# **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 | 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 |
| 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 |

Matrix:

* Construct the normalized tf-idf weights matrix W.

Document frequency (df)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Terms | hello | open | house | mi | casa | hola | proffesor | bienvenido | and | welcome | y |
| df | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Inverse document frequency (idf)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Terms | hello | open | house | mi | casa | hola | proffesor | bienvenido | and | welcome | y |
| df | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 |

Tfidf:

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

Tfidf (l2 normalize):

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Doc1 | Doc2 | Doc2 | Doc4 | Doc5 | Doc6 |
| hello | 0.71 | 0 | 0 | 0 | 0 | 0.78 |
| open | 0 | 0.71 | 0 | 0 | 0 | 0 |
| house | 0 | 0.71 | 0 | 0 | 0 | 0 |
| mi | 0 | 0 | 0.71 | 0 | 0 | 0 |
| casa | 0 | 0 | 0.71 | 0 | 0 | 0 |
| hola | 0 | 0 | 0 | 0.71 | 0.71 | 0 |
| Professor | 0 | 0 | 0 | 0.71 | 0 | 0 |
| bienvenido | 0 | 0 | 0 | 0 | 0.71 | 0 |
| and | 0 | 0 | 0 | 0 | 0 | 0.71 |
| welcome | 0 | 0 | 0 | 0 | 0 | 0.71 |
| y | 0 | 0 | 0 | 0 | 0.71 | 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.

cos(banana, monkey) = 0

cos(banana, orange) = 2/6 = 0.33

cos(banana, elephant) = 0

cos(monkey, orange) = 0

cos(monkey, elephant) = 2/6 = 0.33

cos(orange, elephant) = 0

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

d(banana, monkey) = sqrt(4 + 4 + 4) = 3.46

d(banana, orange) = sqrt(4 + 4) = 2.82

d(banana, elephant) = sqrt(12) = 3.46

d(monkey, orange) = sqrt(12) = 3.46

d(monkey, elephant)= sqrt(8) = 2.82

d(orange, elephant) = sqrt(12) = 3.46

* Find the closest pairs. Justify the semantic rationality against the above vector representation.
* {banana, orange} and {monkey, elephant} are two closet pairs of above vector representation.
* Banana and orange is belonged to fruit category and money and elephant belongs to animal category so that is the reason why the distance between those word pairs are the lowest.