

Information Retrieval

Natural Language Processing

University of Maryland

Evaluation

Example Adapted from Ethen Liu

Collection

```
docs = {0: "the sky is blue",  
        1: "the sun is bright today",  
        2: "the sun in the sky is bright",  
        3: "we can see the shining sun the bright sun"}
```

Build the Vocab ($V = 5$)

```
original_frequency = Counter()
for doc in docs:
    for word in docs[doc].split():
        word_frequency[word.lower()] += 1
vocab = [x for x, y in word_frequency.most_common(5)]
```

Build the Vocab ($V = 5$)

```
original_frequency = Counter()
for doc in docs:
    for word in docs[doc].split():
        word_frequency[word.lower()] += 1
vocab = [x for x, y in word_frequency.most_common(5)]
[('the', 6), ('sun', 4), ('is', 3), ('bright', 3),
 ('sky', 2)]
```

Censor the Vocab

```
['the sky is UNK',  
 'the sun is bright UNK',  
 'the sun UNK the sky is bright',  
 'UNK UNK UNK the UNK sun the bright sun']
```

Doc Frequency

How many docs did each term appear in?

Doc Frequency

How many docs did each term appear in?

Word	Document Frequency
UNK	4
bright	3
is	3
sky	2
sun	3
the	4

Term Frequency

Original Salton paper uses absolute frequency and makes vectors unit length later; let's use raw frequency immediately.

	the	sun	is	bright	sky	UNK
0						the sky is UNK
1						the sun is bright UNK
2						the sun UNK the sky is bright
3						UNK UNK UNK the UNK sun the bright sun

Term Frequency

Original Salton paper uses absolute frequency and makes vectors unit length later; let's use raw frequency immediately.

	the	sun	is	bright	sky	UNK	
0	0.25	0.00	0.25	0.00	0.25	0.25	the sky is UNK
1							the sun is bright UNK
2							the sun UNK the sky is bright
3							UNK UNK UNK the UNK sun the bright sun

Term Frequency

Original Salton paper uses absolute frequency and makes vectors unit length later; let's use raw frequency immediately.

	the	sun	is	bright	sky	UNK	
0	0.25	0.00	0.25	0.00	0.25	0.25	the sky is UNK
1	0.20	0.20	0.20	0.20	0.00	0.20	the sun is bright UNK
2							the sun UNK the sky is bright
3							UNK UNK UNK the UNK sun the bright sun

Term Frequency

Original Salton paper uses absolute frequency and makes vectors unit length later; let's use raw frequency immediately.

	the	sun	is	bright	sky	UNK	
0	0.25	0.00	0.25	0.00	0.25	0.25	the sky is UNK
1	0.20	0.20	0.20	0.20	0.00	0.20	the sun is bright UNK
2	0.29	0.14	0.14	0.14	0.14	0.14	the sun UNK the sky is bright
3							UNK UNK UNK the UNK sun the bright sun

Term Frequency

Original Salton paper uses absolute frequency and makes vectors unit length later; let's use raw frequency immediately.

	the	sun	is	bright	sky	UNK	
0	0.25	0.00	0.25	0.00	0.25	0.25	the sky is UNK
1	0.20	0.20	0.20	0.20	0.00	0.20	the sun is bright UNK
2	0.29	0.14	0.14	0.14	0.14	0.14	the sun UNK the sky is bright
3	0.22	0.22	0.00	0.11	0.00	0.44	UNK UNK UNK the UNK sun the bright sun

tf-idf

$$w_{i,j} = f_{i,j} \log\left(\frac{D}{d_i}\right) \quad (1)$$

Use log base 10

	the	sun	is	bright	sky	UNK
0	the sky is UNK					
1	the sun is bright UNK					
2	the sun UNK the sky is bright					
3	UNK UNK UNK the UNK sun the bright sun					

tf-idf

$$w_{i,j} = f_{i,j} \log\left(\frac{D}{d_i}\right) \quad (1)$$

Use log base 10

	the	sun	is	bright	sky	UNK	
0	0.00	0.00	0.12	0.00	0.30	0.00	the sky is UNK
1							the sun is bright UNK
2							the sun UNK the sky is bright
3							UNK UNK UNK the UNK sun the bright sun

tf-idf

$$w_{i,j} = f_{i,j} \log\left(\frac{D}{d_i}\right) \quad (1)$$

Use log base 10

	the	sun	is	bright	sky	UNK	
0	0.00	0.00	0.12	0.00	0.30	0.00	the sky is UNK
1	0.00	0.12	0.12	0.12	0.00	0.00	the sun is bright UNK
2							the sun UNK the sky is bright
3							UNK UNK UNK the UNK sun the bright sun

tf-idf

$$w_{i,j} = f_{i,j} \log\left(\frac{D}{d_i}\right) \quad (1)$$

Use log base 10

	the	sun	is	bright	sky	UNK	
0	0.00	0.00	0.12	0.00	0.30	0.00	the sky is UNK
1	0.00	0.12	0.12	0.12	0.00	0.00	the sun is bright UNK
2	0.00	0.12	0.12	0.12	0.30	0.00	the sun UNK the sky is bright
3							UNK UNK UNK the UNK sun the bright sun

tf-idf

$$w_{i,j} = f_{i,j} \log\left(\frac{D}{d_i}\right) \quad (1)$$

Use log base 10

	the	sun	is	bright	sky	UNK	
0	0.00	0.00	0.12	0.00	0.30	0.00	the sky is UNK
1	0.00	0.12	0.12	0.12	0.00	0.00	the sun is bright UNK
2	0.00	0.12	0.12	0.12	0.30	0.00	the sun UNK the sky is bright
3	0.00	0.25	0.00	0.12	0.00	0.00	UNK UNK UNK the UNK sun the bright sun

Query Document

bright sun ball

Working out vector:

1. term frequency
2. document frequency
3. vector

Working out vector:

1. term frequency

$$tf^{\text{bright}} = 0.33 \quad (2)$$

$$tf^{\text{sun}} = 0.33 \quad (3)$$

$$tf^{\text{UNK}} = 0.33 \quad (4)$$

(5)

2. document frequency

3. vector

Working out vector:

1. term frequency

$$tf^{\text{bright}} = 0.33 \quad (2)$$

$$tf^{\text{sun}} = 0.33 \quad (3)$$

$$tf^{\text{UNK}} = 0.33 \quad (4)$$

(5)

2. document frequency

$$df^{\text{bright}} = 3 \quad (6)$$

$$df^{\text{sun}} = 3 \quad (7)$$

$$df^{\text{UNK}} = 4 \quad (8)$$

(9)

3. vector

Working out vector:

1. term frequency
2. document frequency

$$df^{\text{bright}} = 3 \quad (2)$$

$$df^{\text{sun}} = 3 \quad (3)$$

$$df^{\text{UNK}} = 4 \quad (4)$$

(5)

3. vector

$$tf-idf^{\text{bright}} = \frac{1}{4} \log_{10} \left(\frac{4}{3} \right) = 0.12 \quad (6)$$

$$tf-idf^{\text{sun}} = \frac{1}{4} \log_{10} \left(\frac{4}{3} \right) = 0.12 \quad (7)$$

$$tf-idf^{\text{UNK}} = \frac{1}{4} \log_{10} \left(\frac{4}{4} \right) = 0.00 \quad (8)$$

(9)

Most similar document?

Use dot product $\sum_i f_i \cdot g_i$

Most similar document?

Use dot product $\sum_i f_i \cdot g_i$

$$\begin{aligned} \text{sim}(q, d_0) &= 0.00 \\ \text{sim}(q, d_1) &= 0.12 \cdot (\text{sun}) + 0.12 \cdot (\text{bright}) = 0.036 \\ &\quad (10) \end{aligned}$$

Exam-Style Question

Consider the source document:

One Fish Two Fish Red Fish Blue Fish

If you have two queries:

- blue
- fish

that have the same similarity to the source document and that “blue” ($b = 10$) and “fish” ($f = 100$) appear in the given number of documents, how many total documents are there (N)?

Solution

$$\frac{1}{8} \left[\log \left(\frac{N}{b} \right) \right]^2 = \frac{1}{2} \left[\log \left(\frac{N}{f} \right) \right]^2 \quad (11)$$

(12)

(13)

(14)

Solution

Representation is term frequency times idf. Blue appears only once in the source document (with eight words), query only has one word, so $1 \cdot \frac{1}{8}$.

$$\frac{1}{8} \left[\log \left(\frac{N}{b} \right) \right]^2 = \frac{1}{2} \left[\log \left(\frac{N}{f} \right) \right]^2 \quad (11)$$

(12)

(13)

(14)

Solution

Fish appears $\frac{4}{8}$ times.

$$\frac{1}{8} \left[\log \left(\frac{N}{b} \right) \right]^2 = \frac{1}{2} \left[\log \left(\frac{N}{f} \right) \right]^2 \quad (11)$$

(12)

(13)

(14)

Solution

The idf for both the query and the source are $\log \frac{N}{\# \text{ docs with type}}$, but it is in both the query and the source, so the idf is squared.

$$\frac{1}{8} \left[\log \left(\frac{N}{b} \right) \right]^2 = \frac{1}{2} \left[\log \left(\frac{N}{f} \right) \right]^2 \quad (11)$$

(12)

(13)

(14)

Solution

Multiply both sides by 8 and take the square root.

$$\frac{1}{8} \left[\log \left(\frac{N}{b} \right) \right]^2 = \frac{1}{2} \left[\log \left(\frac{N}{f} \right) \right]^2 \quad (11)$$

$$\log \left(\frac{N}{b} \right) = 2 \log \left(\frac{N}{f} \right) \quad (12)$$

$$(13)$$

$$(14)$$

Solution

Bring exponent inside

$$\frac{1}{8} \left[\log \left(\frac{N}{b} \right) \right]^2 = \frac{1}{2} \left[\log \left(\frac{N}{f} \right) \right]^2 \quad (11)$$

$$\log \left(\frac{N}{b} \right) = 2 \log \left(\frac{N}{f} \right) \quad (12)$$

$$\log \left(\frac{N}{b} \right) = \log \left(\frac{N^2}{f^2} \right) \quad (13)$$

$$(14)$$

Solution

Exponentiate both sides, solve for N

$$\frac{1}{8} \left[\log \left(\frac{N}{b} \right) \right]^2 = \frac{1}{2} \left[\log \left(\frac{N}{f} \right) \right]^2 \quad (11)$$

$$\log \left(\frac{N}{b} \right) = 2 \log \left(\frac{N}{f} \right) \quad (12)$$

$$\log \left(\frac{N}{b} \right) = \log \left(\frac{N^2}{f^2} \right) \quad (13)$$

$$N = \frac{f^2}{b} \quad (14)$$

Solution

Put in values

$$\frac{1}{8} \left[\log \left(\frac{N}{b} \right) \right]^2 = \frac{1}{2} \left[\log \left(\frac{N}{f} \right) \right]^2 \quad (11)$$

$$\log \left(\frac{N}{b} \right) = 2 \log \left(\frac{N}{f} \right) \quad (12)$$

$$\log \left(\frac{N}{b} \right) = \log \left(\frac{N^2}{f^2} \right) \quad (13)$$

$$N = \frac{f^2}{b} = \frac{100 \cdot 100}{10} = 1000 \quad (14)$$