Lesson 5.06
NLP PART 2

Vectorization Types

- **Count Vectorizer** creates a document-term matrix where the entry of each cell will be a count of the number of times that word occurred in that document.

- **TF-IDF (Term Frequency — Inverse Document Frequency)** is basically a count vectorizer that includes some consideration for the length of the document, and also how common the word is across other text messages.

- **N-Grams** is used to look for groups of adjacent words instead of just looking for single terms.

- They're all just slight modifications of each other, and typically you'll test different vectorization methods depending on your problem, and the results determine which one you will proceed with.

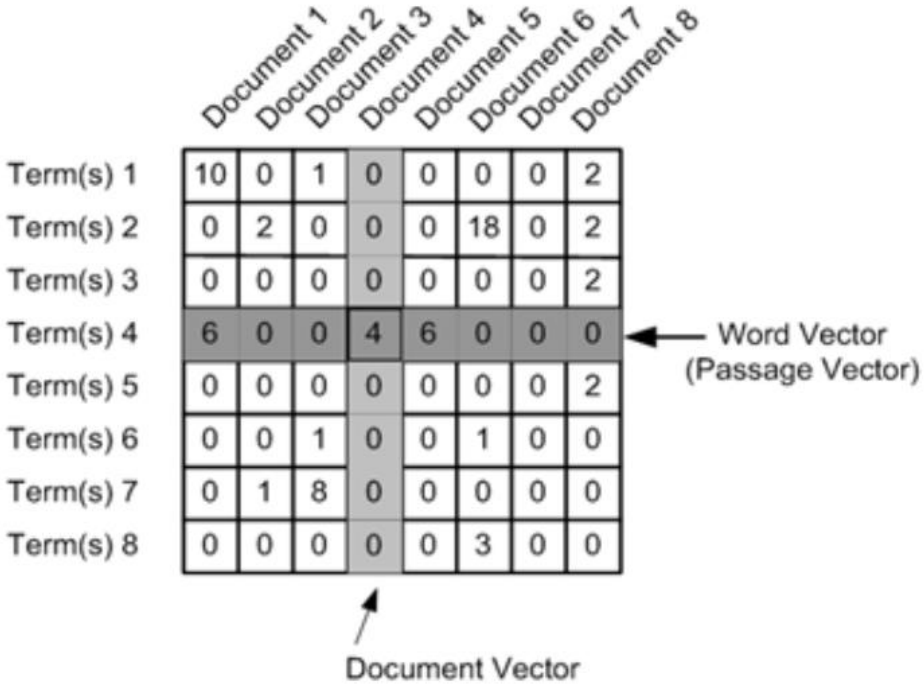
TF-IDF Formula

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents



Document Term Matrix

- A document-term matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents.

- In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms.

Sparse Matrix

- A matrix comprising of mostly zero values
- Only locations and values non-zero elements are stored to save space

```
1  [[1 0 0 1 0 0]
2   [0 0 2 0 0 1]
3   [0 0 0 2 0 0]]
4
5     (0, 0)    1
6     (0, 3)    1
7     (1, 2)    2
8     (1, 5)    1
9     (2, 3)    2
10
```

Normal Matrix
(All values)

Sparse Matrix
(Location and Values)