# **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 | Doc3 | 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 |
| welcome | 0 | 0 | 0 | 0 | 0 | 1 |

C = [1 0 0 0 0 1

0 1 0 0 0 0

0 1 0 0 0 0

0 0 1 0 0 0

0 0 1 0 0 0

0 0 0 1 1 0

0 0 0 1 0 0

0 0 0 0 1 0

0 0 0 0 1 0

0 0 0 0 0 1 ]

* Construct the normalized tf-idf weights matrix W.
* tf

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | hello | open | house | mi | casa | hola | Professor | y | bienvenido | welcome |
| Doc1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Doc2 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Doc3 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| Doc4 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| Doc5 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| Doc6 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

* df

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | hello | open | house | mi | casa | hola | Professor | y | bienvenido | welcome |
| df | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

* idf

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | hello | open | house | mi | casa | hola | Professor | y | bienvenido | welcome |
| idf | 0.48 | 0.78 | 0.78 | 0.78 | 0.78 | 0.48 | 0.78 | 0.78 | 0.78 | 0.78 |

* tf-idf

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | hello | open | house | mi | casa | hola | Professor | y | bienvenido | welcome |
| Doc1 | 0.48 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Doc2 | 0 | 0.78 | 0.78 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Doc3 | 0 | 0 | 0 | 0.78 | 0.78 | 0 | 0 | 0 | 0 | 0 |
| Doc4 | 0 | 0 | 0 | 0 | 0 | 0.48 | 0.78 | 0 | 0 | 0 |
| Doc5 | 0 | 0 | 0 | 0 | 0 | 0.48 | 0 | 0.78 | 0.78 | 0 |
| Doc6 | 0.48 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.78 |

* normalized tf-idf

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | hello | open | house | mi | casa | hola | Professor | y | bienvenido | welcome |
| Doc1 | 0.7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Doc2 | 0 | 0.7 | 0.7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Doc3 | 0 | 0 | 0 | 0.7 | 0.7 | 0 | 0 | 0 | 0 | 0 |
| Doc4 | 0 | 0 | 0 | 0 | 0 | 0.7 | 0.7 | 0 | 0 | 0 |
| Doc5 | 0 | 0 | 0 | 0 | 0 | 0.7 | 0 | 0.7 | 0.7 | 0 |
| Doc6 | 0.7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.7 |

# **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.

rbanana = (2+1+1)/8=0.5

rmonkey= (2+1+1)/8=0.5

rorange = (2+1+1)/8=0.5

relephant = (2+1+1)/8=0.5

sim(banana, monkey)=0

sim(banana, orange)=(1+1)/(\*)=0.33

sim(banana, elephant)=0

sim(monkey, orange)=0

sim(monkey, elephant)=0.33

sim(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) = sqrt(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 from calculations: (banana, orange), (monkey, elephant).
* The result is rational:

“monkey” and “elephant” are both animals

“orange” and “banana” are both fruits.