**What is NLP?**

Natural Language Processing is a field of Artificial Intelligence that gives the machine the ability to read, understand and derive meaning from Human Languages. It is a discipline that focus on the interaction between Data Science and human language and scaling to lots of industries (Healthcare, media, finance, human resources etc.).

**Use Cases of NLP**

* NLP enables the recognition and **prediction of diseases**based on electronic health records and patient’s own speech. This capability is being explored in health conditions that go from cardiovascular diseases to depression and even schizophrenia. For example, Amazon Comprehend Medical is a service that uses NLP to extract disease conditions, medications and treatment outcomes from patient notes, clinical trial reports and other electronic health records.
* Organizations can determine what customers are saying about a service or product by identifying and extracting information in sources like social media. This **Sentiment Analysis**can provide a lot of information about customers choices and their decision drivers.
* Companies like Yahoo and Google filter and classify your emails with NLP by analysing text in emails that flow through their servers and **stopping spam** before they even enter your inbox.
* To help **identifying fake news**, the NLP Group at MIT developed a new system to determine if a source is accurate or politically biased, detecting if a news source can be trusted or not.
* Amazon’s Alexa and Apple’s Siri are examples of intelligent **voice driven interfaces** that use NLP to respond to vocal prompts and do everything like find a particular shop, tell us the weather forecast, suggest the best route to the office or turn on the lights at home.
* NLP is also being used in both the search and selection phases of **talent recruitment,** identifying the skills of potential hires and spotting prospects before they become active on the job market.
* Just like above mentioned scenarios, NLP is vastly being used in many sectors and some industries are yet to implement NLP.

NLP algorithms are machine learning algorithms based. NLP learns by analysing a set of examples (i.e. a large corpus, like a book, down to a collection of sentences), and making a statistical inference, instead of coding large sets of rules. We can organize the massive chunks of text data and solve a wide range of problems such as — automatic summarization, machine translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation etc.

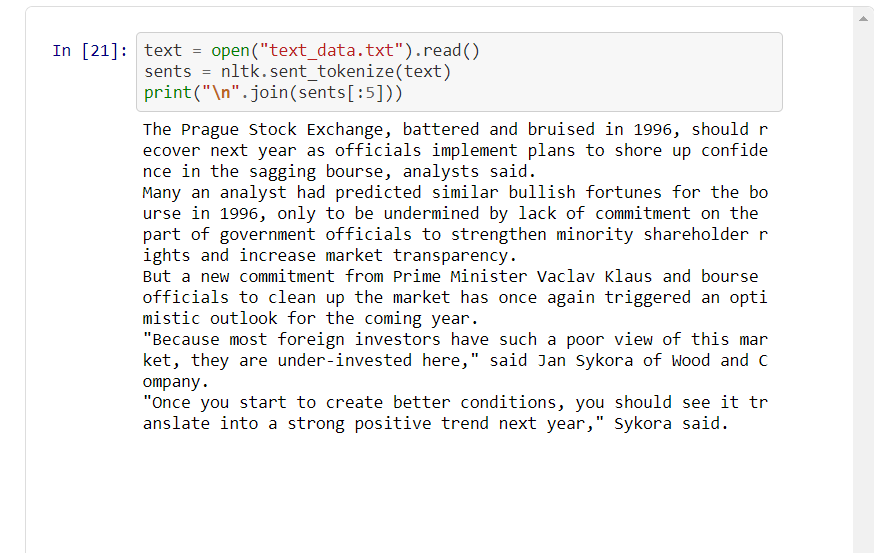
**Let’s dive deeper**

The text is the most unstructured form of all the available data. It is important to cleaning and standardize this text and make it noise free. The idea is to take the raw text and turn into something which can be utilized by an ML algorithm to carry out prediction. We will talk about few important techniques using NLTK.



**Sentence Segmentation**

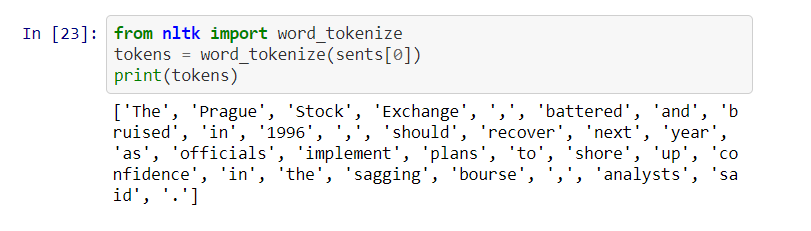
We break the articles into sentences. Often, we must do analysis at sentences level. For example, we want to check the number of sentences in an article and number of words in a sentence.



## Tokenization

Tokenization breaks unstructured data, text, into chunks of information which can be counted as discrete elements. This immediately turns an unstructured string (text document) into a more usable data, which can be further structured, and made more suitable for machine learning. Here we take the first sentence and we get each word as token. Below are two different ways i.e. RegexpTokenizer & Word Tokenize.

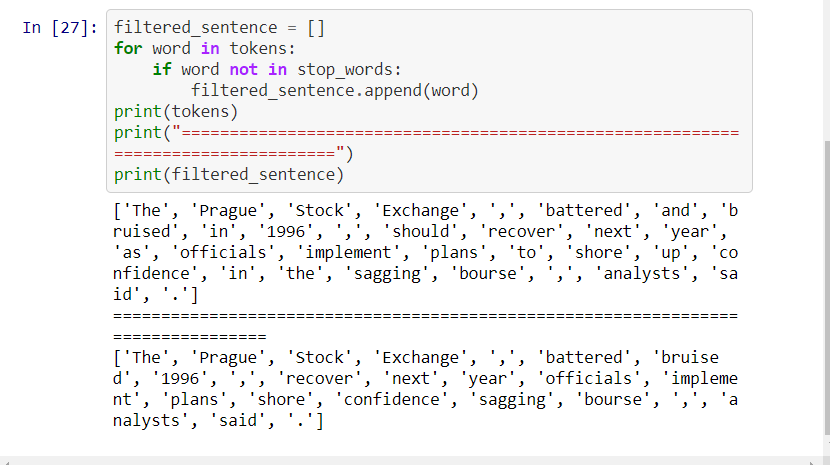




## Stop Words

Consider words like a, an, the, be etc. These words don’t add any extra information in a sentence. Such words can often create noise while modelling. Such words are known as Stop Words. We filter each sentence by removing the stop words as shown below:





## Stemming and Lemmatization

Some words represent the same meaning. For example, Copy, copied, copying. The model might treat them differently, so we tend to strip such words to their core. We can do that by stemming or lemmatization. Stemming and Lemmatization are the basic text processing methods for English text.

## Stemming

It helps to create groups of words which have similar meanings and works based on a set of rules, such as remove “**ing**” if words are ending with “**ing**”. Different types of stemmers in NLTK are PorterStemmer, LancasterStemmer, SnowballStemmer.

## Lemmatization

Ituses a knowledgebase called WordNet. Because of knowledge, lemmatization can even convert words which are different and can’t be solved by stemmers, for example converting “**came**” to “**come**”.

