# **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 |
| Bienvenido | 0 | 0 | 0 | 0 | 1 | 0 |
| Hola | 0 | 0 | 0 | 1 | 1 | 0 |
| professor | 0 | 0 | 0 | 1 | 0 | 0 |
| welcome | 0 | 0 | 0 | 0 | 0 | 1 |
| and | 0 | 0 | 0 | 0 | 0 | 1 |
| y | 0 | 0 | 0 | 0 | 1 | 0 |

* Construct the normalized tf-idf weights matrix W.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Hello | Open | House | Mi | Casa | Hola | Professor | Welcome | And | y | bienvenido |
| Df | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Idf:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Hello | Open | House | Mi | Casa | Hola | Professor | Welcome | And | y | Bienvenido |
| Idf | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 |

Tf-Idf:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Hello | Open | House | Mi | Casa | Hola | Professor | Welcome | And | y | Bienvenido |
| Doc1 | 0.78 | 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 | 0.78 | 0.78 | 0 | 0 | 0 | 0 |
| Doc5 | 0 | 0 | 0 | 0 | 0 | 0.78 | 0 | 0 | 0 | 0.78 | 0.78 |
| Doc6 | 0.78 | 0 | 0 | 0 | 0 | 0 | 0 | 0.78 | 0.78 | 0 | 0 |

Tf-idf (I2 normalized):

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Hello | Open | House | Mi | Casa | Hola | Professor | Welcome | And | y | Bienvenido |
| Doc1 | 0.708 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Doc2 | 0 | 0.708 | 0.708 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Doc3 | 0 | 0 | 0 | 0.708 | 0.708 | 0 | 0 | 0 | 0 | 0 | 0 |
| Doc4 | 0 | 0 | 0 | 0 | 0 | 0.708 | 0.708 | 0 | 0 | 0 | 0 |
| Doc5 | 0 | 0 | 0 | 0 | 0 | 0.708 | 0 | 0 | 0 | 0.708 | 0.708 |
| Doc6 | 0.708 | 0 | 0 | 0 | 0 | 0 | 0 | 0.708 | 0.708 | 0 | 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.
  + Sim(banana,monkey)=0
  + Sim(banana,orange)=0.33
  + Sim(banana,elephant)=0
  + Sim(monkey,orange)=0
  + Sim(monkey,elephant)=0.33
  + Sim(orange,elephant)=0
* Compute distance of each pair of words using euclide distance.
  + D(banana,monkey)= = 3.46
  + D(banana,orange)=2.82
  + D(banana,elephant)=3.46
  + D(monkey,orange)=3.46
  + D(monkey,elephant)=2.82
  + D(orange,elephant)=3.46
* Find the closest pairs. Justify the semantic rationality against the above vector representation.
  + Closet pairs from calculations: (banana, orange); (monkey, elephant).
  + The results is rational:
    - Monkey and elephant are both animals.
    - Orange and Banana are both fruits.