# **Week 5 –** **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 |
| and | 0 | 0 | 0 | 0 | 0 | 1 |
| welcome | 0 | 0 | 0 | 0 | 0 | 1 |

* Construct the normalized tf-idf weights matrix W.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | idf | Doc1 | Doc2 | Doc3 | Doc4 | Doc5 | Doc6 |
| hello | 1.58 | 1.58 | 0 | 0 | 0 | 0 | 1.58 |
| open | 2.58 | 0 | 2.58 | 0 | 0 | 0 | 0 |
| house | 2.58 | 0 | 2.58 | 0 | 0 | 0 | 0 |
| mi | 2.58 | 0 | 0 | 2.58 | 0 | 0 | 0 |
| casa | 2.58 | 0 | 0 | 2.58 | 0 | 0 | 0 |
| hola | 1.58 | 0 | 0 | 0 | 1.58 | 1.58 | 0 |
| Professor | 2.58 | 0 | 0 | 0 | 2.58 | 0 | 0 |
| y | 2.58 | 0 | 0 | 0 | 0 | 2.58 | 0 |
| bienvenido | 2.58 | 0 | 0 | 0 | 0 | 2.58 | 0 |
| and | 2.58 | 0 | 0 | 0 | 0 | 0 | 2.58 |
| welcome | 2.58 | 0 | 0 | 0 | 0 | 0 | 2.58 |

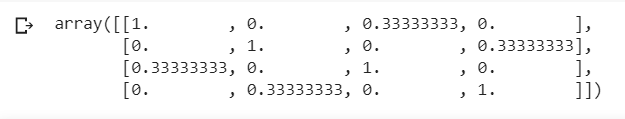
# **Exercise 02: Words Representation**

Given some words with their semantic vectors as following:

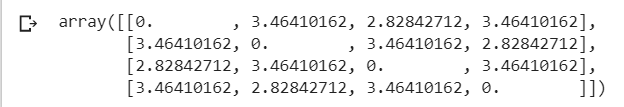
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| banana | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 1 |
| monkey | 2 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| orange | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 1 |
| elephant | 0 | 0 | 2 | 0 | 1 | 0 | 1 | 0 |

* 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.

Cosine similarities:



Euclidean distances:



The closest pairs are:

* “banana” and “orange” (2.828) - fruits
* “monkey” and “elephant” (2.828) - animals