# **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.

Tf:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 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 |
| welcom | 0 | 0 | 0 | 0 | 0 | 1 |

* Construct the normalized tf-idf weights matrix W.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | hello | open | house | mi | casa | hola | professor | y | bienvenido | and | welcom |
| df | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |

Idf = log()

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | hello | open | House | mi | casa | hola | professor | y | bienvenido | and | welcom |
| idf | 0.477 | 0.778 | 0.778 | 0.778 | 0.778 | 0.477 | 0.778 | 0.778 | 0.778 | 0.778 | 0.778 |

Tf-idf = tf \* idf

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Doc1 | Doc2 | Doc3 | Doc4 | Doc5 | Doc6 |
| hello | 0.447 | 0 | 0 | 0 | 0 | 1.09 |
| open | 0 | 0.778 | 0 | 0 | 0 | 0 |
| house | 0 | 0.778 | 0 | 0 | 0 | 0 |
| mi | 0 | 0 | 0.778 | 0 | 0 | 0 |
| casa | 0 | 0 | 0.778 | 0 | 0 | 0 |
| hola | 0 | 0 | 0 | 0.447 | 0.447 | 0 |
| professor | 0 | 0 | 0 | 0.778 | 0 | 0 |
| y | 0 | 0 | 0 | 0 | 0.778 | 0 |
| bienvenido | 0 | 0 | 0 | 0 | 0.778 | 0 |
| and | 0 | 0 | 0 | 0 | 0 | 0.778 |
| welcom | 0 | 0 | 0 | 0 | 0 | 0.778 |

Normalized tf-idf

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | hello | open | house | mi | casa | hola | Professor | y | bienvenido | and | welcome |
| Doc1 | 0.477 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Doc2 | 0 | 0.707 | 0.707 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Doc3 | 0 | 0 | 0 | 0.707 | 0.707 | 0 | 0 | 0 | 0 | 0 | 0 |
| Doc4 | 0 | 0 | 0 | 0 | 0 | 0.523 | 0.853 | 0 | 0 | 0 | 0 |
| Doc5 | 0 | 0 | 0 | 0 | 0 | 0.4 | 0 | 0.649 | 0.649 | 0 | 0 |
| Doc6 | 0.4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.649 | 0.649 |

# **Exercise 02: Words Representation**

Given some words with their semantic vectors as following:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| V1 | banana | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 1 |
| V2 | monkey | 2 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| V3 | orange | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 1 |
| V4 | elephant | 0 | 0 | 2 | 0 | 1 | 0 | 1 | 0 |

* Compute the cosine similarities of each pair of words.
* cos(v1,v2) =
* cos(v1,v3) =
* cos(v1,v4) =
* cos(v2,v3) =
* cos(v2,v4) =
* cos(v3,v4) =
* Compute distance of each pair of words using euclide distance.
* d(v1,v2) = =
* d(v1,v3) = =
* d(v1,v4) = =
* d(v2,v3) = =
* d(v2,v4) = =
* d(v3,v4) = =
* Find the closest pairs. Justify the semantic rationality against the above vector representation.
* Closest pairs: (banana, orange), (monkey, elephant)