Exercise: Syntactical 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						
open						
…						
y						

- Construct the normalized tf-idf weights matrix W.

Exercise: 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.
- Compute distance of each pair of words using euclide distance.
- Find the closest pairs. Justify the semantic rationality against the above vector representation.