

with Dr. Mahdi Roozbahani & Wafa Louhichi



Learning Objectives

In this lesson, you will learn another discrete text representation

- Explain text data into numerical format using TF-IDF
- Understand the advantages and disadvantages of TF-IDF



Why Do We Need TF-IDF?

 We learned that the bag-of-words approach does not provide a logical importance for words

For example: "This is the NLP class."

- All the words have the same importance here. "the" is as important as "NLP" word
- TF-IDF will help us assign more logical importance to a vector of words for each document



Why Do We Need TF-IDF?

 We learned that the bag-of-words approach does not provide a logical importance for words

For example: "This is the NLP class."

- All the words have the same importance here. "the" is as important as "NLP" word
- TF-IDF will help us assign more logical importance to a vector of words for each document



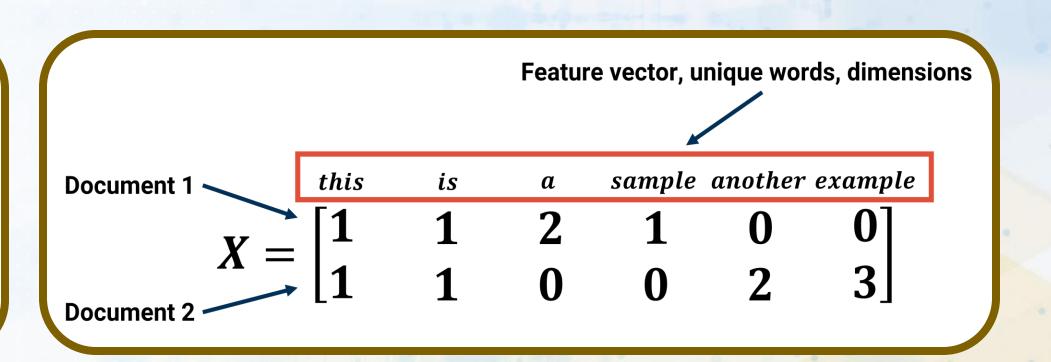
What is TF-IDF and when to use it?

- A word's importance score in a document, among N documents
- Everywhere you use "word count", you can likely use TF-IDF



Doc	ument 1
Term	Term Count
this	1
is	1
а	2
sample	1

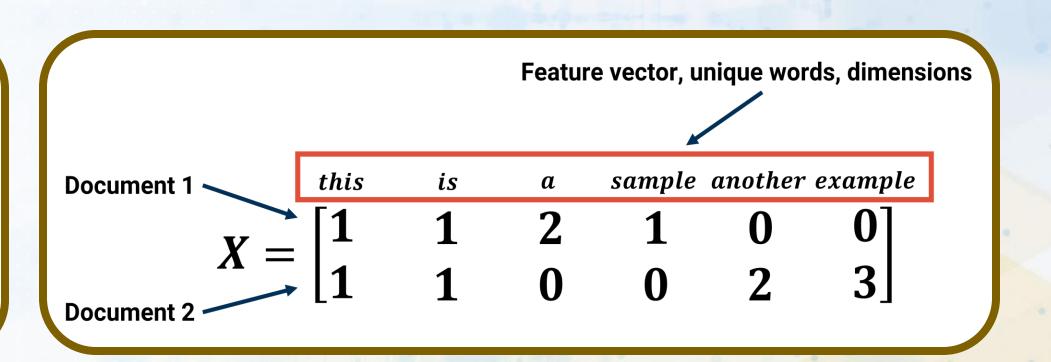
Document 2	
Term Count	
1	
1	
2	
3	





Doc	ument 1
Term	Term Count
this	1
is	1
а	2
sample	1

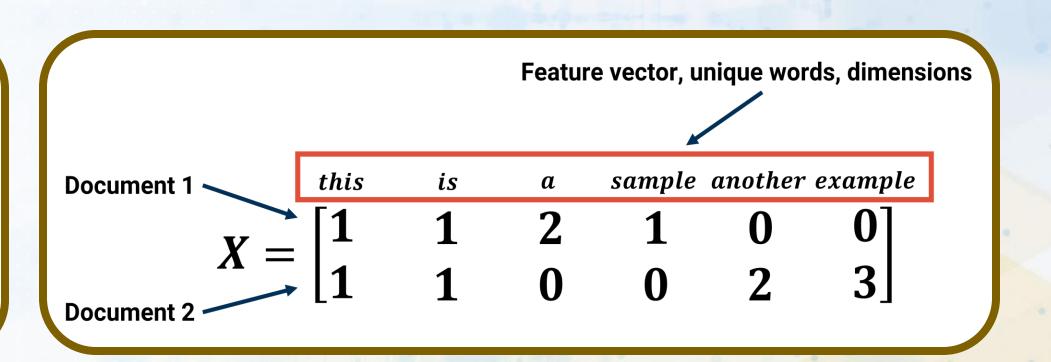
Document 2	
Term Count	
1	
1	
2	
3	





Doc	ument 1
Term	Term Count
this	1
is	1
а	2
sample	1

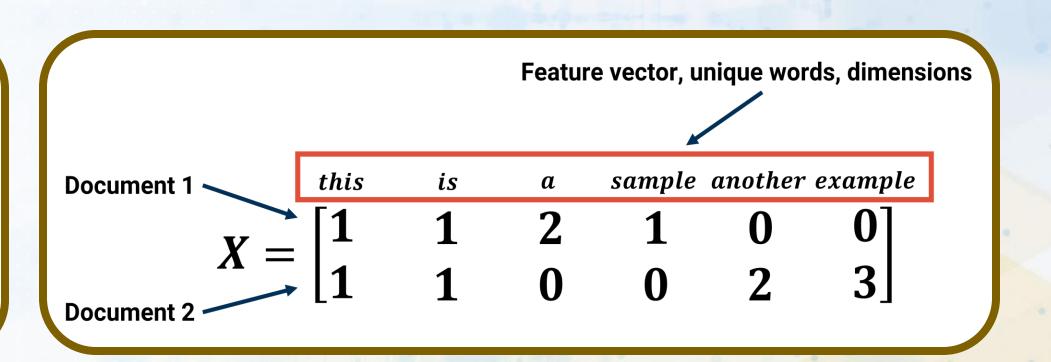
Document 2	
Term Count	
1	
1	
2	
3	





Doc	ument 1
Term	Term Count
this	1
is	1
а	2
sample	1

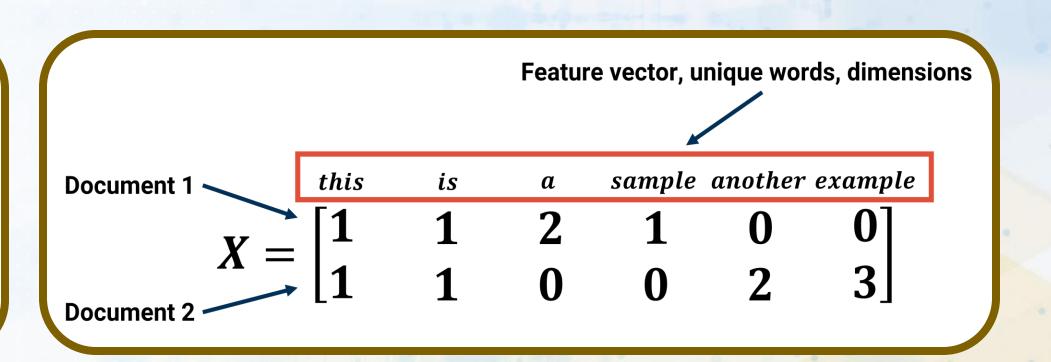
Document 2	
Term Count	
1	
1	
2	
3	





Doc	ument 1
Term	Term Count
this	1
is	1
а	2
sample	1

Document 2	
Term Count	
1	
1	
2	
3	





TF (Term Frequency)

The number of appearances of a term in a document. It will be high, if terms appear many times in this document

For example; we have the term (words) count tables of a corpus for just two documents:

Document 1

Term	Term Count
this	1
is	1
а	2
sample	1

1
1
2
3

$$ext{tf("this"}, d_1) = rac{1}{5} = 0.2 \ ext{tf("this"}, d_2) = rac{1}{7} pprox 0.14$$



TF (Term Frequency)

The number of appearances of a term in a document. It will be high, if terms appear many times in this document

For example; we have the term (words) count tables of a corpus for just two documents:

Document 1

Term	Term Count
this	1
is	1
а	2
sample	1

1
1
2
3

$$ext{tf("this"}, d_1) = rac{1}{5} = 0.2 \ ext{tf("this"}, d_2) = rac{1}{7} pprox 0.14$$



TF (Term Frequency)

The number of appearances of a term in a document. It will be high, if terms appear many times in this document

For example; we have the term (words) count tables of a corpus for just two documents:

Document 1

Term	Term Count
this	1
is	1
а	2
sample	1

1
1
2
3

$$ext{tf("this"}, d_1) = rac{1}{5} = 0.2 \ ext{tf("this"}, d_2) = rac{1}{7} pprox 0.14$$



IDF (Inverse Document Frequency)

$$IDF = \log(\frac{N}{The number of documents containing that term})$$

$$= \log(\frac{N}{The \ number \ of \ documents \ containing \ that \ \textbf{term}})$$

$$\operatorname{idf}("\mathsf{this}",D) = \log\!\left(rac{2}{2}
ight) = 0$$



IDF (Inverse Document Frequency)

$$IDF = \log(\frac{N}{The number of documents containing that term})$$

$$= \log(\frac{N}{The \ number \ of \ documents \ containing \ that \ \textbf{term}})$$

$$\operatorname{idf}("\mathsf{this}",D) = \log\!\left(rac{2}{2}
ight) = 0$$



TF-IDF

A word's importance score in a document, among N documents

Final score = TF * IDF (higher score → more "characteristic")

Document 1

Term	Term Count
this	1
is	1
а	2
sample	1

Term	Term Count
this	1
is	1
another	2
example	3

$$ext{tfidf}(" ext{this}", d_1, D) = 0.2 imes 0 = 0 \ ext{tfidf}(" ext{this}", d_2, D) = 0.14 imes 0 = 0 \$$



Advantages and Disadvantages of TF-IDF

- Advantages:
 - Simple and easy to implement
 - Higher score means "more characteristic". Common words will have very small scores such as "the", "a", "this",...
 - TF-IDF is a good technique to search for document, find similar documents, or cluster documents
- Disadvantages:
 - TF-IDF does NOT consider the position of the words because of creating the document-term matrix. Other methods such as Bag of Words also suffers from this issue



Summary

- Use TF-IDF as a better alterative for discrete text representation comparing to one-hot-encoding and bag of words
- Understand the advantages and disadvantages of this method

