# **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.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Word** | **IDF** | **Doc 1** | **Doc 2** | **Doc 3** | **Doc 4** | **Doc 5** | **Doc 6** |
| hello | 1.585 | 1.585 | 0 | 0 | 0 | 0 | 1.585 |
| open | 2.585 | 0 | 2.585 | 0 | 0 | 0 | 0 |
| house | 2.585 | 0 | 2.585 | 0 | 0 | 0 | 0 |
| mi | 2.585 | 0 | 0 | 2.585 | 0 | 0 | 0 |
| casa | 2.585 | 0 | 0 | 2.585 | 0 | 0 | 0 |
| hola | 1.585 | 0 | 0 | 0 | 1.585 | 1.585 | 0 |
| Professor | 2.585 | 0 | 0 | 0 | 2.585 | 0 | 0 |
| y | 2.585 | 0 | 0 | 0 | 0 | 2.585 | 0 |
| bienvenido | 2.585 | 0 | 0 | 0 | 0 | 2.585 | 0 |

# **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.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | banana | monkey | orange | elephant |
| banana | 1 | 0 | 0.33 | 0 |
| monkey | 0 | 1 | 0 | 0.33 |
| orange | 0.33 | 0 | 1 | 0 |
| elephant | 0 | 0.33 | 0 | 1 |

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | banana | monkey | orange | elephant |
| banana | 0 | 3.46 | 2.83 | 3.46 |
| monkey | 3.46 | 0 | 3.46 | 2.83 |
| orange | 2.83 | 3.46 | 0 | 3.46 |
| elephant | 3.46 | 2.83 | 3.46 | 0 |

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

Closest pairs: banana – orange and monkey – elephant.

Banana and orange are both fruits, while monkey and elephant are both animal, which paired together.