BÀI TẬP LÝ THUYẾT 1

Vũ Cao Nguyên – 18600187

Bài 1:

const **Stopwords** = ["how", "and", "or", "an", "a", "the", "there", "that", "of", "for", "to", "is", "are", "can", "has", "with", "within", "in", "on", "about", "as", "well", "very",];

Stemming:

- **Chuyển danh từ thành thể số ít**
- **Chuyển động từ về thể nguyên bản**

Viết thường, không xét ký tự ngoài chữ cái

Từ được sắp xếp theo thứ tự từ điển

d1: information retrieval model govern document query represent relevance document user query define.

d2: Information retrieval area study concern search document, information document, metadata document, search structure storage, relational database, World Wide Web.

d3: Web search become important information age. Increase exposure page Web result significant financial gain fame organization individual.

Bài 2:

d1: I do not like green eggs and ham.

$\Longrightarrow$ I do, do not, not like, like green, green egg, egg ham

d2: I do not like them, Sam I am.

$\Longrightarrow$ I do, do not, not like, like them, them sam, sam I, I am

$sim(\mathbf{d1}, \mathbf{d2}) = 3/10 = 0.33$

Bài 3:

id1: breakthrough drug for schizophrenia
1 2 3 4

id2: new schizophrenia drug
1 2 3

id3: new approach for treatment of schizophrenia
1 2 3 4 5 6

id4: new hopes for schizophrenia patients
1 2 3 4 5

V = { breakthrough , drug, schizophrenia, new, approach, treatment, hopes, patiens}.

Breakthrough: id1	Breakthrough: <id1, 1, [1]
Drug: id1, id2	Drug: <id1, 1, [2]>, <id2, 1 [3]>
Schizophrenia: id1, id2, id3, id4	Schizophrenia: <id1, 1 , [4]>, <id2,1,[2] >, <id3,1,[6]>, <id4,1,[4]>
New: id2, id3, id4	New: <id2,1,[1]>, <id3,1,[1]>, <id4,1,[1]>
Approach: id3	Approach: <id3,1,[2]>
Treatment: id3	Treatment: <id3,1,[4]>
Hopes: id4	Hopes: <id4,1,[2]>
Patiens: id4	Patiens: <id4,1,[5]>

- Truy vấn: schizophrenia AND drug
  - Bước 1:
    Schizophrenia: <id1, 1 , [4]>, <id2,1,[2] >, <id3,1,[6]>, <id4,1,[4]>
    Drug: <id1, 1, [2]>, <id2, 1 [3]>
  - Bước 2:
    → id1, id2
  - Bước 3:
    → rank(id2) > rank(id1)
    → **KQ**: new schizophrenia drug

- Truy vấn: for AND NOT(drug OR approach)
  - Bước 1:
    Drug: <id1, 1, [2]>, <id2, 1 [3]>
    Approach: <id3,1,[2]>
  - Bước 2:
    → Không có tài liệu chứa 2 từ
    → **KQ**: ∅ (rỗng)