# **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 | 0.477 | 0.778 | 0.778 | 0.778 | 0.778 | 0.477 | 0.778 | 0.778 | 0.778 | 0.778 | 0.778 |

Tf-idf:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Doc1 | Doc2 | Doc3 | Doc4 | Doc5 | Doc6 |
| hello | 0.477 | 0 | 0 | 0 | 0 | 0.477 |
| 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 | 0.477 | 0.477 | 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 |
| welcome | 0 | 0 | 0 | 0 | 0 | 0.778 |
| And | 0 | 0 | 0 | 0 | 0 | 0.778 |

Normalized tf-idf:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Doc1 | Doc2 | Doc3 | Doc4 | Doc5 | Doc6 |
| hello | 0.707 | 0 | 0 | 0 | 0 | 0.707 |
| open | 0 | 0.707 | 0 | 0 | 0 | 0 |
| house | 0 | 0.707 | 0 | 0 | 0 | 0 |
| mi | 0 | 0 | 0.707 | 0 | 0 | 0 |
| casa | 0 | 0 | 0.707 | 0 | 0 | 0 |
| hola | 0 | 0 | 0 | 0.707 | 0.707 | 0 |
| professor | 0 | 0 | 0 | 0.707 | 0 | 0 |
| y | 0 | 0 | 0 | 0 | 0.707 | 0 |
| bienvenido | 0 | 0 | 0 | 0 | 0.707 | 0 |
| welcome | 0 | 0 | 0 | 0 | 0 | 0.707 |
| And | 0 | 0 | 0 | 0 | 0 | 0.707 |

# **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:0

Banana-orange: 2/(2.45\*2.45) = 1/3

Banana-elephant: 0

Monkey-orange: 0

Monkey-elephant: 2/(2.45\*2.45) = 1/3

Orange-elephant: 0

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

Banana-monkey: 3.46

Banana-orange: 2.83

Banana-elephant: 3.46

Monkey-orange: 3.46

Monkey-elephant: 2.83

Orange-elephant: 3.46

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

Banana-orange, monkey-elephant