# **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 |
| and | 0 | 0 | 0 | 0 | 0 | 1 |
| welcome | 0 | 0 | 0 | 0 | 0 | 1 |

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

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Terms | hello | open | house | mi | casa | hola | Professor | y | bienvendio | and | welcome |
| Idf | 0.47 | 0.78 | 0.78 | 0.78 | 0.78 | 0.47 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 |

* Tf-Idf:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Terms | hello | open | house | mi | casa | hola | | Professor | y | bienvenido | and | welcome |
| Doc1 | 0.47 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 | 0 |
| Doc2 | 0 | 0.78 | 0.78 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 | 0 |
| Doc3 | 0 | 0 | 0 | 0.78 | 0.78 | | 0 | 0 | 0 | 0 | 0 | 0 |
| Doc4 | 0 | 0 | 0 | 0 | 0 | | 0.47 | 0.78 | 0 | 0 | 0 | 0 |
| Doc5 | 0 | 0 | 0 | 0 | 0 | | 0.47 | 0 | 0.78 | 0.78 | 0 | 0 |
| Doc6 | 0.47 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 0.78 | 0.78 |

# **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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Terms | 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 euclidean distance.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Terms | banana | monkey | orange | elephant |
| Banana | 1 | 3.46 | 3 | 3.46 |
| Monkey | 3.46 | 1 | 3.46 | 3 |
| Orange | 3 | 3 | 1 | 3.46 |
| Elephant | 3.46 | 3 | 3.46 | 1 |

* Find the closest pairs. Justify the semantic rationality against the above vector representation.
* With Cosine Similarity : We have 2 pairs have largest similarity (Banana, Orange), (Monkey, Elephant).
* With Euclidean distance : We have 4 pairs with most distance (Banana, Monkey), (Banana, Elephant), (Monkey, Orange), (Orange, Elephant).
* With 2 method above, some of pairs in method 1 are not correct in method 2. Each method apply in difference situation . In my case, because data we have is not continuous data so I recommend you use cosine similarity.