# Co-Occurrence Matrix

A screenshot of a computer

AI-generated content may be incorrect.

We are provided with a Corpus which has multiple documents. The documents are in the form of a list (tokenized). As shown below each document has a word with indexes. In this case below, Document 1 and 2 both have 7 words with index from 0 to 1.

A white board with black text

AI-generated content may be incorrect.

First, we extract all unique words into a single vector and order them. Lets call it Corpus words. The logic to extract the words is simple:

1. Loop over the documents in the Corpus
   1. Loop every word in the document,
      1. Check if the word exists in the Corpus words list
      2. Add the word if it does not exist
   2. End Loop
2. End Loop

A close-up of a number

AI-generated content may be incorrect.

Create a word2Ind dictionary:

The “key” of the dictionary is the unique word and the “value” is the index.

A close-up of a sign

AI-generated content may be incorrect.

The co-occurrence Matrix is a square matrix which has both the rows and columns same as the width of the word dictionary.

Context Window = the words before of a fixed window size excluding that word.

1. Loop over the documents in the Corpus
2. Assign the **idx (index)** and **word** from the enumeration of the documents. Now there will be 2 lookups. One before and after:
   1. Loop through the 2 words **before** idx **(idx is used for the back and forth)**
      1. For the word, fetch the index from word2ind and assign to context index. *Before you can find the index, you have to get the word from document using idx.* ***(idx is used for fetching the context word)***
      2. Update the M[center Index][context Index] by 1.
   2. End Loop
   3. Loop through the 2 words **after** idx **(idx is used for the back and forth)**
      1. For the word, fetch the index from word2ind and assign to context index. *Before you can find the index, you have to get the word from document using idx.* ***(idx is used for fetching the context word)***
      2. Update the M[center Index][context Index] by 1.
3. End Loop

A whiteboard with colorful text

AI-generated content may be incorrect.

A graph on a white board

AI-generated content may be incorrect.

Make the co-occurrence dimensions smaller and yet have all the knowledge.

<https://davetang.org/file/Singular_Value_Decomposition_Tutorial.pdf>