Applied Text Analytics & Natural Language Processing

with Dr. Mahdi Roozbahani
& Wafa Louhichi

TF-IDF

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

BoW Example:

$$Vocabulary\ V = [this, is, a, sample, another, example]$$

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

Feature vector, unique words, dimensions

$$\begin{array}{l} \\ \text{Document 1} \\ \\ \text{Document 2} \end{array} X = \begin{array}{c} \begin{array}{cccccc} this & is & a & sample & another & example \end{array} \\ \begin{bmatrix} 1 & 1 & 2 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 2 & 3 \end{bmatrix} \end{array}$$

BoW Example:

$$Vocabulary\ V = [this, is, a, sample, another, example]$$

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

Feature vector, unique words, dimensions

$$\begin{array}{l} \\ \text{Document 1} \\ \\ \text{Document 2} \end{array} X = \begin{array}{c} \begin{array}{cccccc} this & is & a & sample & another & example \end{array} \\ \begin{bmatrix} 1 & 1 & 2 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 2 & 3 \end{bmatrix} \end{array}$$

BoW Example:

$$Vocabulary\ V = [this, is, a, sample, another, example]$$

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

Feature vector, unique words, dimensions

$$\begin{array}{l} \\ \text{Document 1} \\ \\ \text{Document 2} \end{array} X = \begin{array}{c} \begin{array}{cccccc} this & is & a & sample & another & example \end{array} \\ \begin{bmatrix} 1 & 1 & 2 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 2 & 3 \end{bmatrix} \end{array}$$

BoW Example:

$$Vocabulary\ V = [this, is, a, sample, another, example]$$

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

Feature vector, unique words, dimensions

$$\begin{array}{l} \\ \text{Document 1} \\ \\ \text{Document 2} \end{array} X = \begin{array}{c} \begin{array}{cccccc} this & is & a & sample & another & example \end{array} \\ \begin{bmatrix} 1 & 1 & 2 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 2 & 3 \end{bmatrix} \end{array}$$

BoW Example:

$$Vocabulary\ V = [this, is, a, sample, another, example]$$

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

Feature vector, unique words, dimensions

$$\begin{array}{l} \\ \text{Document 1} \\ \\ \text{Document 2} \end{array} X = \begin{array}{c} \begin{array}{cccccc} this & is & a & sample & another & example \end{array} \\ \begin{bmatrix} 1 & 1 & 2 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 2 & 3 \end{bmatrix} \end{array}$$

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
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

$$\mathrm{tf}(''\text{this}'', d_1) = \frac{1}{5} = 0.2$$

$$\mathrm{tf}(''\text{this}'', d_2) = \frac{1}{7} \approx 0.14$$

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
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

$$\mathrm{tf}(''\text{this}'', d_1) = \frac{1}{5} = 0.2$$

$$\mathrm{tf}(''\text{this}'', d_2) = \frac{1}{7} \approx 0.14$$

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
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

$$\mathrm{tf}(''\text{this}'', d_1) = \frac{1}{5} = 0.2$$

$$\mathrm{tf}(''\text{this}'', d_2) = \frac{1}{7} \approx 0.14$$

IDF (Inverse Document Frequency)

$$\mathrm{IDF} = \log\left(\frac{N}{\textit{The number of documents containing that } \boldsymbol{term}}\right)$$

$$= \log\left(\frac{N}{\textit{The number of documents containing that } \boldsymbol{term}}\right)$$

$$\mathrm{idf}(\text{"this"}, D) = \log\left(\frac{2}{2}\right) = 0$$

IDF (Inverse Document Frequency)

$$\mathrm{IDF} = \log\left(\frac{N}{\textit{The number of documents containing that } \boldsymbol{term}}\right)$$

$$= \log\left(\frac{N}{\textit{The number of documents containing that } \boldsymbol{term}}\right)$$

$$\mathrm{idf}(\text{"this"}, D) = \log\left(\frac{2}{2}\right) = 0$$

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
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

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

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