TD02 Corrected - Information Retrieval

Third Year ISIL - University of M'Sila

Exercise 01

Documents:

- d1 = Big cats are nice and funny
- d2 = Small dogs are better than big dogs
- d3 = Small cats are afraid of small dogs
- d4 = Big cats are not afraid of small dogs
- d5 = Funny cats are not afraid of small dogs

(a) Tokenization

- d1 = {Big, cats, are, nice, and, funny}
- d2 = {Small, dogs, are, better, than, big, dogs}
- d3 = {Small, cats, are, afraid, of, small, dogs}
- d4 = {Big, cats, are, not, afraid, of, small, dogs}
- d5 = {Funny, cats, are, not, afraid, of, small, dogs}

(b) Normalization (lowercase + singular)

- d1 = {big, cat, are, nice, and, funny}
- d2 = {small, dog, are, better, than, big, dog}
- d3 = {small, cat, are, afraid, of, small, dog}
- d4 = {big, cat, are, not, afraid, of, small, dog}
- d5 = {funny, cat, are, not, afraid, of, small, dog}

(c) Dictionary

$$\{big, cat, are, nice, and, funny, small, dog, better, than, afraid, of, not\}$$

Pseudo-code

Tokenization

```
function tokenize(document):
    tokens = split(document, " ")
    return tokens
```

Normalization

```
function normalize(tokens):
    for each token in tokens:
        token = toLowerCase(token)
        if token ends with 's':
            token = removeLastChar(token)
    return tokens
```

Dictionary Construction

```
function buildDictionary(documents):
    dictionary = empty set
    for each doc in documents:
        tokens = normalize(tokenize(doc))
        for each token in tokens:
            add token to dictionary
    return dictionary
```

—

Exercise 02

Term-Document Incidence Matrix

Term	d1	d2	d3	d4	d5
big	1	1	0	1	0
cat	1	0	1	1	1
dog	0	1	1	1	1
funny	1	0	0	0	1
nice	1	0	0	0	0
small	0	1	1	1	1
afraid	0	0	1	1	1
not	0	0	0	1	1

—

Exercise 03

Query q1 = funny AND dog

funny = {d1, d5}
dog = {d2, d3, d4, d5}

$$Result = \{d5\}$$

Query q2 = nice OR dog

nice = {d1}
dog = {d2, d3, d4, d5}

$$Result = \{d1, d2, d3, d4, d5\}$$

Query q3 = big AND dog AND NOT funny

big = {d1, d2, d4}
dog = {d2, d3, d4, d5}
funny = {d1, d5}

$$big \cap dog = \{d2, d4\}$$

$$NOT\ funny = \{d2, d3, d4\}$$

$$Result = \{d2, d4\}$$

—

Exercise 04

Stop words removed: is, and, in, the, was, each, other

Filtered Documents

- Doc1 = italy world champion 2006
- Doc2 = germany italy played semifinal
- Doc3 = germany semifinal 2006
- Doc4 = germany won semifinal italy 1990

Inverted Index

- italy → {1,2,4}
- world → {1}
- champion → {1}
- 2006 → {1,3}
- germany → {2,3,4}
- played → {2}
- semifinal → {2,3,4}
- won → {4}
- 1990 → {4}

—

Exercise 05

Documents

- Doc1 breakthrough drug for schizophrenia
- Doc2 new schizophrenia drug
- Doc3 new approach for treatment of schizophrenia
- Doc4 new hopes for schizophrenia patients

Incidence Matrix

Term	D1	D2	D3	D4
breakthrough	1	0	0	0
drug	1	1	0	0
new	0	1	1	1
approach	0	0	1	0
treatment	0	0	1	0
hopes	0	0	0	1
patients	0	0	0	1
schizophrenia	1	1	1	1
for	1	0	1	1
of	0	0	1	0

Inverted Index

- breakthrough → {1}
- drug → {1,2}
- new → {2,3,4}
- approach → {3}
- treatment → {3}
- hopes → {4}
- patients → {4}
- schizophrenia → {1,2,3,4}
- for → {1,3,4}
- of → {3}

Queries

1. schizophrenia AND drug

$$schizophrenia = \{1, 2, 3, 4\}, \quad drug = \{1, 2\}$$

$$Result = \{1, 2\}$$

2. for AND NOT(drug OR approach)

$$drug \cup approach = \{1, 2, 3\}$$

$$NOT = \{4\}$$

$$for = \{1, 3, 4\}$$

$$Result = \{4\}$$

—