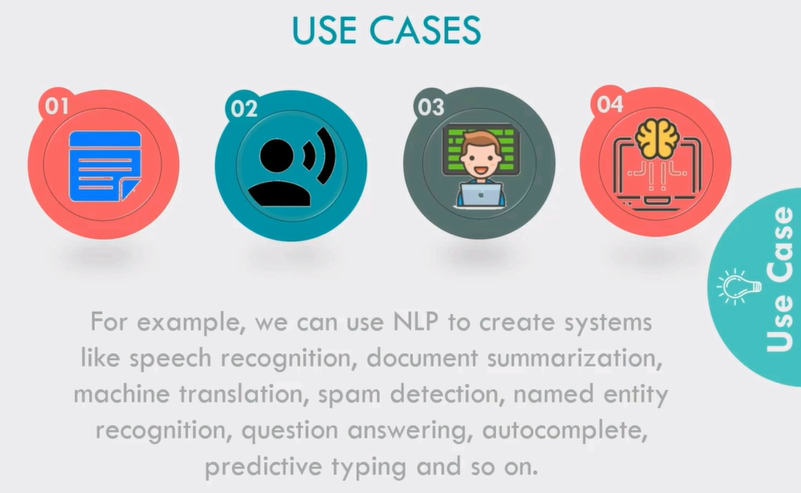
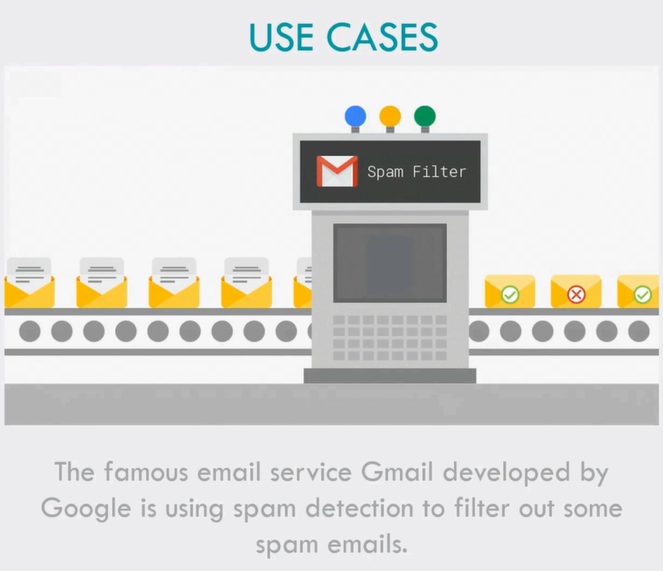
**NLP (Natural Language Processing):**

Is the subfield of Computer Science and Artificial Intelligence concerned with interactions between computers and humans(natural) languages. It is used to apply machine learning algorithms to text and speech.

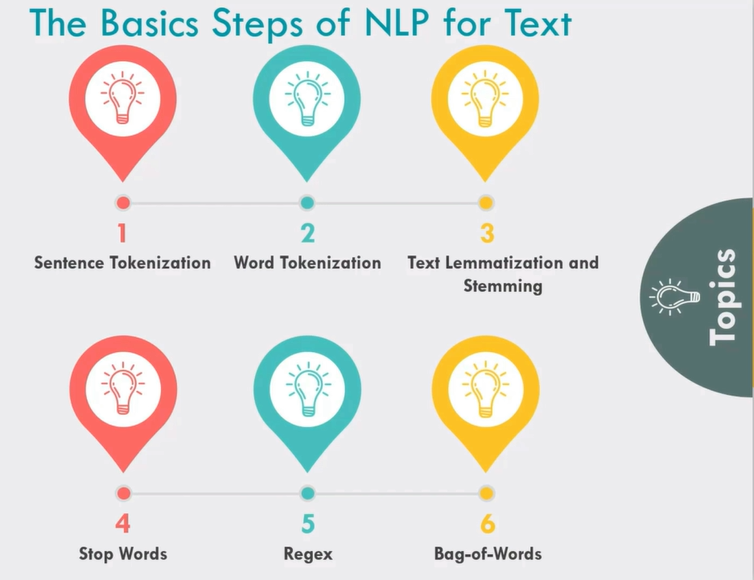




**NLTK (Natural Language Toolkit):**

Is the leading platform for building Python Programs to work with human language data. It provides easy to use interfaces to many corpora and lexical resources. Also it contains a suit of text processing libraries for classification, tokenization, stemming, tagging, parsing and semantic reasoning.

**Basics Steps of NLP For Text:**



**Sentence and word Tokenization:** process of tokenizing or splitting a string, text into a list of tokens. One can think of token as parts like a word is a token in a sentence, and a sentence is a token in a paragraph.

**Text Lemmatization:** Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.

**Stemming:**  Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes.

**Stop Words:** These are actually the most common words in any language (like articles, prepositions, pronouns, conjunctions, etc.) and does not add much information to the text. Examples of a few stop words in English are “the”, “a”, “an”, “so”, “what”.

**Bag-of-Words:** A bag of words is a representation of text that describes the occurrence of words within a document. We just keep track of word counts and disregard the grammatical details and the word order. It is called a “bag” of words because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document.