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"}
```

Doc Frequency

How many docs did each term appear in?

Doc Frequency

How many docs did each term appear in?

Doc	Frequency
blue	1.000000
bright	3.000000
can	1.000000
in	1.000000
is	3.000000
see	1.000000
shining	1.000000
sky	2.000000
sun	3.000000
the	4.000000
today	1.000000
we	1.000000

Term Frequency

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

Term Frequency

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

blue	0.25	0.00	0.00	0.00
bright	0.00	0.20	0.14	0.11
can	0.00	0.00	0.00	0.11
in	0.00	0.00	0.14	0.00
is	0.25	0.20	0.14	0.00
see	0.00	0.00	0.00	0.11
shining	0.00	0.00	0.00	0.11
sky	0.25	0.00	0.14	0.00
sun	0.00	0.20	0.14	0.22
the	0.25	0.20	0.29	0.22
today	0.00	0.20	0.00	0.00
we	0.00	0.00	0.00	0.11

tf-idf

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

Use log base 10

tf-idf

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

Use log base 10

sky	0.08	0.00	0.04	0.00
sun	0.00	0.02	0.02	0.03
can	0.00	0.00	0.00	0.07
bright	0.00	0.02	0.02	0.01
blue	0.15	0.00	0.00	0.00
shining	0.00	0.00	0.00	0.07
see	0.00	0.00	0.00	0.07
we	0.00	0.00	0.00	0.07
is	0.03	0.02	0.02	0.00
in	0.00	0.00	0.09	0.00
the	0.00	0.00	0.00	0.00
today	0.00	0.12	0.00	0.00

tf-idf

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

Use log base 10

sky	0.08	0.00	0.04	0.00
sun	0.00	0.02	0.02	0.03
can	0.00	0.00	0.00	0.07
bright	0.00	0.02	0.02	0.01
blue	0.15	0.00	0.00	0.00
shining	0.00	0.00	0.00	0.07
see	0.00	0.00	0.00	0.07
we	0.00	0.00	0.00	0.07
is	0.03	0.02	0.02	0.00
in	0.00	0.00	0.09	0.00
the	0.00	0.00	0.00	0.00
today	0.00	0.12	0.00	0.00

tf-idf

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

Use log base 10

sky	0.08	0.00	0.04	0.00
sun	0.00	0.02	0.02	0.03
can	0.00	0.00	0.00	0.07
bright	0.00	0.02	0.02	0.01
blue	0.15	0.00	0.00	0.00
shining	0.00	0.00	0.00	0.07
see	0.00	0.00	0.00	0.07
we	0.00	0.00	0.00	0.07
is	0.03	0.02	0.02	0.00
in	0.00	0.00	0.09	0.00
the	0.00	0.00	0.00	0.00
today	0.00	0.12	0.00	0.00

tf-idf

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

Use log base 10

sky	0.08	0.00	0.04	0.00
sun	0.00	0.02	0.02	0.03
can	0.00	0.00	0.00	0.07
bright	0.00	0.02	0.02	0.01
blue	0.15	0.00	0.00	0.00
shining	0.00	0.00	0.00	0.07
see	0.00	0.00	0.00	0.07
we	0.00	0.00	0.00	0.07
is	0.03	0.02	0.02	0.00
in	0.00	0.00	0.09	0.00
the	0.00	0.00	0.00	0.00
today	0.00	0.12	0.00	0.00

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

$$\text{tf}^{\text{the}} = 0.33 \tag{2}$$

$$\text{tf}^{\text{shining}} = 0.33 \tag{3}$$

$$\text{tf}^{\text{sky}} = 0.33 \tag{4}$$

2. document frequency
3. vector

Working out vector:

1. term frequency

$$\text{tf}^{\text{the}} = 0.33 \tag{2}$$

$$\text{tf}^{\text{shining}} = 0.33 \tag{3}$$

$$\text{tf}^{\text{sky}} = 0.33 \tag{4}$$

2. document frequency

$$\text{df}^{\text{the}} = 4.00 \tag{5}$$

$$\text{df}^{\text{shining}} = 1.00 \tag{6}$$

$$\text{df}^{\text{sky}} = 2.00 \tag{7}$$

3. vector

Working out vector:

1. term frequency
2. document frequency

$$\text{df}^{\text{the}} = 4.00 \tag{2}$$

$$\text{df}^{\text{shining}} = 1.00 \tag{3}$$

$$\text{df}^{\text{sky}} = 2.00 \tag{4}$$

3. vector

$$\text{tf-idf}^{\text{the}} = \frac{1}{3} \log_1 0 \left(\frac{4}{4.00} \right) = 0.000000 \tag{5}$$

$$\text{tf-idf}^{\text{shining}} = \frac{1}{3} \log_1 0 \left(\frac{4}{1.00} \right) = 0.200486 \tag{6}$$

$$\text{tf-idf}^{\text{sky}} = \frac{1}{3} \log_1 0 \left(\frac{4}{2.00} \right) = 0.100243 \tag{7}$$

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 The sky is blue 0.008
1 The sun is bright today 0.0
2 The sun in the sky is bright 0.004
3 We can see the shining sun the bright sun 0.013
```

What we left out!

- UNK token
- Making vectors unit length
- Efficient computation

Exam-Style Question

Consider the source document (edited so it would have ten words):

we hold these truths self evident that all are equal

If you have two queries:

- `self sealing truths`
- `equal`

that have the same similarity to the source document and that self (1000), truths (500), and equal (50) appear in the given number of documents, how many total documents are there?