# **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 |  |  |  |  |  |  |
| open |  |  |  |  |  |  |
| house |  |  |  |  |  |  |
| mi |  |  |  |  |  |  |
| casa |  |  |  |  |  |  |
| hola |  |  |  |  |  |  |
| professor |  |  |  |  |  |  |
| y |  |  |  |  |  |  |
| bienvenido |  |  |  |  |  |  |
| welcome |  |  |  |  |  |  |
| and |  |  |  |  |  |  |

* Construct the normalized tf-idf weights matrix W.

TF:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 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 |
| And | 0 | 0 | 0 | 0 | 0 | 1 |

Idf:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Terms | Hello | Open | House | Mi | Casa | Hola | professor | y | bienvenido | welcome | and |
| idf | 0.544 | 0.845 | 0.845 | 0.845 | 0.845 | 0.544 | 0.845 | 0.845 | 0.845 | 0.845 | 0.845 |

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 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.
* The clossets pairs are banana – orange and monkey – elephant. The vector attributes are represent the pairing between animals with fruits. The larger number is better that show the closer between them.