# **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 |
| bienvenido | 0 | 0 | 0 | 0 | 1 | 0 |
| and | 0 | 0 | 0 | 0 | 0 | 1 |
| welcome | 0 | 0 | 0 | 0 | 0 | 1 |
| Y | 0 | 0 | 0 | 0 | 1 | 0 |

* Construct the normalized tf-idf weights matrix W.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | IDF | Doc1 | Doc2 | Doc3 | Doc4 | Doc5 | Doc6 |
| hello | 1.5849 | 1.5849 | 0 | 0 | 0 | 0 | 1.5849 |
| open | 2.5849 | 0 | 2.5849 | 0 | 0 | 0 | 0 |
| house | 2.5849 | 0 | 2.5849 | 0 | 0 | 0 | 0 |
| mi | 2.5849 | 0 | 0 | 2.5849 | 0 | 0 | 0 |
| casa | 2.5849 | 0 | 0 | 2.5849 | 0 | 0 | 0 |
| hola | 1.5849 | 0 | 0 | 0 | 1.5849 | 1.5849 | 0 |
| professor | 2.5849 | 0 | 0 | 0 | 2.5849 | 0 | 0 |
| bienvenido | 2.5849 | 0 | 0 | 0 | 0 | 2.5849 | 0 |
| and | 2.5849 | 0 | 0 | 0 | 0 | 0 | 2.5849 |
| welcome | 2.5849 | 0 | 0 | 0 | 0 | 0 | 2.5849 |
| Y | 2.5849 | 0 | 0 | 0 | 0 | 2.5849 | 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.
* Compute distance of each pair of words using euclide distance.
* Find the closest pairs. Justify the semantic rationality against the above vector representation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Banana | Monkey | Orange | Elephant |
| Banana | 1 | 0 | 0.3333 | 0 |
| Monkey | 0 | 1 | 0 | 0.3333 |
| Orange | 0.3333 | 0 | 1 | 0 |
| Elephant | 0 | 0.3333 | 0 | 1 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Banana | Monkey | Orange | Elephant |
| Banana | 0 | 3.464 | 2.828 | 3.464 |
| Monkey | 3.564 | 0 | 3.464 | 2.828 |
| Orange | 2.828 | 3.464 | 0 | 3.464 |
| Elephant | 3.464 | 2.828 | 3.464 | 0 |

The closest pairs are Banana – Orange (Fruit), Monkey – Elephant (Animal), which is valid.