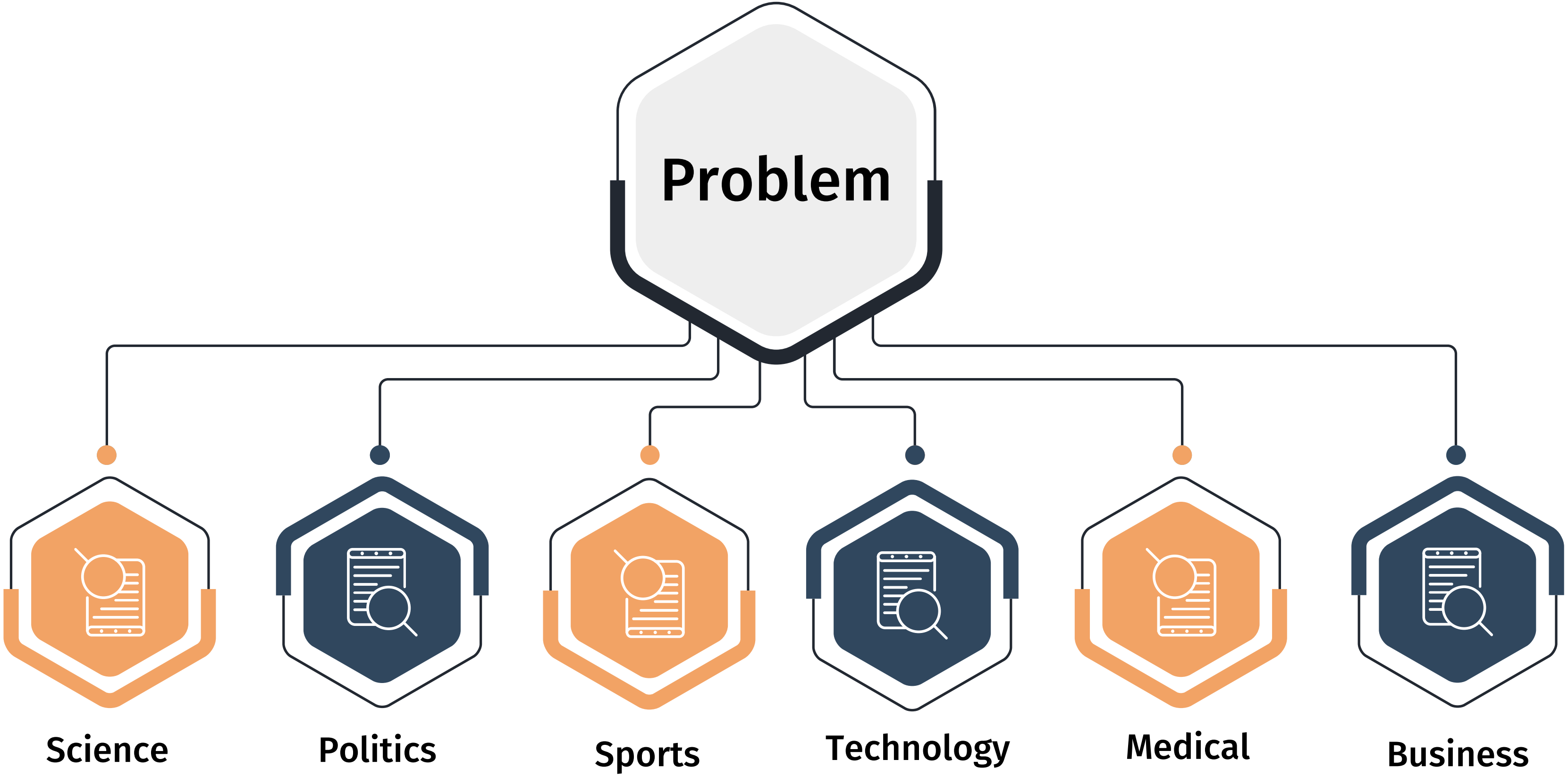
Topic Modeling

Latent Dirichlet Allocation



Problem
Science
Politics
Sports
Technology
Medical
Business

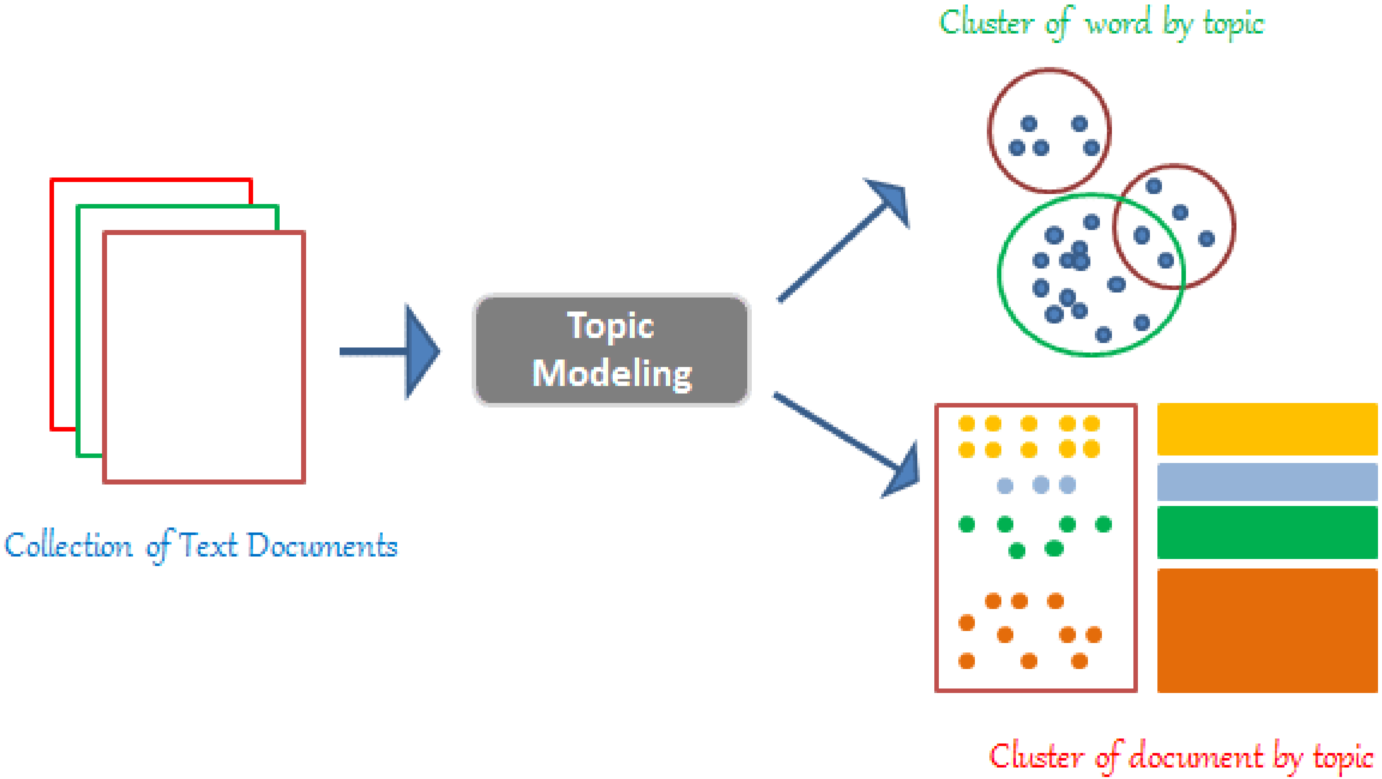
Latent Dirichlet Allocation (LDA)

It is an algorithm that has the ability to classify documents into topics.

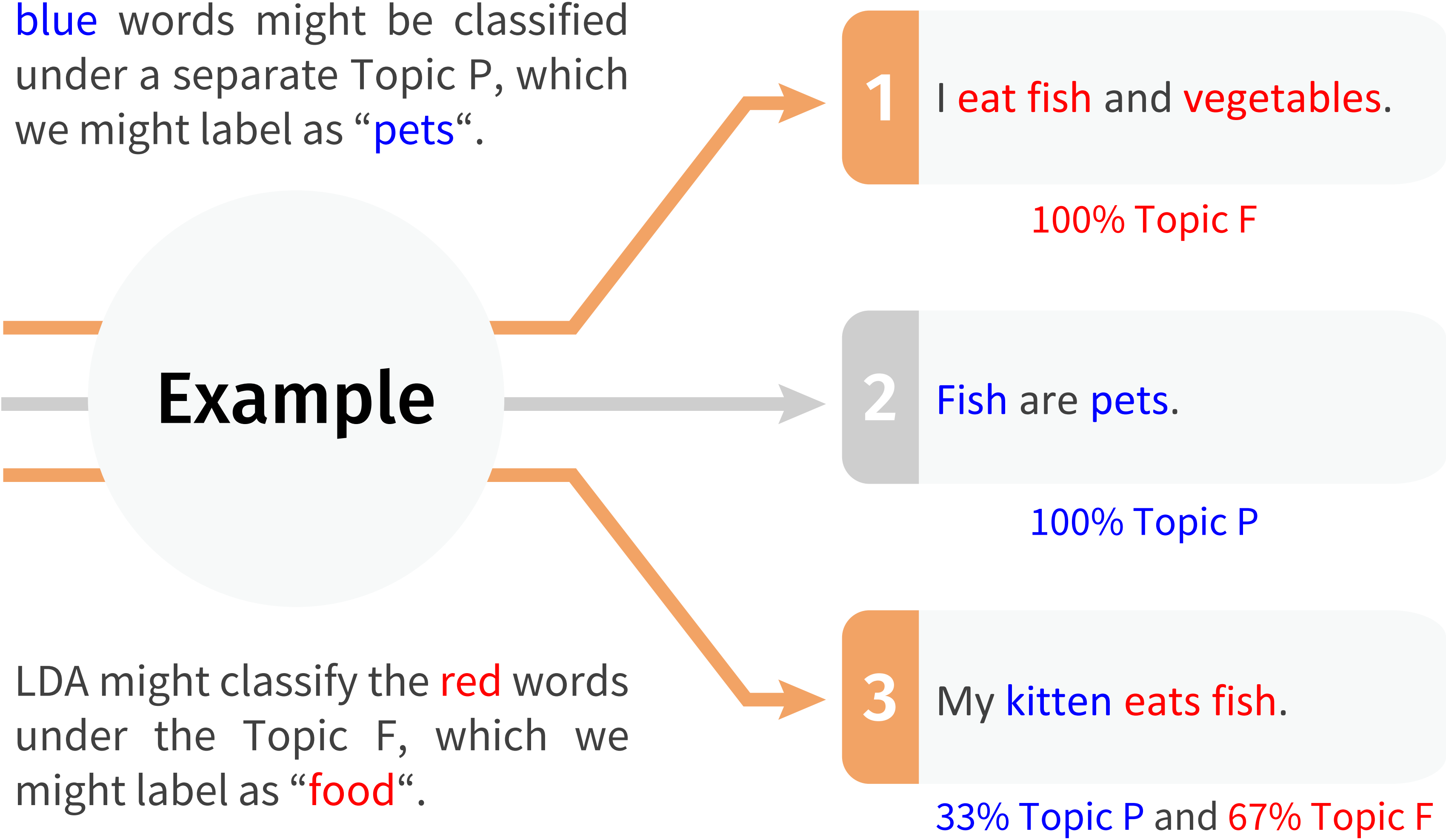
Latent topics: The topics present within a specific article. A specific article may contain 10 latent topics.

Similar articles use somewhat similar words.

LDA


Cluster of word by topic
Topic Modeling
Collection of Text Documents
Cluster of document by topic

blue words might be classified under a separate Topic P, which we might label as “pets“.



Example

1 I eat fish and vegetables.

100% Topic F

2 Fish are pets.

100% Topic P

3 My kitten eats fish.

33% Topic P and 67% Topic F

LDA might classify the red words under the Topic F, which we might label as “food“.

Steps

01

You tell the algorithm how many topics you think there are.

02

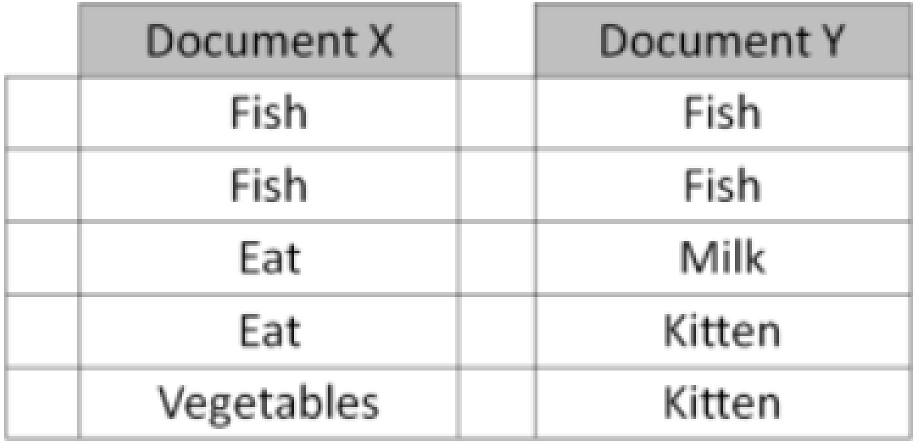
The algorithm will assign every word to a temporary topic.

03

The algorithm will check and update topic assignments, iteratively

Example (cont.)

- Imagine you have 2 documents with the following words:

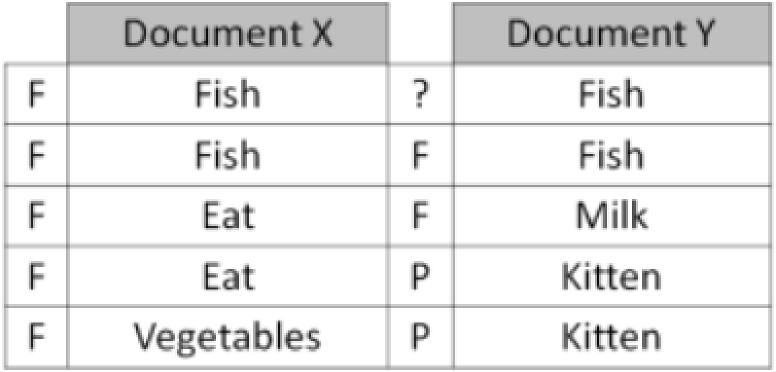


	Document X	Document Y
	Fish	Fish
	Fish	Fish
	Eat	Milk
	Eat	Kitten
	Vegetables	Kitten

Example (Cont.)

- For each word, its topic assignment is updated based on two criteria:
- How prevalent is that word across topics?
- How prevalent are topics in the document?

Imagine that we are now checking the topic assignment for the word "fish" in Doc Y:

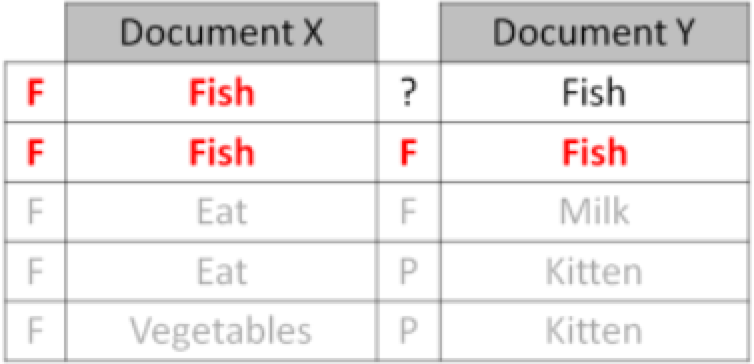


	Document X		Document Y
F	Fish	?	Fish
F	Fish	F	Fish
F	Eat	F	Milk
F	Eat	P	Kitten
F	Vegetables	P	Kitten

Example (cont.)

- **How prevalent is that word across topics?**

Since “fish” words across both documents comprise nearly half of remaining Topic F words but 0% of remaining Topic P words, a “fish” word picked at random would more likely be about Topic F.

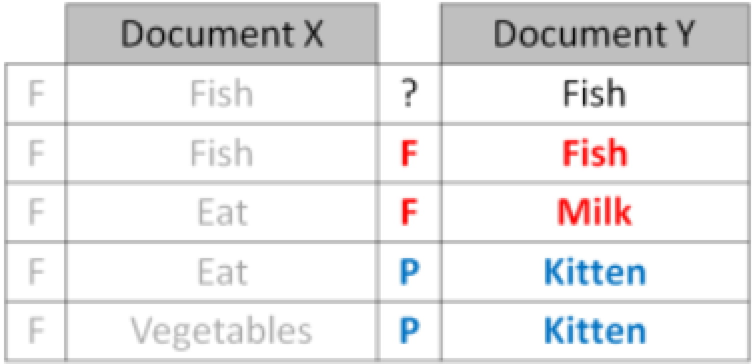


	Document X		Document Y
F	Fish	?	Fish
F	Fish	F	Fish
F	Eat	F	Milk
F	Eat	P	Kitten
F	Vegetables	P	Kitten

Example (cont.)

- **How prevalent are topics in the document?**

Since the words in Doc Y are assigned to Topic F and Topic P in a 50-50 ratio, the remaining “fish” word seems equally likely to be about either topic.



	Document X		Document Y
F	Fish	?	Fish
F	Fish	F	Fish
F	Eat	F	Milk
F	Eat	P	Kitten
F	Vegetables	P	Kitten

- Weighing conclusions from the two criteria, we would assign the “fish” word of Doc Y to Topic F. Doc Y might then be a document on what to feed kittens.

What is BERT?

- BERT (“Bidirectional Encoder Representations from Transformers”) is a popular large language model created and published in 2018. BERT is widely used in research and production settings Google even implements BERT in its search engine.
- By 2020, BERT had become a standard benchmark for NLP applications with over 150 citations. At its core, it is built like many transformer models. The main difference between transformer models and Recurrent Neural Networks (RNNs), another classic in the NLP toolkit, is that they process the input all at once.
- The original BERT language model was trained on over 800 million words from BooksCorpus and over 2.5 billion words from Wikipedia. It was originally trained on two tasks: language modeling and next sentence prediction.

What is BERTopic?

- BERTopic is an open-source library that uses a BERT model to do Topic Detection with class-based TF-IDF procedure.
    - TF-IDF stands for “Term Frequency - Inverse Document Frequency”. TF-IDF is an algorithm that weights the importance of words in a corpus, exactly as the name implies. The more frequently a word appears in a document, the more important it is. However, the more you see that word across documents, the less important it becomes.

Apply topic modeling using python

USING LDA

- Link of colab

USING BERTTOPIC

- Link of colab

Thanks for your attention