# **Week 7 –** **Textual Analysis**

# **Exercise 01: Syntatical analysis**

Assume you have a set of documents each of which is in either English or in Spanish. The collection is given in below Table 01:

|  |  |
| --- | --- |
| **DocID** | **Document Text** |
| 1 | hello |
| 2 | open house |
| 3 | mi casa |
| 4 | hola Professor |
| 5 | hola y bienvenido |
| 6 | hello and welcome |

* Construct the appropriate term-document matrix C to use for a collection consisting of these documents.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Doc1 | Doc2 | Doc2 | Doc4 | Doc5 | Doc6 |
| hello | 1 | 0 | 0 | 0 | 0 | 1 |
| open | 0 | 1 | 0 | 0 | 0 | 0 |
| house | 0 | 1 | 0 | 0 | 0 | 0 |
| Mi | 0 | 0 | 1 | 0 | 0 | 0 |
| Casa | 0 | 0 | 1 | 0 | 0 | 0 |
| Hola | 0 | 0 | 0 | 1 | 1 | 0 |
| Professor | 0 | 0 | 0 | 1 | 0 | 0 |
| Y | 0 | 0 | 0 | 0 | 1 | 0 |
| Bienvenido | 0 | 0 | 0 | 0 | 1 | 0 |
| and | 0 | 0 | 0 | 0 | 0 | 1 |
| welcome | 0 | 0 | 0 | 0 | 0 | 1 |

* Construct the normalized tf-idf weights matrix W.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| hello | open | house | hi | casa | hola | Professor | y | bienvenido | and | welcom |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| hello | open | house | hi | casa | hola | Professor | y | bienvenido | and | welcom |
| 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 |

Normalized

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Doc1 | Doc2 | Doc2 | Doc4 | Doc5 | Doc6 |
| hello | 0.707 | 0 | 0 | 0 | 0 | 0.707 |
| open | 0 | 0.707 | 0 | 0 | 0 | 0 |
| house | 0 | 0.707 | 0 | 0 | 0 | 0 |
| mi | 0 | 0 | 0.707 | 0 | 0 | 0 |
| casa | 0 | 0 | 0.707 | 0 | 0 | 0 |
| hola | 0 | 0 | 0 | 0.707 | 0.707 | 0 |
| Professor | 0 | 0 | 0 | 0.707 | 0 | 0 |
| y | 0 | 0 | 0 | 0 | 0.707 | 0 |
| bienvenido | 0 | 0 | 0 | 0 | 0.707 | 0 |
| and | 0 | 0 | 0 | 0 | 0 | 0.707 |
| welcome | 0 | 0 | 0 | 0 | 0 | 0.707 |

# **Exercise 02: Words Representation**

Given some words with their semantic vectors as following:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| banana | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 1 |
| monkey | 2 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| orange | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 1 |
| elephant | 0 | 0 | 2 | 0 | 1 | 0 | 1 | 0 |

* Compute the cosine similarities of each pair of words.

similarity(banana, monkey) = 0

similarity(banana, orange) = 0.33

similarity(banana, elephant) = 0

similarity(monkey, orange) = 0

similarity(monkey, elephant) = 0.33

similarity(orange, elephant) = 0

* Compute distance of each pair of words using euclide distance.

dist(banana, monkey) = sqrt(12) = 3.46

dist(banana, orange) = sqrt(8) = 2.828

dist(banana, elephant) = sqrt(12) = 3.46

dist(monkey, orange) = squrt(12) = 3.46

dist(monkey, elephant) = sqrt(8) = 2.828

dist(orange, elephant) = sqrt(12) = 3.46

Find the closest pairs. Justify the semantic rationality against the above vector representation.

Closest pairs: (banana, orange), (monkey, elephant).

The result is appropriate based on calculation and semantic angle: “monkey” and “elephant” are both animals and “orange” and “banana” are both fruit.