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	2

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	0.00	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	0.00	0.00	0.25	0.00	0.25	0.25	The sky is UNK
1	0.00	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	0.00	0.00	0.25	0.00	0.25	0.25	The sky is UNK
1	0.00	0.20	0.20	0.20	0.00	0.20	The sun is bright UNK
2	0.14	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	0.00	0.00	0.25	0.00	0.25	0.25	The sky is UNK
1	0.00	0.20	0.20	0.20	0.00	0.20	The sun is bright UNK
2	0.14	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

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

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

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

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

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

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

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

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	0.30	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

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

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	0.30	0.12	0.12	0.12	0.30	0.00	The sun UNK the sky is bright
3	0.60	0.25	0.00	0.12	0.00	0.00	UNK UNK UNK the UNK sun the bright sun

Query Document

The shining sky ball

Don’t use UNK token—just make unknown zero (but will in HW)

Working out vector:

1. term frequency
2. document frequency
3. vector

Working out vector:

1. term frequency

$$\text{tf}^{\text{UNK}} = 0.50 \tag{2}$$

$$\text{tf}^{\text{the}} = 0.25 \tag{3}$$

$$\text{tf}^{\text{sky}} = 0.25 \tag{4}$$

$$\tag{5}$$

2. document frequency
3. vector

Working out vector:

1. term frequency

$$\text{tf}^{\text{UNK}} = 0.50 \tag{2}$$

$$\text{tf}^{\text{the}} = 0.25 \tag{3}$$

$$\text{tf}^{\text{sky}} = 0.25 \tag{4}$$

$$\tag{5}$$

2. document frequency

$$\text{df}^{\text{UNK}} = 4 \tag{6}$$

$$\text{df}^{\text{the}} = 2 \tag{7}$$

$$\text{df}^{\text{sky}} = 2 \tag{8}$$

$$\tag{9}$$

3. vector

Working out vector:

1. term frequency
2. document frequency

$$\text{df}^{\text{UNK}} = 4 \tag{2}$$

$$\text{df}^{\text{the}} = 2 \tag{3}$$

$$\text{df}^{\text{sky}} = 2 \tag{4}$$

$$\tag{5}$$

3. vector

$$\text{tf-idf}^{\text{UNK}} = \frac{2}{4} \log_{10}\left(\frac{4}{4}\right) = 0.00 \tag{6}$$

$$\text{tf-idf}^{\text{the}} = \frac{1}{4} \log_{10}\left(\frac{4}{2}\right) = 0.30 \tag{7}$$

$$\text{tf-idf}^{\text{sky}} = \frac{1}{4} \log_{10}\left(\frac{4}{2}\right) = 0.30 \tag{8}$$

$$\tag{9}$$

Most similar document?

Use dot product $\sum_i f_i \cdot g_i$

Most similar document?

Use dot product $\sum_i f_i \cdot g_i$

$$0.30 .^{(\text{sky})} 0.30 = 0.09 \quad = 0.000.30 .^{(\text{the})} 0.60 + 0.30 .^{(\text{sky})} 0.30 = 0.270.60 .^{(\text{t}} \tag{10}$$

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

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

(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 $1 \cdot \frac{1}{8}$.

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

(12)

(13)

(14)

Solution

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

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

(12)

(13)

(14)

Solution

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

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

(12)

(13)

(14)

Solution

Multiply both sides by 8 and take the square root.

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

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

(13)

(14)

Solution

Bring exponent inside

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

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

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

$$\tag{14}$$

Solution

Exponentiate both sides, solve for N

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

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

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

$$N = \frac{f^2}{b} \tag{14}$$

Solution

Put in values

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

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

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

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