Introduction

An overview of natural language processing and word embeddings.

In this section, you will learn about using word embeddings to give numeric vector representations of words. Word embeddings are an extremely important part of natural language processing, since they allow us to use text with deep learning models.

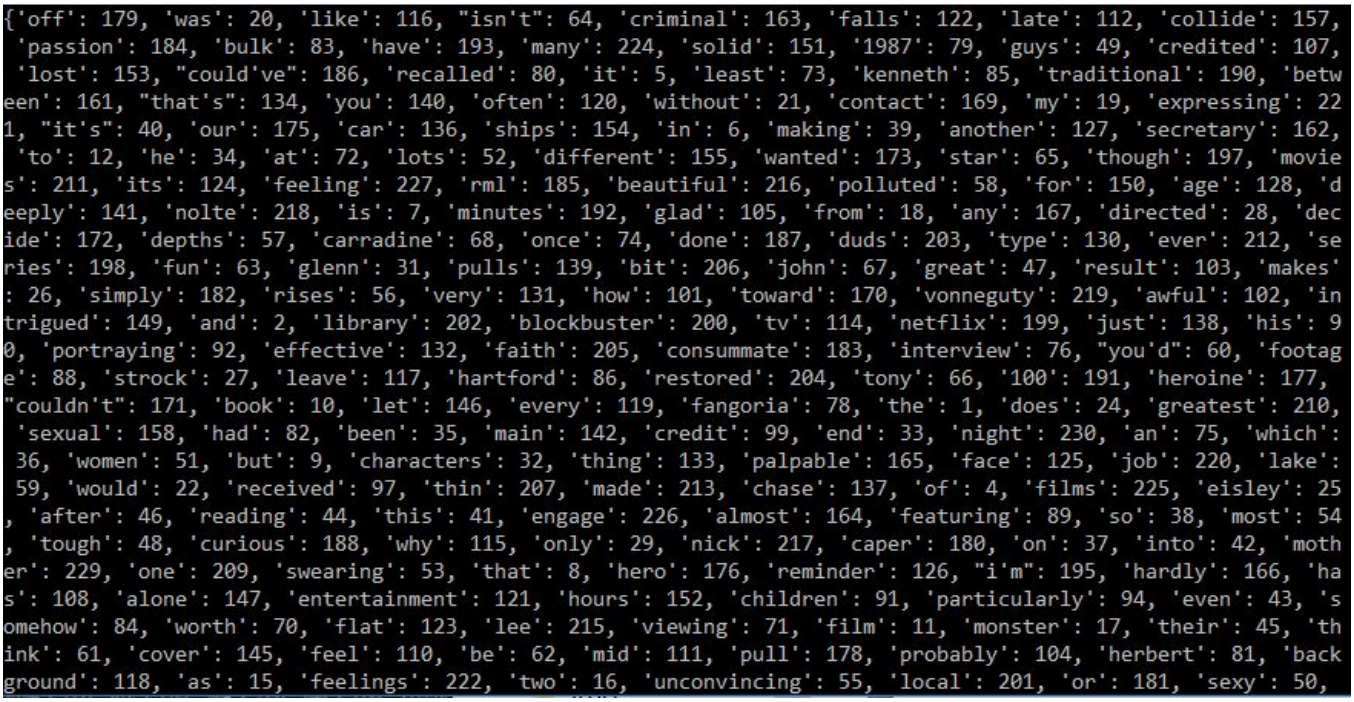
A. What is NLP?

Natural language processing (NLP) encompasses any task related to machines dealing with natural language, i.e. human spoken or written language. Hence, it is one of the most important fields for machine learning and artificial intelligence. Tasks such as translating between languages, speech recognition, text analysis, and automatic text generation all fall under the scope of NLP. Without NLP we wouldn't have voice assistants like Siri and Alexa, or even search engines such as Google and Bing.

While machine learning for NLP is the focus of this course, not all NLP tasks require machine learning. For example, search engines rely largely on non-ML algorithms to find the most relevant documents or web pages for a given search query. Nevertheless, machine learning is becoming more and more widespread in NLP, and this trend will likely continue for many years to come.

B. Using text data

Natural language deals with two main categories of data: spoken or written data. While spoken language data is heavily used when building conversational agents like Siri and Alexa, written data (i.e. text data) is much more prevalent in industry NLP tasks. Raw text data is almost always unusable in NLP applications. The data is basically just a mass of strings without any real meaning for a machine to process. It is up to the engineer to first convert the raw text data into usable machine data, which can then be used as input for NLP algorithms.



Example of processed text data.

In the following chapters you'll be introduced to an easy and efficient way to process raw text, and then you'll use the processed text to run a machine learning algorithm.