一 互 联 世 界 的 记 忆 一

搭建投资与创业者的分析与决策平台

AN ANALYTIC PLATFORM FOR THIRD MARKET INVESTORS

知识检索与搜索

郑锦光

搭建投资与创业者的分析与决策平台

Recap

- **知识抽取 –** Statistical Methods, Pattern/RegEx Methods

- **知识表示 –** RDF, JSON, JSON-LD, etc.
-
-
- **知识存储 –** PostgreSQL, OrientDB, Neo4j

New Topic

- **知识抽取 –** Statistical Methods, Pattern/RegEx Methods

- **知识表示 –** RDF, JSON, JSON-LD, etc.
-
-
- **知识存储 –** PostgreSQL, OrientDB, Neo4j
-
-
- **知识检索 –** Lucene, Elasticsearch
-

知识检索

- Free text search – fast, scalable
- If design/model appropriately, it is enough to support simple relation query
-
-
- Not for complex knowledge models/queries
-
-
-
-

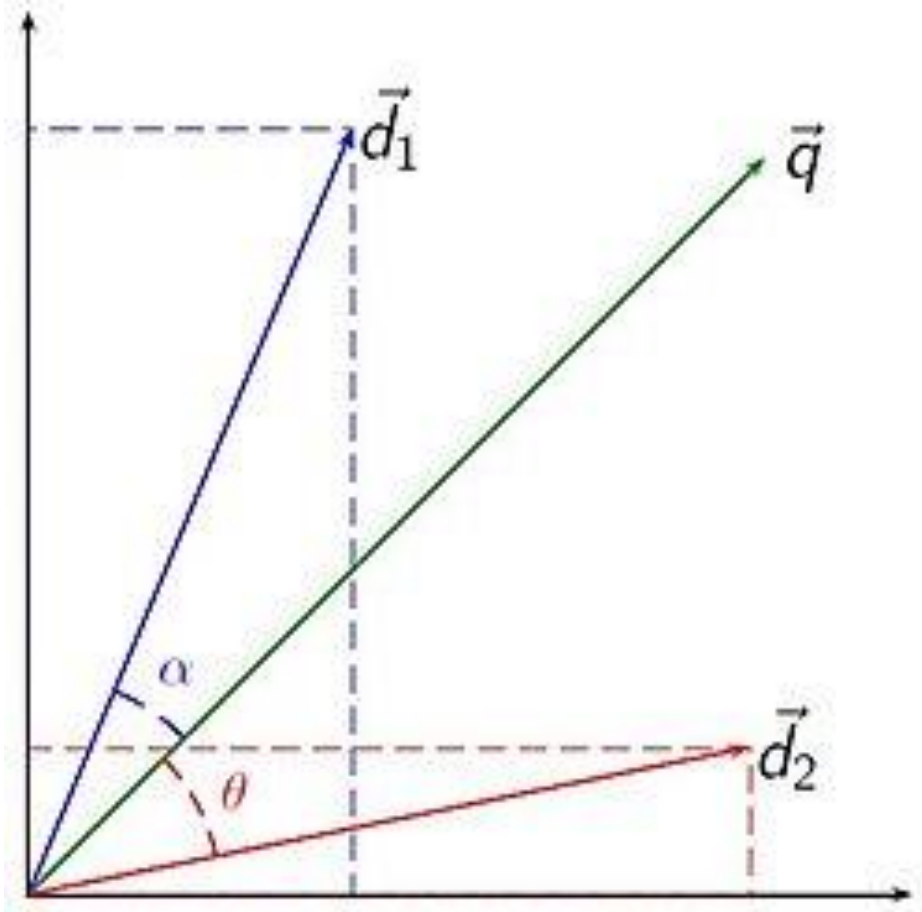
知识检索

-
- 倒排索引 (Inverted Index)
  - an inverted index (also referred to as postings file or inverted file) is an index data structure storing a mapping from content, such as words or numbers, to its locations in a database file, or in a document or a set of documents
-
- 向量空间模型 (Vector Space Model)
  - an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms.
-
-
-
-
-

倒排索引

单词ID	单词	文档频率	倒排列表（DocID;TF;<POS>）
1	谷歌	5	(1;1;<1>), (2;1;<1>), (3;2;<1;6>), (4;1;<1>), (5;1;<1>)
2	地图	5	(1;1;<2>), (2;1;<2>), (3;1;<2>), (4;1;<2>), (5;1;<2>)
3	之父	4	<1;1;<3>), (2;1;<3>), (4;1;<3>), (5;1;<3>)
4	跳槽	2	(1;1;<4>), (4;1;<4>)
5	Facebook	5	(1;1;<5>), (2;1;<5>), (3;1;<8>), (4;1;<5>), (5;1;<8>)
6	加盟	3	(2;1;<4>), (3;1;<7>), (5;1;<5>)
7	创始人	1	(3;1;<3>)
8	拉斯	2	(3;1;<4>), (5;1;<4>)
9	离开	1	(3;1;<5>)
10	与	1	(4;1;<6>)

向量空间模型

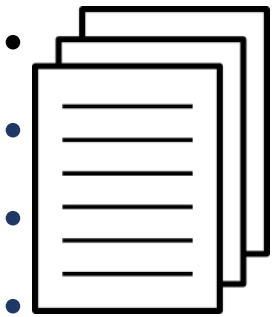


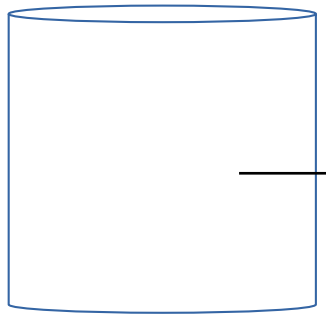
- TF*IDF (term frequency * inverse document frequency)

$$w_{t,d} = \mathrm{tf}_{t,d} \cdot \log \frac{|D|}{|\{d' \in D \mid t \in d'\}|}$$

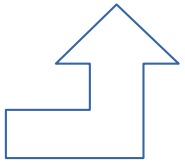
单词ID	单词	文档频率	倒排列表（DocID;TF;<POS>）
1	谷歌	5	(1;1;<1>), (2;1;<1>), (3;2;<1;6>), (4;1;<1>), (5;1;<1>)
2	地图	5	(1;1;<2>), (2;1;<2>), (3;1;<2>), (4;1;<2>), (5;1;<2>)
3	之父	4	<1;1;<3>), (2;1;<3>), (4;1;<3>), (5;1;<3>)
4	跳槽	2	(1;1;<4>), (4;1;<4>)
5	Facebook	5	(1;1;<5>), (2;1;<5>), (3;1;<8>), (4;1;<5>), (5;1;<8>)
6	加盟	3	(2;1;<4>), (3;1;<7>), (5;1;<5>)
7	创始人	1	(3;1;<3>)
8	拉斯	2	(3;1;<4>), (5;1;<4>)
9	离开	1	(3;1;<5>)
10	与	1	(4;1;<6>)

Overview

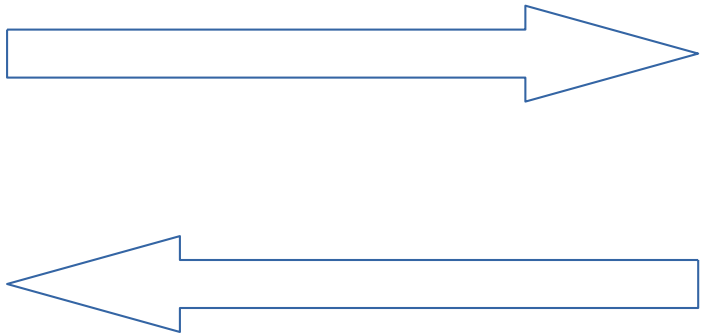


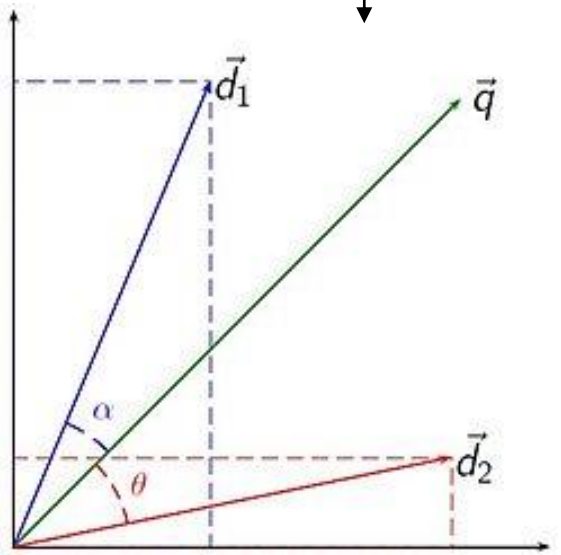


单词ID	单词	文档频率	倒排列表（DocID;TF;<POS>）
1	谷歌	5	(1;1;<1>), (2;1;<1>), (3;2;<1;6>), (4;1;<1>), (5;1;<1>)
2	地图	5	(1;1;<2>), (2;1;<2>), (3;1;<2>), (4;1;<2>), (5;1;<2>
3	之父	4	<1;1;<3>), (2;1;<3>), (4;1;<3>), (5;1;<3>
4	跳槽	2	(1;1;<4>), (4;1;<4>
5	Facebook	5	(1;1;<5>), (2;1;<5>), (3;1;<8>), (4;1;<5>), (5;1;<8>
6	加盟	3	(2;1;<4>), (3;1;<7>), (5;1;<5>
7	创始人	1	(3;1;<3>
8	拉斯	2	(3;1;<4>), (5;1;<4>
9	离开	1	(3;1;<5>
10	与	1	(4;1;<6>



© Can Stock Photo - csp17657725





- **Lucene and Elasticsearch**

Lucene

-

Apache LuceneTM is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform.

-
- 向量空间模型 (Vector Space Model)
  - an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms.
-
-
-
-

THANKS!

MEMECT
文因互联
一互联世界的记忆一
http://memect.com
contact@memect.com