Text Mining in Search Engines

By: DJ Ambler
With special thanks to the Internet

Overview

- What is text mining?
- How is it used in search engines?

Text Mining Definition

- A way to extract meaning from text
- Structuring, deriving patterns, then evaluating
- “High quality” in text mining

Text Mining Tasks

- Text categorization
- Text clustering
- Concept/entity extraction
- Production of granular taxonomies
- Sentiment analysis
- Document summarization
- Entity relation modeling

Parts of a Search Engine

- Crawler
- Indexer
- Ranker

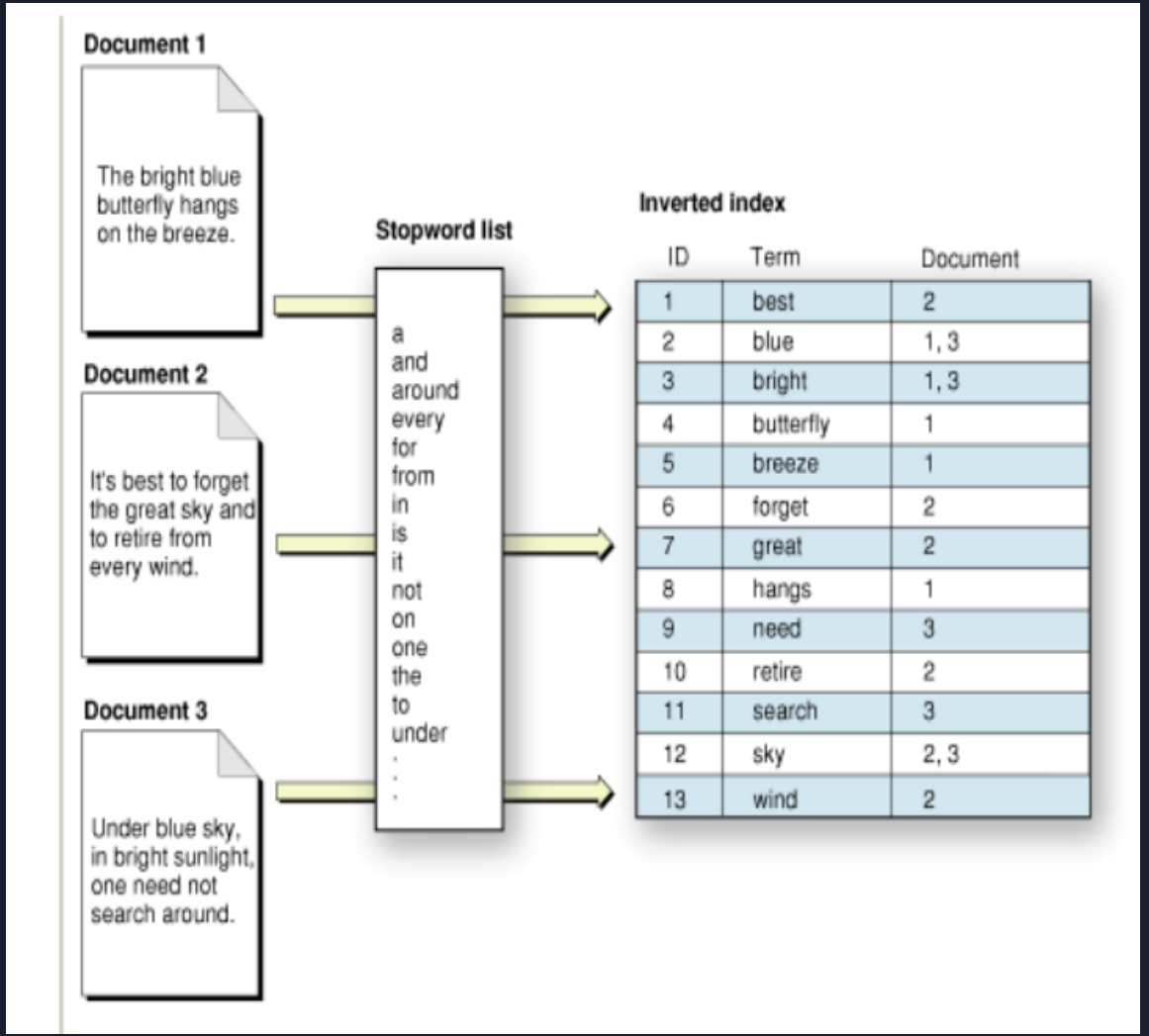
Crawler (Spider)

Issues in crawling:

1. What to crawl?
2. How much to crawl?
3. How often to crawl?

Indexer

- Stop words
- Stemming
- Issues



Ranker

- Receives query
- Searches index
- Ranks the pages based on complex algorithms

Ranking Criteria

- Number of matching query words in the page
- Proximity of matching words to one another
- Location of terms within the page
- Location of terms within tags e.g. <title>, <h1>, link text, body text, etc...
- Frequency of terms on the page and in general
- How “fresh” is the page

Sources

- Cong, G. (n.d.). Introduction to Text Mining and Web Search. Retrieved November 3, 2017.
- Joshi, H. (n.d.). Search Engines - Text Mining in Action. Retrieved November 03, 2017, from https://www.scribd.com/document/176948623/Search-Engines-Text-Mining-in-Action