BÀI TẬP LÝ THUYẾT 2

Vũ Cao Nguyên – 18600187

Bài 1:

- Cho tập hợp các tài liệu **d***i* và truy vấn **q** như sau:
  d1: new york times
  d2: new york post
  d3: los angeles times
  q: new new times
- Đếm số lần xuất hiện của mỗi từ trong **d***i* và **q**

	new	york	times	post	los	angeles
d1	1	1	1	0	0	0
d2	1	1	0	1	0	0
d3	0	0	1	0	1	1
q	2	0	1	0	0	0

- Giả sử rằng mỗi thành phần trong vector trọng số là số lần xuất hiện của từ tương ứng. Như vậy, ta có:
  d1 = [1, 1, 1, 0, 0 ,0]
  d2 = [1, 1, 0, 1, 0, 0]
  d3 = [0, 0, 1, 0, 1, 1]
  q = [2, 0, 1, 0, 0, 0]
- Xếp hạng các tài liệu theo giá trị **Cosine similarity** của mỗi tài liệu đối với **q** trong mô hình không gian vector.
  $cosine(\mathbf{d1}, \mathbf{q}) = 0.775$
  $cosine(\mathbf{d2}, \mathbf{q}) = 0.516$
  $cosine(\mathbf{d3}, \mathbf{q}) = 0.258$