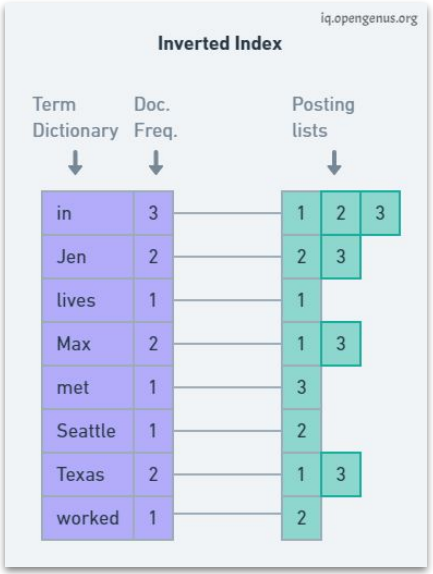
30. 倒排索引 (Inverted Index)



倒排索引(Inverted Index)

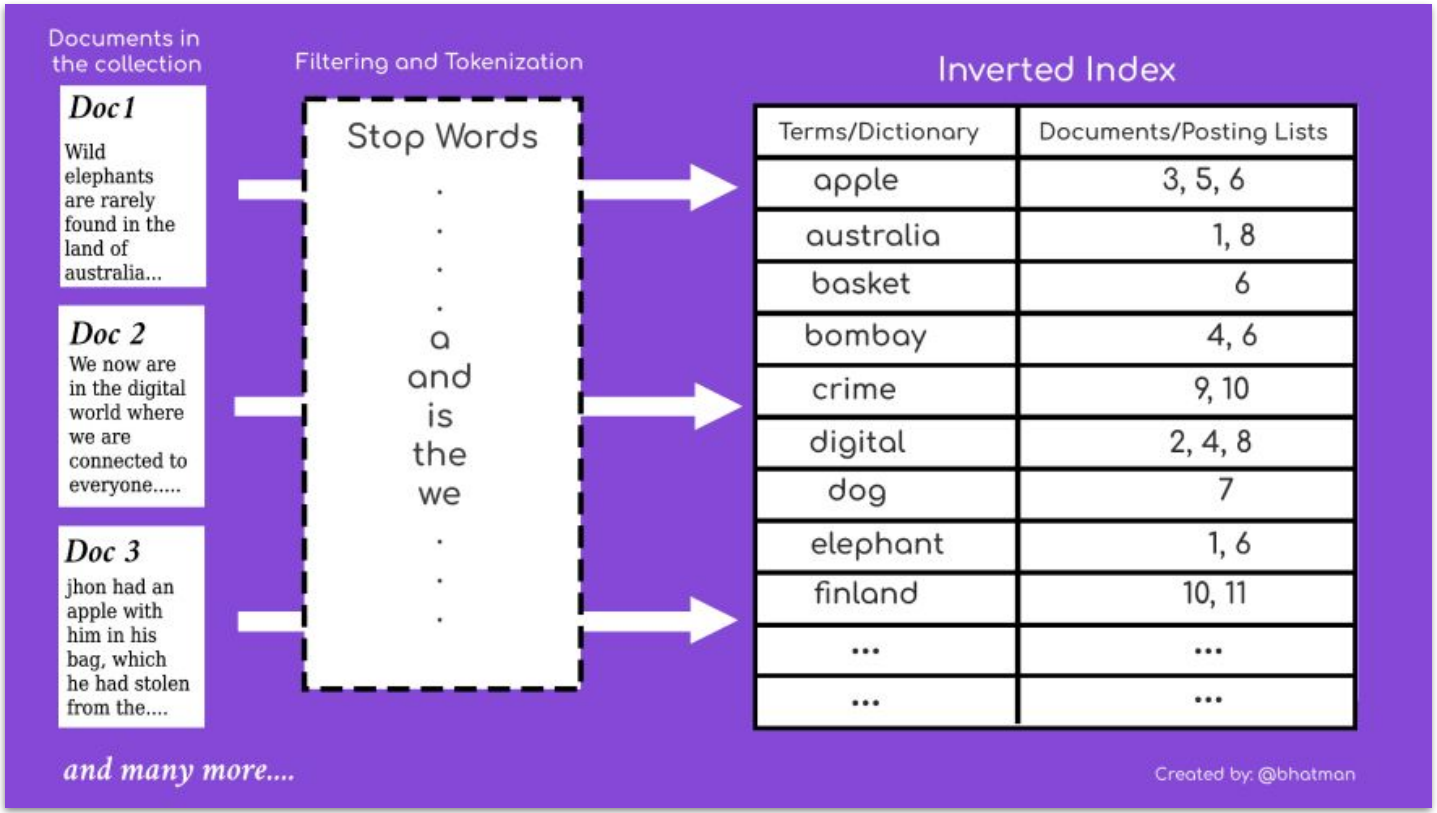
- 一种特殊的字典数据结构，将terms映射到包含它的文档
  - Analyzer -> Token
  - 其它 -> Term

两个文本字段：

“Hello, World!”
“Hello, Mate”

Word	Doc Num
hello	1, 2
world	1
mate	2

复杂一点的例子


Documents in the collection
Doc 1
Wild elephants are rarely found in the land of australia...
Doc 2
We now are in the digital world where we are connected to everyone.....
Doc 3
jhon had an apple with him in his bag, which he had stolen from the....
and many more....
Filtering and Tokenization
Stop Words
a
and
is
the
we
Inverted Index
Terms/Dictionary
Documents/Posting Lists
apple 3, 5, 6
australia 1, 8
basket 6
bombay 4, 6
crime 9, 10
digital 2, 4, 8
dog 7
elephant 1, 6
finland 10, 11
Created by: @bhatman

倒排索引(Inverted Index)

- 一种特殊的字典数据结构，将terms映射到包含它的文档
  - Analyzer -> Token
  - 其它 -> Term
- Terms根据字典序进行排序
- 只是简化表示，实际除了Terms/DocIds，还存储更多信息，例如词频等(relevance scoring)
- 对于一个文档中的每一个Text字段，都会建立对应的倒排索引
- 其它字段类型使用不同的数据结构存储
  - BKD树用于数值/日期/地理位置等，方便范 围查询

倒排索引小节

- 文本字段的值经过分析，存储为倒排索引(Inverted Index)
- 一种适用于根据terms进行快速查询的数据结构
- 每个文本字段有一个独立的Inverted Index
- 类似字典数据结构，terms -> doc映射
- Terms按照字典序排列，快速查询
- 底层基于Apache Lucene
- 除了terms/docId，还包括relevance scoring相关信息(后续讲解)
- ES(Lucene)还支持其它存储结构，BKD用于数值/日期/地理位置