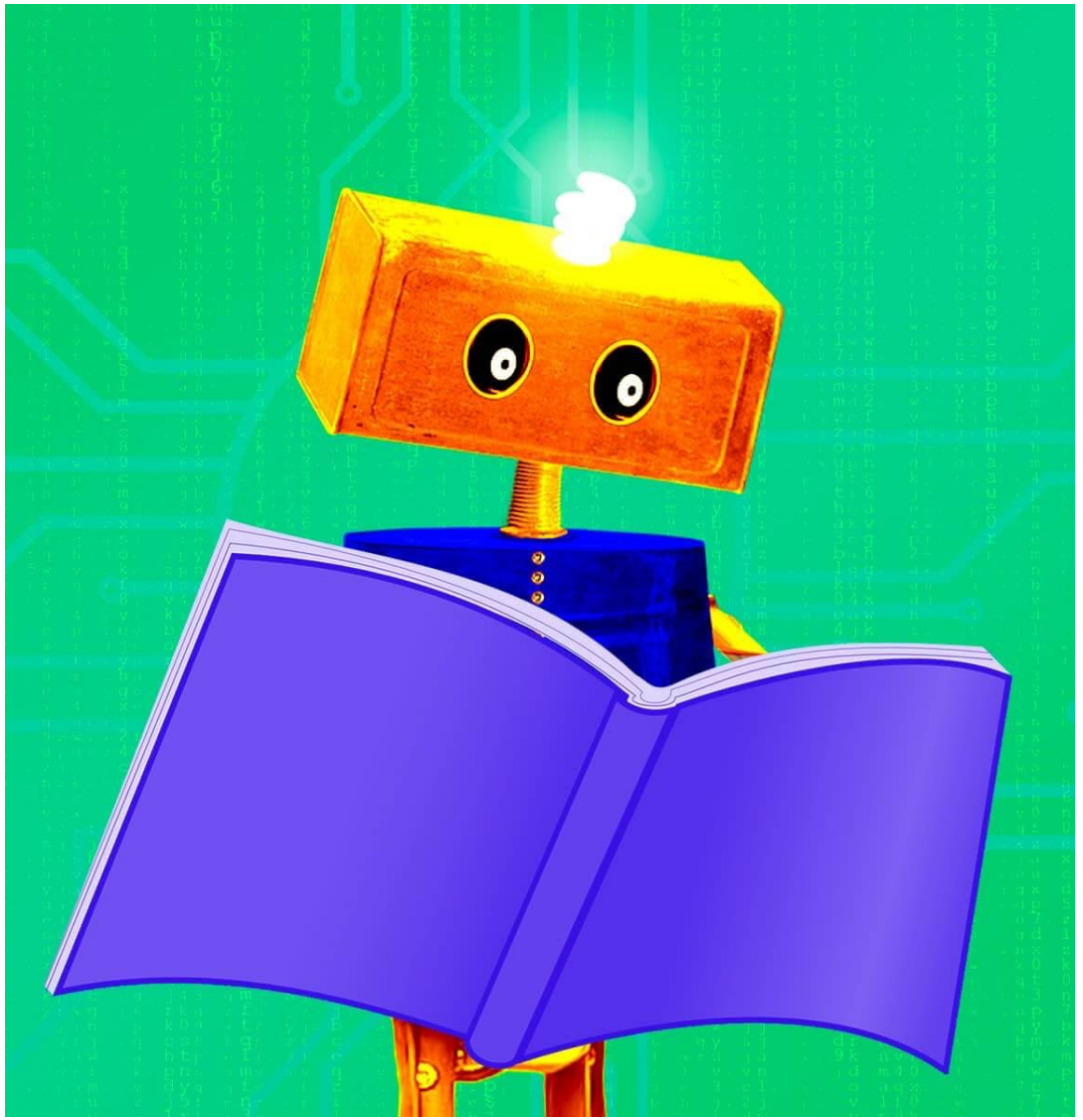
Text Analytics Workshop Part II:

Preprocessing and Tokenization

Author: Liliya Akhtyamova, PhD

How do computers understand words?

- Through numbers (and then bits)
- Need to encode words (or parts of words, sentences, whole documents) into numbers. That process is called text vectorization, and numerical representations of texts are called embeddings
- Most popular types of embeddings are word embeddings



Funny video – how automated systems can struggle to understand human language:
Scottish Elevator - Voice Recognition - ELEVEN !

Text Vectorization Techniques

- Bag Of Words (Count Vectorizer)
- Term Frequency and Inverse Document Frequency (TF-IDF)
- Word2Vec
- Transformers

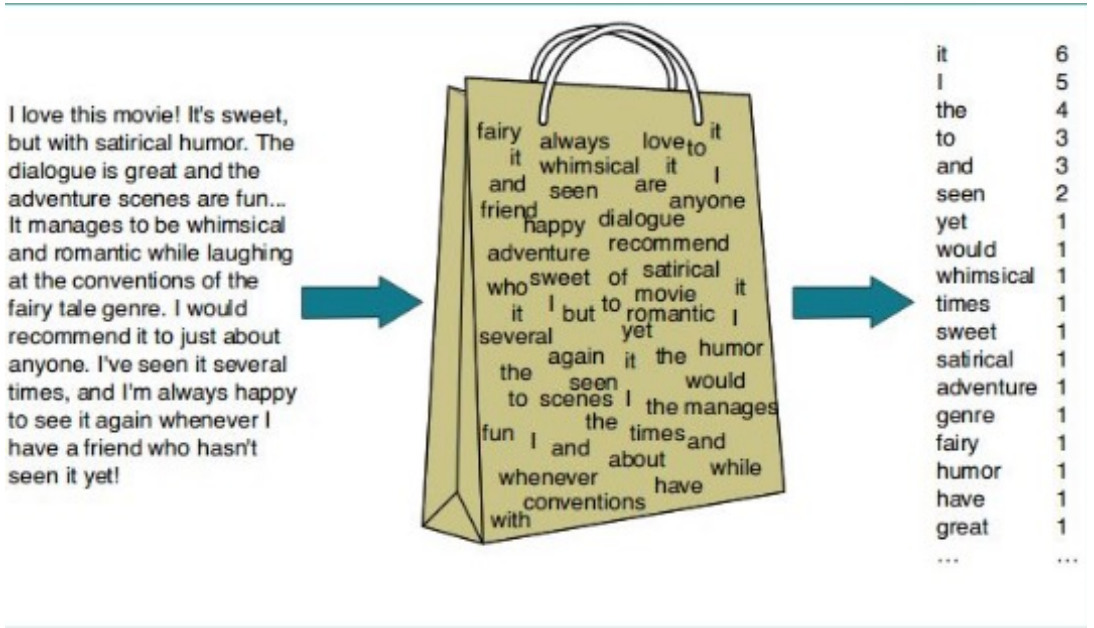
Bag of Words

- Bag of Words (BoW) is a model that is represented as an unordered set of words included in the processed text.

- Simple, but

"I love dogs, hate cats" == "I love cats, hate dogs" for BoW

- Still, may suffice for the global context: movie sentiment analysis, restaurant feedback, etc as the details of feedback message are less important



TF-IDF (Term Frequency – Inverse Document Frequency)

TF-IDF (Term Frequency - Inverse Document Frequency) is an algorithm that uses the frequency of words to determine how relevant those words are to a given document.

The weight of a word is proportional to the frequency of occurrence of this word in the document and inversely proportional to the frequency of occurrence of the word in all documents in the collection.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

Preprocessing

Problems:

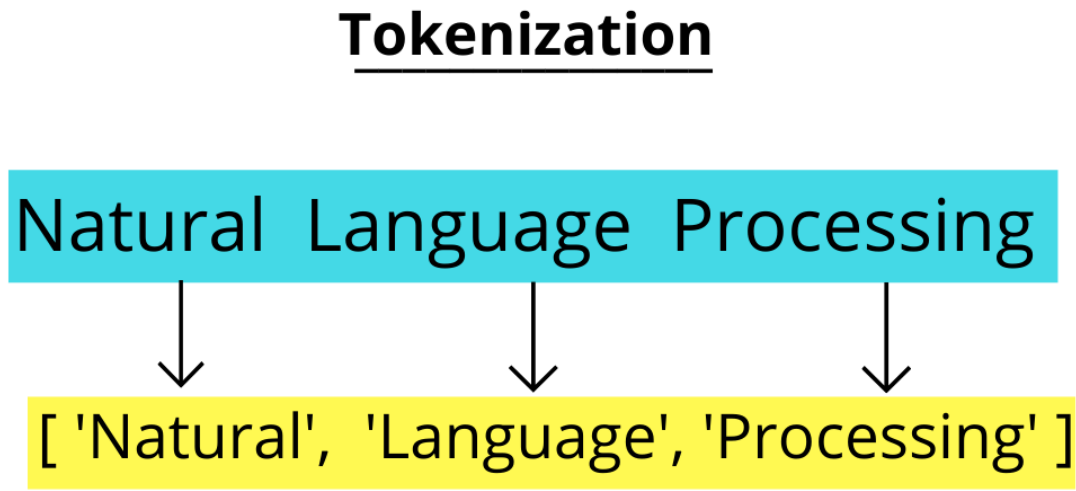
- long vectors (word embeddings) => expensive computation
- Same words with little difference in writing are considered different, i.e. “cup” vs “cups”, “finalize” vs “finalise”

Types of text preprocessing

- Tokenization,
- Noise removal (stop words, lowercasing, punctuation, etc)
- Stemming,
- Lemmatization.

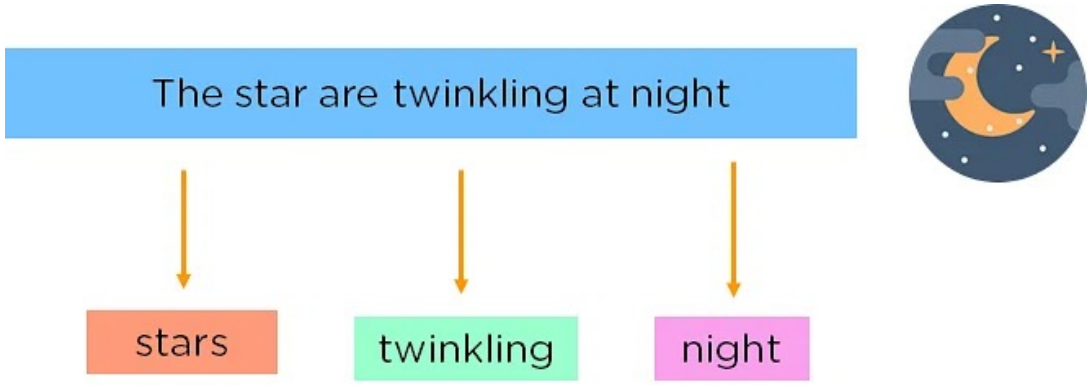
Most important – tokenization.

It’s the process of breaking a stream of textual data into words, terms, sentences, symbols, or some other meaningful elements called tokens



Stop words

- Stop words are words that are thrown out of the text during text processing. When we apply machine learning to texts, such words can add a lot of noise, so it is necessary to get rid of irrelevant words.
- Stop words are usually articles, interjections, conjunctions, etc., which do not carry a semantic meaning.
- At the same time, one must understand that there is no universal list of stop words, everything depends on the specific case.



Word normalization

- Stemming is the process of finding the stem of a word for a given source word. The stem of the word does not necessarily coincide with the morphological root of the word and does not have to be an existing word in the language. Stemming is a crude heuristic process that cuts off "excess" from the root of words, often resulting in the loss of derivational suffixes.

- Lemmatization brings all occurring word forms to one, normal dictionary form. Lemmatization uses vocabulary and morphological analysis to eventually bring the word to its canonical form, the lemma.

Thank you!