Information Retrieval

NLP: Jordan Boyd-Graber

University of Maryland

Exercise

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	1.00
bright	3.00
can	1.00
in	1.00
is	3.00
see	1.00
shining	1.00
sky	2.00
sun	3.00
the	4.00
today	1.00
we	1.00

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

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

tf-idf

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

Use log base 10

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

tf-idf

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

Use log base 10

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

tf-idf

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

Use log base 10

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

Query Document

The shining sky ball

Don’t use UNK token (but will in HW)

Query Document

The shining sky ball

Don’t use UNK token (but will in HW)

```
Query:
{'the': 0.0, 'shining': 0.2, 'sky': 0.1}
```

Term Frequencies

Term Frequencies

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

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

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

Document Frequencies

Document Frequencies

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

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

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

tf-idf

tf-idf

$$\text{tf-idf}^{\text{the}} = \frac{1}{3}\log_{10}\left(\frac{4}{4.00}\right) = 0.000000 \tag{8}$$

$$\text{tf-idf}^{\text{shining}} = \frac{1}{3}\log_{10}\left(\frac{4}{1.00}\right) = 0.200486 \tag{9}$$

$$\text{tf-idf}^{\text{sky}} = \frac{1}{3}\log_{10}\left(\frac{4}{2.00}\right) = 0.100243 \tag{10}$$

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
```