Text Mining

Claire Tayco

Project SPARTA Subject Matter Expert

Head of Research and Analytics, CirroLytix

What is Text Mining?

- Automatic process of extracting valuable insights from unstructured text
- Text mining techniques:
  - ➢Information Extraction – identifying key words, phrases, and relationships within text. Includes tokenization, identification of named entities, sentence segmentation, and part-of-speech tagging
  - ➢Concept extraction - process of searching documents or unstructured text for ideas and topics (e.g. topic modeling, sentiment analysis)
  - ➢Categorization – assignment of texts to predefined classes based on their content (e.g. spam detection)
  - ➢Clustering – finding groups of documents with similar contents

Sentiment Analysis

- Involves taking a piece of text (e.g. sentence, a comment, or an entire document) and returning a “score” that measures how positive or negative the text is.
- Applications:
  - ➢Review-related websites
  - ➢Antagonistic, heated language detection in mails
  - ➢Business and Government Intelligence (e.g. knowing consumer attitudes on products and services, knowing public opinion for political leaders)

Sentiment Analysis

- There are broadly two categories of sentiment analysis :

1. **Lexical Methods**: These techniques employ dictionaries of words annotated with their semantic polarity and sentiment strength. This is then used to calculate a score for the polarity and/or sentiment of the document.
2. **Machine Learning Methods**: Such techniques require creating a model by training the classifier with labeled examples. This means that you must first gather a dataset with examples for positive, negative and neutral classes, extract the features from the examples and then train the algorithm based on the examples.

VADER for Sentiment Analysis

- **VADER (Valence Aware Dictionary and sEntiment Reasoner)** is a lexicon and rule-based sentiment analysis tool
- It has been found to be quite successful when dealing with social media texts, NY Times editorials, movie reviews, and product reviews.
- It gives 4 scores:
  - ➢The Positive, Negative and Neutral scores represent the proportion of text that falls in these categories. All these add up to 1.
  - ➢The Compound score is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1 (most extreme negative) and +1 (most extreme positive).

Demonstration: Hotel Reviews Dataset

https://github.com/fstayco/shopee-code-league-2021

Topic Modeling

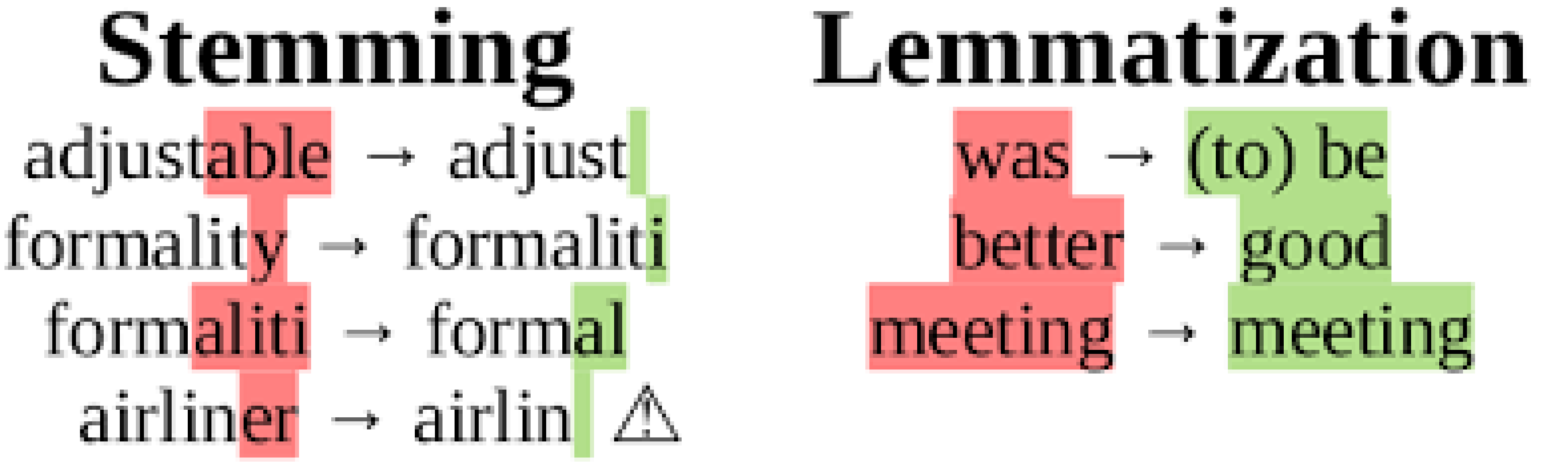
- A way to analyze large volumes of unlabeled text to discover topics or themes based on repeating patterns of co-occurrences of words in a set of documents (e.g. mentions).
- Topics – cluster of words that frequently occur together in documents

Topic Modeling: Text Preprocessing

- **Regular Expression/Normalization** — lowercase the words, remove punctuation and remove numbers
- **Tokenization** — a process of splitting the text into smaller pieces called tokens
- **Stop Words Removal** — a set of commonly used words in any language
- **Lemmatization** — a process of grouping together the inflected forms of a word so they can be analyzed as a single item
- **Stemming** — a process of grouping together the inflected forms of a word so they can be analyzed as a single item

Topic Modeling: Text Preprocessing

- **Lemmatization vs. Stemming:**



Stemming

adjustable → adjust

formality → formaliti

formaliti → formal

airliner → airlin ⚠

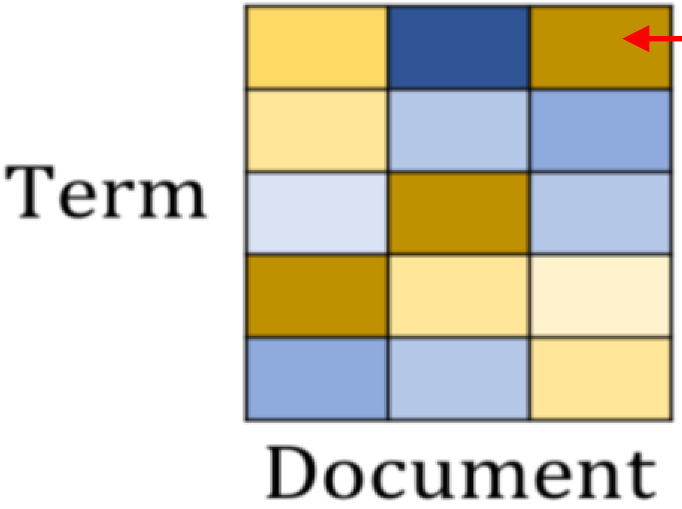
Lemmatization

was → (to) be

better → good

meeting → meeting

Topic Modeling: Feature Creation



$$TF-IDF_{t,d}=\underbrace{\frac{\text{Count of the occurences of the token t in document d}}{\text{Total number of tokens in document d}}}_{\boxed{\text{The more frequent the token appeared in a document, the higher the value.}}}\times\log\underbrace{\left(\frac{\text{Total number of documents in a corpus}}{\text{Number of documents where the token t appears}}\right)}_{\boxed{\text{The more the document appeared in the corpus, the lower the value}}}$$

TF-IDF measures the importance of the token, to a particular document and to all document. The higher the TF-IDF score of a token, the rarer the token.