Natural Language Processing

A **corpus** is a large collection of text. Before we can apply data mining methods to a corpus, we need to first extract features (attributes) from the corpus that we can then use to data mine the corpus. Text feature extraction involves some or all of the following steps:

removing punctuation Remove punctuation and convert all letters to lower case.
tokenization Map individual words (terms) in the corpus to distinct numbers.
establishing the vocabulary Identify the set of unique words in the corpus.
filtering words Remove **stop words**, i.e. words like *a* and *the* that are too common to be useful.
vectorization Map individual documents in the corpus to vectors.

The basic approach that we will use to vectorize a document in the corpus is known as the **bag-of-words** approach. Each document in the corpus is treated as a bag of words. The order of the words in a document is important, but for simplicity, word order is ignored. Treating a corpus as a bags of words is used to create a **document term matrix**.

45 Example (Document Term Matrix)
Consider a corpus consisting of three documents. The words *goal, bill, stocks* and *intense* are in the vocabulary of the corpus. The table below lists the number of times each word in the vocabulary appears in each document. Observe that the words are listed in alphabetic order.

Document	bills	goals	intense	stocks
article 1	3	1	1	0
article 2	2	2	1	3
article 3	0	2	1	0

A **document term frequence matrix** is a document term matrix normalized by the number of words in each document.

46 Example (Document Term Frequency Matrix)
Divide each row of the term matrix by its row sum to get the term frequency matrix:

Document	bills	goals	intense	stocks
article 1	3/5	1/5	1/5	0
article 2	2/8	2/8	1/8	3/8
article 3	0	2/3	1/3	0

tf-idf vectorization is widely used in data mining of text. **tf** is the term frequency matrix of a collection of documents in a corpus. **idf** stands for Inverse Document Frequency. **idf** is used as a measure of the importance of a word in a vocabulary. **tf-idf** is the *element-wise product* of the **tf** and **idf** matrices.

47 Definition (Inverse Document Frequency (idf))
Sklearn defines the idf of a term (word) to equal:

$$\text{idf(term)} = \overset{\text{ln}}{\log}\left(\frac{\text{total number of documents} + 1}{\text{number of documents containing the term} + 1}\right) + 1$$

When a large matrix has only a few non-zero values, it is more efficient to convert the matrix into a sparse matrix.

48 Example (Dense vs Sparse Matrix)

Dense Matrix

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 4 & 0 & 0 & 7 \end{pmatrix}$$

Sparse Matrix

(1,2)	1
(2,1)	4
(2,4)	7