module3 · text analytics

introduction to 🅃 text analytics

Introduction

Where text data sources from:

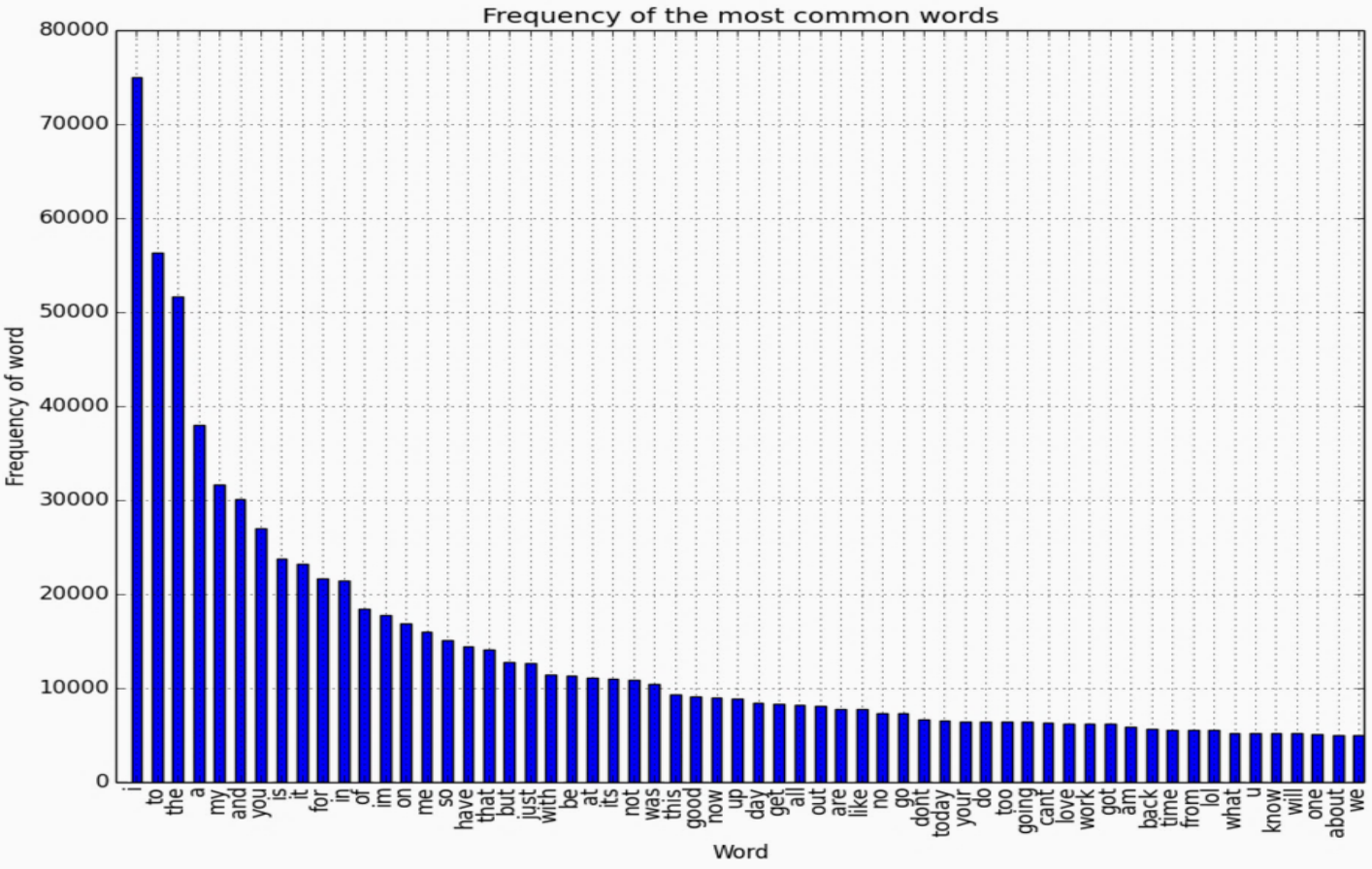
- Almost everywhere on the internet: webpages, social media pages, tweets, messaging, email, product reviews
- News articles, magazine articles
- Books, reference manuals
- Product descriptions, maintenance logs, customer service logs

What can be done with text data:

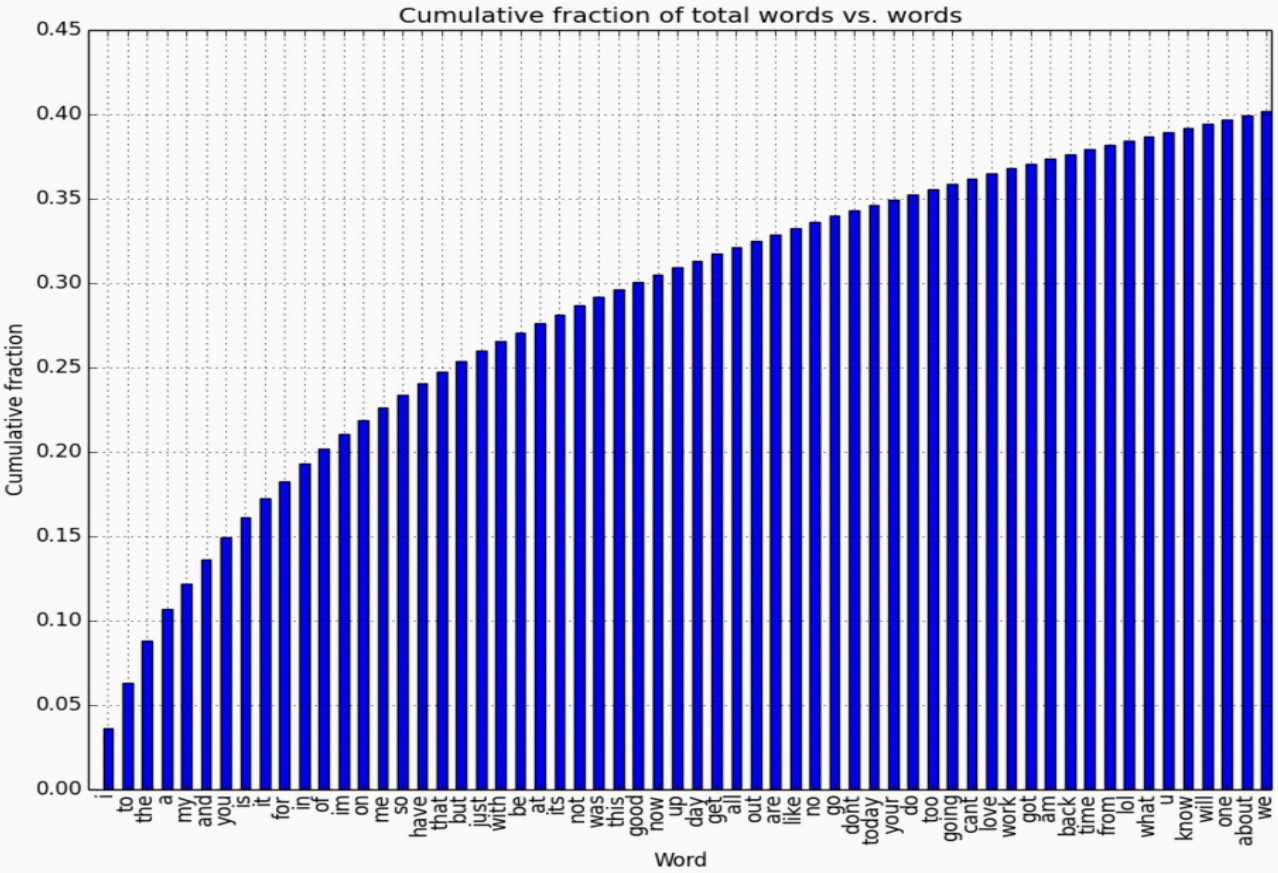
- Summarize it, or represent it in a useful way
- Classify it
- Search for things within it: information retrieval

Word Frequency

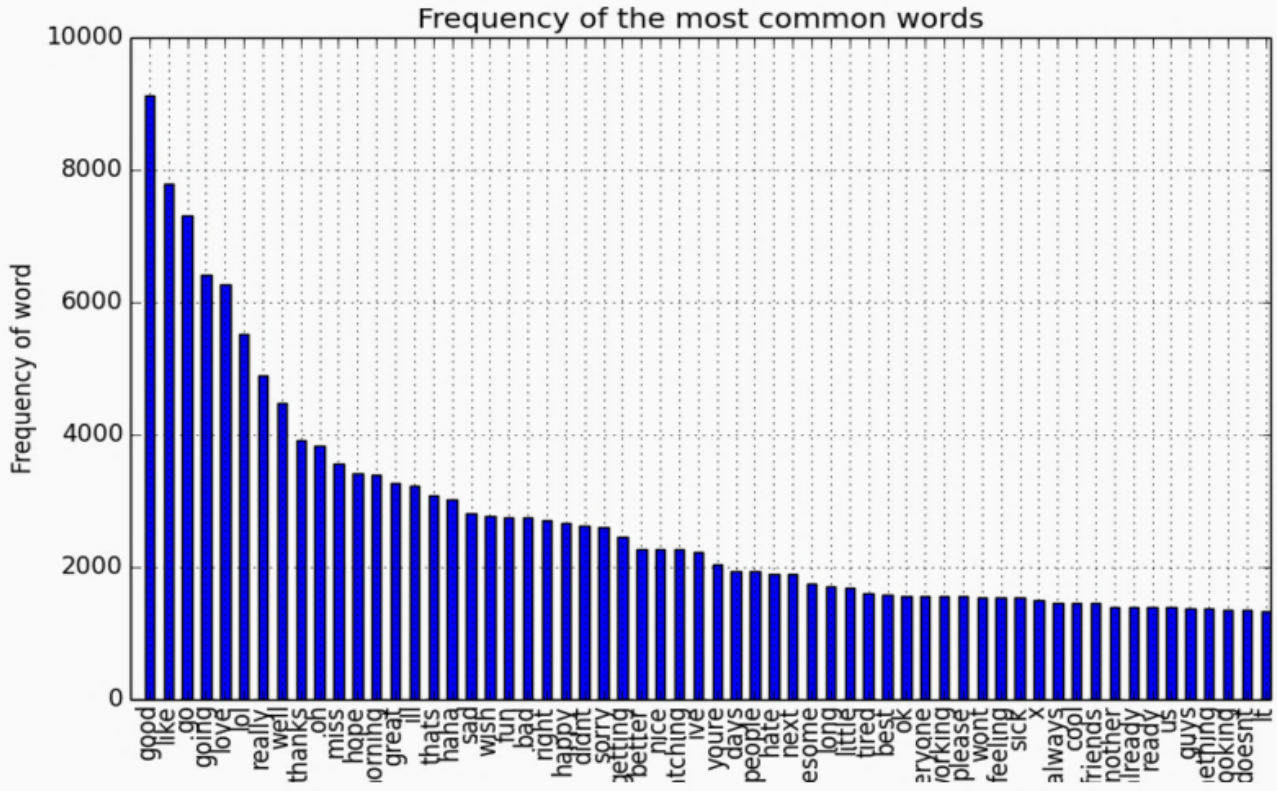
Pareto charts are simple bar charts that have ordered bins amongst the categories:



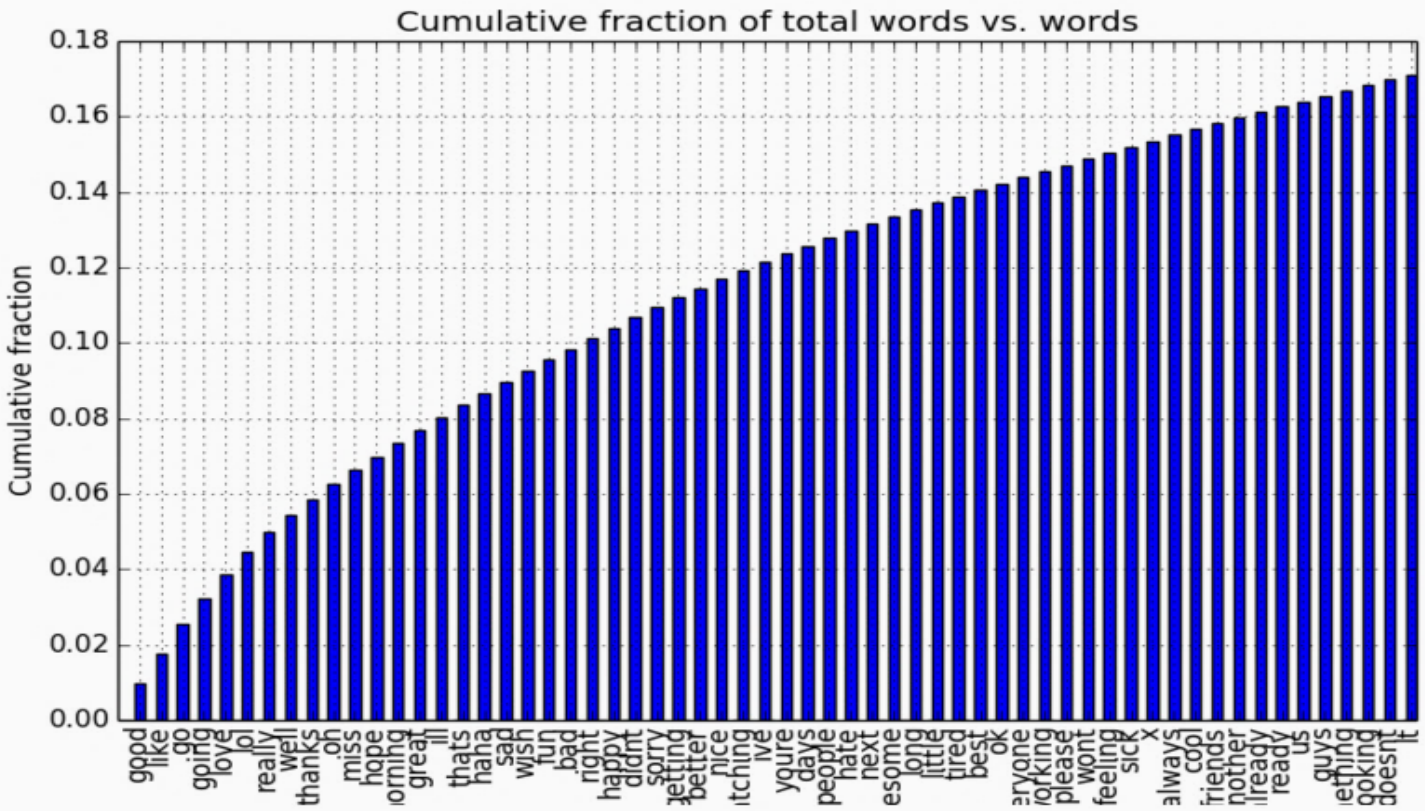
The Pareto chart above illustrates the frequency of most common used works from a Twitter Feed.



A cumulative plot of word frequency demonstrates how common words sum up the total amount of words used. For example, “I” represents ~4% with “I” and “The” representing ~9% cumulatively.



After removing the **stop words** (*the*, *is*, *at*, *which*, that, and *on*). **Stop words** are those that are not very useful for most **natural language processing applications**.



A cumulative Pareto Chart of word frequency after the **stop words** have been removed should words like *good*, *like, love, thanks, and wish* make up the cumulative total of all sampled Twitter words.

When obtaining raw text data, there are certain preprocessing steps to prepare raw data for input: