Bag of Words

Bag of Words

- A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things:
  - A vocabulary of known words.
  - A measure of the presence of known words.
- Any information about the order or structure of words in the document is discarded.
- The model is only concerned with whether known words occur in the document, not where in the document.

Example of the Bag-of-Words Model

- Step 1: Collect Data
- Below is a snippet of the first few lines of text from the book “A Tale of Two Cities” by Charles Dickens, taken from Project Gutenberg.

  It was the best of times,
  it was the worst of times,
  it was the age of wisdom,
  it was the age of foolishness,

Step 2: Design the Vocabulary

- Make a list of all of the words in our model vocabulary.
- The unique words here (ignoring case and punctuation) are:
  - “it”
  - “was”
  - “the”
  - “best”
  - “of”
  - “times”
  - “worst”
  - “age”
  - “wisdom”
  - “foolishness”

Step 3: Create Document Vectors

- The next step is to score the words in each document.
- The objective is to turn each document of free text into a vector that we can use as input or output for a machine learning model.
- Because we know the vocabulary has 10 words, we can use a fixed-length document representation of 10, with one position in the vector to score each word.
- The simplest scoring method is to mark the presence of words as a boolean value, 0 for absent, 1 for present.
- Example “It was the best of times“) and convert it into a binary vector.
- The scoring of the document would look as follows

Continued..

The scoring of the document would look as follows:

“it” = 1

“was” = 1

“the” = 1

“best” = 1

“of” = 1

“times” = 1

“worst” = 0

“age” = 0

“wisdom” = 0

“foolishness” = 0

As a binary vector, this would look as follows:

[1, 1, 1, 1, 1, 1, 0, 0, 0, 0]

Continued..

- The other three documents would look as follows:
- "it was the worst of times" = [1, 1, 1, 0, 1, 1, 1, 0, 0, 0]
- "it was the age of wisdom" = [1, 1, 1, 0, 1, 0, 0, 1, 1, 0]
- "it was the age of foolishness" = [1, 1, 1, 0, 1, 0, 0, 1, 0, 1]