Approaches to Information Retrieval

Naive way to Information Retrieval

Naive way to Information Retrieval

Query: I want a book on Analytics

Naive way to Information Retrieval

Query: I want a book on Analytics

Collection: 4 documents related to either Analytics or Big-data or both

Naive way to Information Retrieval

Query: I want a book on Analytics

Collection: 4 documents related to either Analytics or Big-data or both

Processed query:
Documents containing the words Book and Analytics but not Big-Data

Naive way to Information Retrieval

Query: Book AND Analytics AND NOT Big-Data

Naive way to Information Retrieval

Query: Book AND Analytics AND NOT Big-Data

Solution:

- Scan all the documents

Analytics Vidhya

Naive way to Information Retrieval

Query: Book AND Analytics AND NOT Big-Data

Solution:

- Scan all the documents
- Find the documents containing the words Book AND Analytics

Analytics Vidhya

Naive way to Information Retrieval

Query: Book AND Analytics AND NOT Big-Data

Solution:

- Scan all the documents
- Find the documents containing the words Book AND Analytics
- Then remove the documents containing the word Big-Data

Analytics Vidhya

Challenge: Naive way to Information Retrieval

- Slow for large corpus of data

Term - Document Incidence Matrix

Query: Book AND Analytics AND NOT Big-Data

Term - Document Incidence Matrix

Term / Words

Analytics				
Big-Data				
Book				
for				
The				
Tree				

Term - Document Incidence Matrix

Documents

Term / Words	Doc 1	Doc 2	Doc 3	Doc 4
Analytics				
Big-Data				
Book				
for				
The				
Tree				

Term - Document Incidence Matrix

Documents

Term / Words	Doc 1	Doc 2	Doc 3	Doc 4
Analytics	1	1	0	0
Big-Data				
Book				
for				
The				
Tree				

Term - Document Incidence Matrix

Documents

Term / Words	Doc 1	Doc 2	Doc 3	Doc 4
Analytics	1	1	0	0
Big-Data	0	1	0	0
Book				
for				
The				
Tree				

Term - Document Incidence Matrix

Documents

Term / Words	Doc 1	Doc 2	Doc 3	Doc 4
Analytics	1	1	0	0
Big-Data	0	1	0	0
Book	1	1	1	0
for	1	1	1	1
The	1	1	1	1
Tree	0	1	1	0

Term - Document Incidence Matrix

	Doc 1	Doc 2	Doc 3	Doc 4
Analytics	1	1	0	0
Big-Data	0	1	0	0
Book	1	1	1	0
for	1	1	1	1
The	1	1	1	1
Tree	0	1	1	0

Term - Document Incidence Matrix

Book

AND

Analytics

AND NOT

Big-Data

	Doc 1	Doc 2	Doc 3	Doc 4
Analytics	1	1	0	0
Big-Data	0	1	0	0
Book	1	1	1	0
for	1	1	1	1
The	1	1	1	1
Tree	0	1	1	0

Term - Document Incidence Matrix

Book

AND

Analytics

AND NOT

Big-Data

	Doc 1	Doc 2	Doc 3	Doc 4
Analytics	1	1	0	0
Big-Data	0	1	0	0
Book	1	1	1	0
for	1	1	1	1
The	1	1	1	1
Tree	0	1	1	0

Term - Document Incidence Matrix

Book

AND

Analytics

AND NOT

Big-Data

	Doc 1	Doc 2	Doc 3	Doc 4
Analytics	1	1	0	0
- Big-Data	1	0	1	1
Book	1	1	1	0
for	1	1	1	1
The	1	1	1	1
Tree	0	1	1	0

Term - Document Incidence Matrix

Book

AND

Analytics

AND NOT

Big-Data

	Doc 1	Doc 2	Doc 3	Doc 4
Analytics	1	1	0	0
- Big-Data	1	0	1	1
Book	1	1	1	0
for	1	1	1	1
The	1	1	1	1
Tree	0	1	1	0
AND				

Term - Document Incidence Matrix

Book

AND

Analytics

AND NOT

Big-Data

	Doc 1	Doc 2	Doc 3	Doc 4
Analytics	1	1	0	0
- Big-Data	1	0	1	1
Book	1	1	1	0
for	1	1	1	1
The	1	1	1	1
Tree	0	1	1	0
AND	1			

Term - Document Incidence Matrix

Book

AND

Analytics

AND NOT

Big-Data

	Doc 1	Doc 2	Doc 3	Doc 4
Analytics	1	1	0	0
- Big-Data	1	0	1	1
Book	1	1	1	0
for	1	1	1	1
The	1	1	1	1
Tree	0	1	1	0
AND	1	0	0	0

Term - Document Incidence Matrix

Book

AND

Analytics

AND NOT

Big-Data

	Doc 1	Doc 2	Doc 3	Doc 4
Analytics	1	1	0	0
- Big-Data	1	0	1	1
Book	1	1	1	0
for	1	1	1	1
The	1	1	1	1
Tree	0	1	1	0
AND	1	0	0	0

Term - Document Incidence Matrix

Book

AND

Tree

AND NOT

Analytics

	Doc 1	Doc 2	Doc 3	Doc 4
Analytics	1	1	0	0
Big-Data	0	1	0	0
Book	1	1	1	0
for	1	1	1	1
The	1	1	1	1
Tree	0	1	1	0

Term - Document Incidence Matrix

Book

AND

Tree

AND NOT

Analytics

	Doc 1	Doc 2	Doc 3	Doc 4
- Analytics	0	0	1	1
Big-Data	0	1	0	0
Book	1	1	1	0
for	1	1	1	1
The	1	1	1	1
Tree	0	1	1	0

Term - Document Incidence Matrix

Book

AND

Tree

AND NOT

Analytics

	Doc 1	Doc 2	Doc 3	Doc 4
- Analytics	0	0	1	1
Big-Data	0	1	0	0
Book	1	1	1	0
for	1	1	1	1
The	1	1	1	1
Tree	0	1	1	0
AND				

Term - Document Incidence Matrix

Book

AND

Tree

AND NOT

Analytics

	Doc 1	Doc 2	Doc 3	Doc 4
- Analytics	0	0	1	1
Big-Data	0	1	0	0
Book	1	1	1	0
for	1	1	1	1
The	1	1	1	1
Tree	0	1	1	0
AND	0	0	1	0

Term - Document Incidence Matrix

Book

AND

Tree

AND NOT

Analytics

	Doc 1	Doc 2	Doc 3	Doc 4
- Analytics	0	0	1	1
Big-Data	0	1	0	0
Book	1	1	1	0
for	1	1	1	1
The	1	1	1	1
Tree	0	1	1	0
AND	0	0	1	0

Challenges: Term - Document Incidence Matrix

Challenges: Term - Document Incidence Matrix

- Inefficient with bigger collection

Challenges: Term - Document Incidence Matrix

- Inefficient with bigger collection
  - Number of documents: 1 million = 10^6

Challenges: Term - Document Incidence Matrix

- Inefficient with bigger collection
  - Number of documents: 1 million = 10^6
  - Number of unique words: 5,000

Challenges: Term - Document Incidence Matrix

- Inefficient with bigger collection
  - Number of documents: 1 million = 10^6
  - Number of unique words: 5,000
  - Size of incidence matrix = 5,000 X 10^6

Thank You

Analytics
Vidhya