|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | idf |
| jack | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | log(8/3) |
| jill | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |  |
| went | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |  |
| up | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | log4 |
| hill | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |  |
| fetch | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |  |
| pail | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |  |
| water | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |  |
| fell | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |  |
| down | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |  |
| broke | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |  |
| crown | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |  |
| came | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |  |
| tumbling | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |  |
| after | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |  |
| got | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |  |
| home | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |  |
| trot | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |  |
| fast | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |  |
| he | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |  |
| could | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |  |
| caper | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |  |
| old | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |  |
| dame | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |  |
| dob | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |  |
| who | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |  |
| patched | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |  |
| nob | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |  |
| with | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |  |
| brown | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |  |
| paper | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |  |
| vinegar | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |  |

(1)

Query “Jack”

Stop Words to be taken out: and, the, to, a, of, as, his, did, he

Binary Vectors of Q and D1-D8:

Q=<1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0>

D1=<1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0>

D2=<0,0,0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0>

D3=<1,0,0,0,0,0,0,0,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0>

D4=<0,1,0,0,0,0,0,0,0,0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0>

D5=<1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0>

D6=<0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0>

D7=<0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,0,0,0,0>

D8=<0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1>

Inner Product for each document:

D1 \* Q = 1\*1 + 1\*0 + 1\*0 + 1\*0 + 1\*0 = 1  
D2 \* Q = 1\*0 + 1\*0 + 1\*0 = 0  
D3 \* Q = 1\*1 + 1\*0 + 1\*0 + 1\*0 + 1\*0 = 1  
D4 \* Q = 1\*0 + 1\*0 + 1\*0 + 1\*0 = 0  
D5 \* Q = 1\*0 + 1\*1 + 1\*0 + 1\*0 + 1\*0 = 1  
D6 \* Q = 1\*0 + 1\*0 + 1\*0 = 0  
D7 \* Q = 1\*0 + 1\*0 + 1\*0 + 1\*0 + 1\*0 + 1\*0 = 0  
D8 \* Q = 1\*0 + 1\*0 + 1\*0 + 1\*0 = 0

(2)

Cosine Similarity for each document:

D1: Q\*D1/(|Q|\*|D1|) = 1/(sqrt(1^2)\*sqrt(5\*(1^2))) = 1/sqrt5 = 0.4472

D2: Q\*D2/(|Q|\*|D2|) = 0/(sqrt(1^2)\*sqrt(3\*(1^2))) = 0/sqrt3 = 0

D3: Q\*D3/(|Q|\*|D3|) = 1/(sqrt(1^2)\*sqrt(5\*(1^2))) = 1/sqrt5 = 0.4472

D4: Q\*D4/(|Q|\*|D4|) = 0/(sqrt(1^2)\*sqrt(4\*(1^2))) = 0/sqrt4 = 0

D5: Q\*D5/(|Q|\*|D5|) = 1/(sqrt(1^2)\*sqrt(5\*(1^2))) = 1/sqrt5 = 0.4472

D6: Q\*D6/(|Q|\*|D6|) = 0/(sqrt(1^2)\*sqrt(3\*(1^2))) = 0/sqrt3 = 0

D7: Q\*D7/(|Q|\*|D7|) = 0/(sqrt(1^2)\*sqrt(6\*(1^2))) = 0/sqrt6 = 0

D8: Q\*D8/(|Q|\*|D8|) = 0/(sqrt(1^2)\*sqrt(4\*(1^2))) = 1/sqrt4 = 0

(3)

This is a special case since the query is simple vector, and also that documents 2,4,6,7 and 8 don’t have the query keyword “Jack”. Documents 1,3 and 5 and coincidently they all have 5 words (excluding stop words).

(4)

Using query “Jill” would make two algorithms result differently.

D1 and D4 would result in 1 by inner product, however by cosine similarity we have:

D1: Q\*D1/(|Q|\*|D1|) = 1/(sqrt(1^2)\*sqrt(5\*(1^2))) = 1/sqrt5 = 0.4472

D4: Q\*D4/(|Q|\*|D4|) = 1/(sqrt(1^2)\*sqrt(4\*(1^2))) = 1/sqrt5 = 0.5

From the above, we see that the cosine similarity algorithm differentiate the results of D1 and D4, as they are not equivalent.

(5)

TF-IDF for D1:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Jack | Jill | Went | Up | Hill |
| Q | 1 | 0 | 0 | 0 | 0 |
| D1 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |
| D2 | 0 | 0 | 0 | 0 | 0 |
| D3 | 0.2 | 0 | 0 | 0 | 0 |
| D4 | 0 | 0.25 | 0 | 0 | 0 |
| D5 | 0.2 | 0 | 0 | 0 | 0 |
| D6 | 0 | 0 | 0 | 0 | 0 |
| D7 | 0 | 0 | 0 | 0 | 0 |
| D8 | 0 | 0 | 0 | 0 | 0 |
| dfi | 3 | 2 | 1 | 1 | 1 |
| D/dfi | 2.6667 | 4 | 8 | 8 | 8 |
| IDF | 0.426 | 0.6021 | 0.9031 | 0.9031 | 0.9031 |

Jack: 1/5\*0.426=0.0852

Jill: 1/5\*0.6021=0.1204

Went: 1/5\*0.9031=0.1806

Up: 1/5\*0.9031=0.1806

Hill: 1/5\*0.9031=0.1806