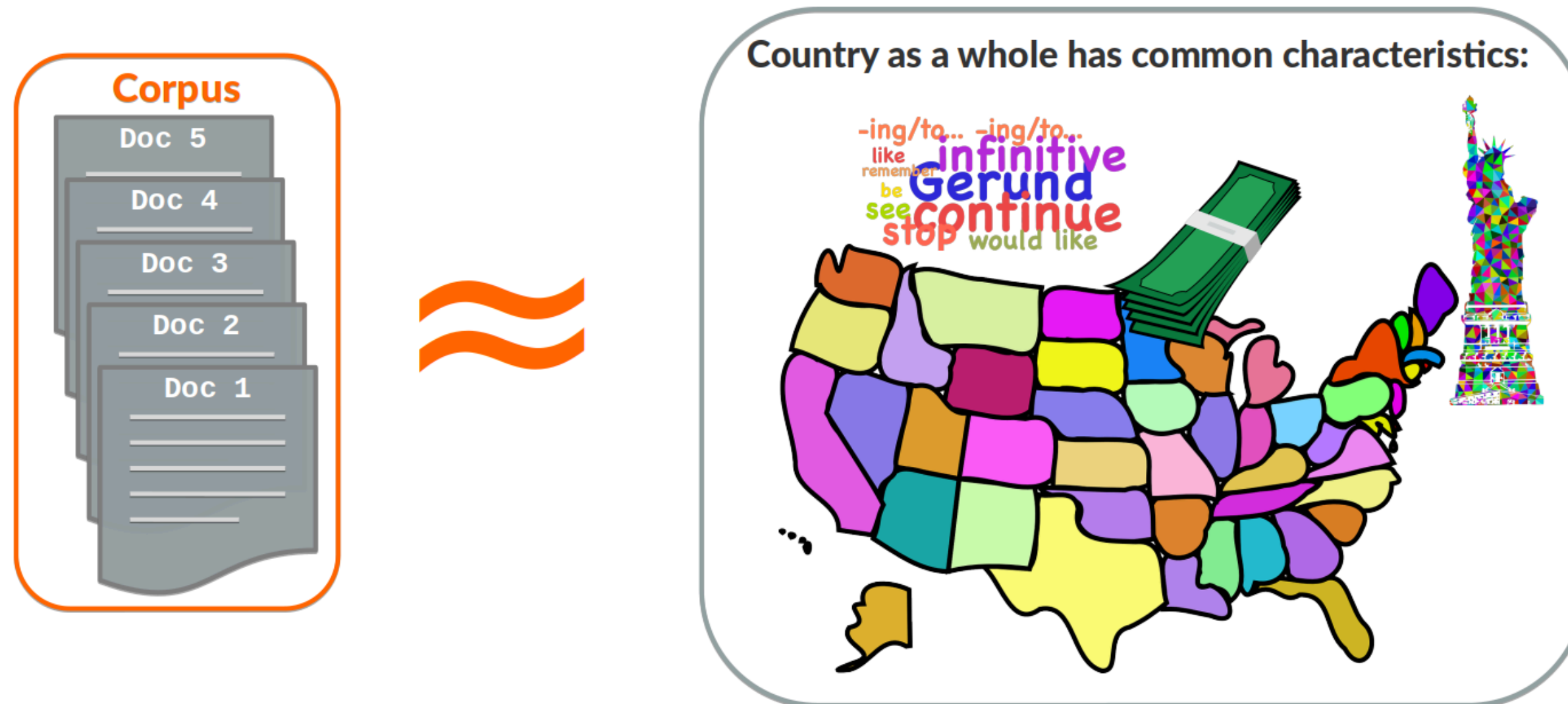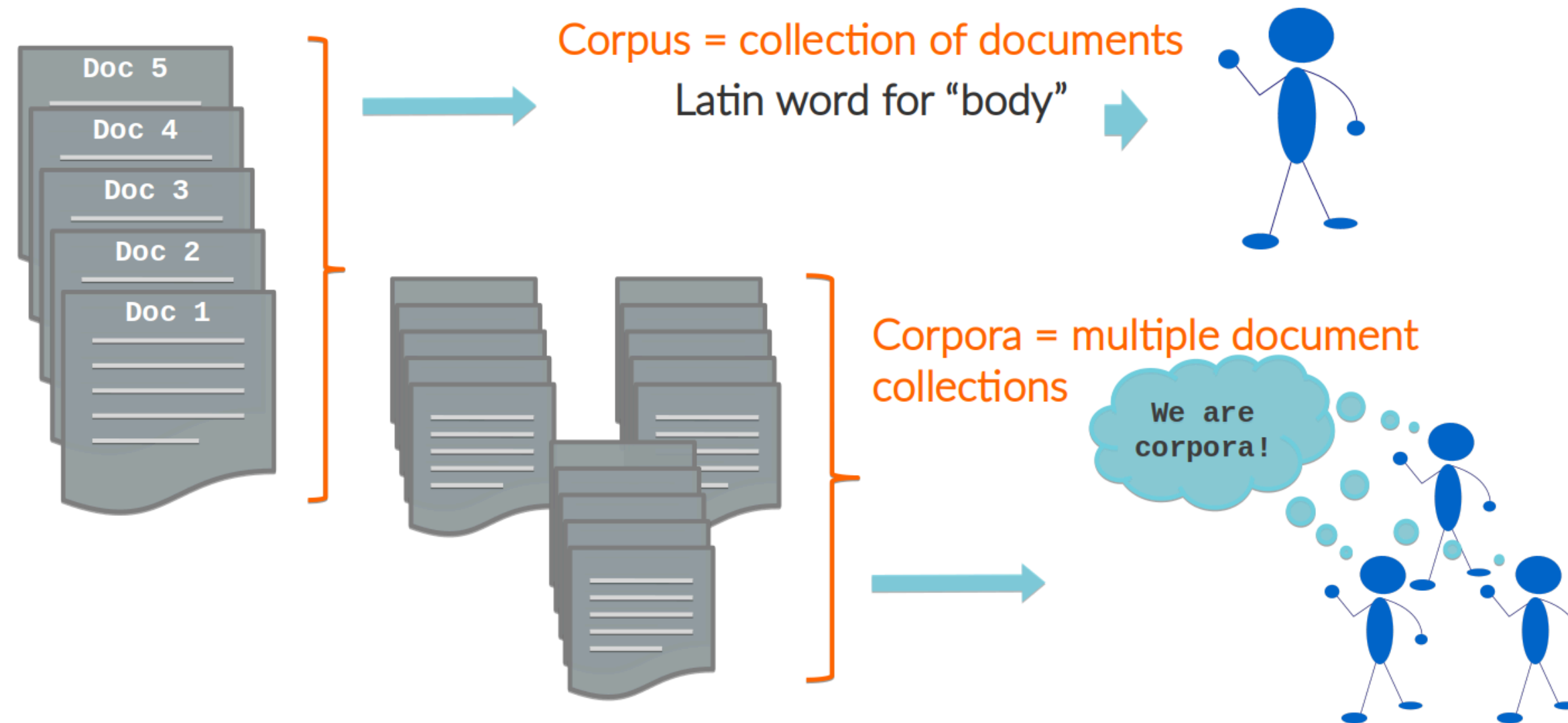
Knowledge check

Module completion checklist

Objective	Complete
Explain the concept of NLP and how it is used in industry	✔
Discuss what a corpus is based on use cases	✔

Intro to NLP: Topic summary

In this part of the course, we have covered:

- NLP definition and use cases
- Text analysis terminology

meldR

Congratulations on completing this module!

meIdR