# **Week 7 –** **Textual Analysis**

# **Exercise 01: Syntactical 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 |  |

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Doc1 | Doc2 | Doc3 | Doc4 | Doc5 | Doc6 |
| hello | 1 | 0 | 0 | 0 | 0 | 1 |
| open | 0 | 1 | 0 | 0 | 0 | 0 |
| hola | 0 | 0 | 0 | 1 | 1 | 0 |
| y | 0 | 0 | 0 | 0 | 1 | 0 |
| House | 0 | 1 | 0 | 0 | 0 | 0 |
| Mi | 0 | 0 | 1 | 0 | 0 | 0 |
| Casa | 0 | 0 | 1 | 0 | 0 | 0 |
| professor | 0 | 0 | 0 | 1 | 0 | 0 |
| bienvenido | 0 | 0 | 0 | 0 | 1 | 0 |
| welcome | 0 | 0 | 0 | 0 | 0 | 1 |

* Construct the normalized tf-idf weights matrix W.

Document frequency

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Hello | Open | Hola | Y | House | Mi | Casa | Professor | bienvenido | welcome |
| df | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Idf

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Hello | Open | Hola | Y | House | Mi | Casa | Professor | bienvenido | welcome |
| idf | 0.77 | 0.77 | 0.77 | 0.77 | 0.77 | 0.77 | 0.77 | 0.77 | 0.77 | 0.77 |

Tf-idf

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Hello | Open | Hola | Y | House | Mi | Casa | Professor | bienvenido | welcome |
| Doc1 | 0.77 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Doc2 | 0 | 0.77 | 0 | 0 | 0.77 | 0 | 0 | 0 | 0 | 0 |
| Doc3 | 0 | 0 | 0 | 0 | 0 | 0.77 | 0.77 | 0 | 0 | 0 |
| Doc4 | 0 | 0 | 0.77 | 0 | 0 | 0 | 0 | 0.77 | 0 | 0 |
| Doc5 | 0 | 0 | 0.77 | 0.77 | 0 | 0 | 0 | 0 | 0.77 | 0 |
| Doc6 | 0.77 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.77 |

Tf-idf(normalize 2)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Hello | Open | Hola | Y | House | Mi | Casa | Professor | bienvenido | welcome |
| Doc1 | 0.707 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Doc2 | 0 | 0.707 | 0 | 0 | 0.707 | 0 | 0 | 0 | 0 | 0 |
| Doc3 | 0 | 0 | 0 | 0 | 0 | 0.707 | 0.707 | 0 | 0 | 0 |
| Doc4 | 0 | 0 | 0.707 | 0 | 0 | 0 | 0 | 0.707 | 0 | 0 |
| Doc5 | 0 | 0 | 0.707 | 0.707 | 0 | 0 | 0 | 0 | 0.707 | 0 |
| Doc6 | 0.707 | 0 | 0 | 0 | 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.
* Compute distance of each pair of words using euclide distance.
* Find the closest pairs. Justify the semantic rationality against the above vector representation.

Compute the cosine similarities:

sim(banana,monkey)=0

sim(banana,orange)==0.33

sim(banana,elephant)= 0

sim(monkey,orange)= 0

sim(monkey,elephant) = 0.33

sim(orange,elephant)= 0

Euclide distance

D(banana,monkey)== 3.46

D(banana,orange)= 2.82

D(banana,elephant)= 3.46

D(monkey,orange)= 3.46

D(monkey,elephant) =2.82

D(orange,elephant)= 3.46

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.