Karen Spärk Jones

Term frequency - Inverse document frequency

Roadmap

1. **Karen Spärk Jones**
2. Finding relevant web pages
3. Term frequency - inverse document frequency (tf-idf)
    a. Term frequency
    b. Inverse document frequency
4. Information theoretic interpretation of tf-idf
    a. Entropy
    b. Mutual information

Karen Spärk Jones

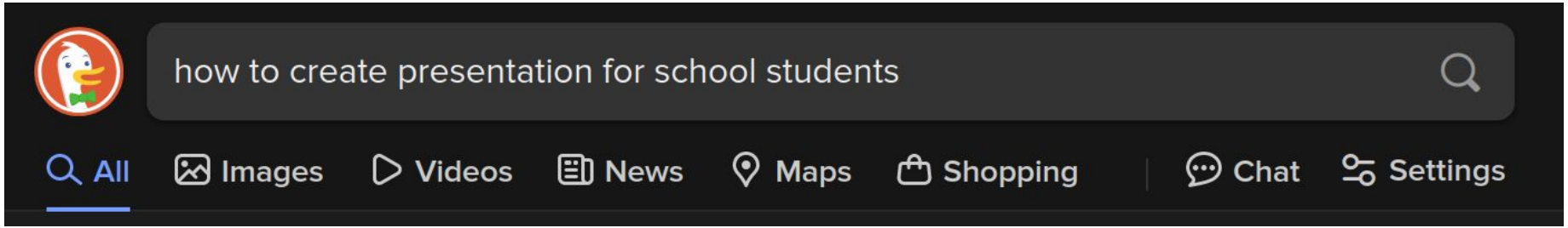
Studied history at university, and then a year in moral sciences (philosophy), worked as a teacher, and then worked at the CLRU.

“...the thing she did do was that she was willing to employ people if she thought they could do something that was worthwhile, without bothering too much about whether they’d got the right formal qualifications…””

Roadmap

1. Karen Spärk Jones
2. **Finding relevant web pages**
3. Term frequency - inverse document frequency (tf-idf)
    a. Term frequency
    b. Inverse document frequency
4. Information theoretic interpretation of tf-idf
    a. Entropy
    b. Mutual information

Finding relevant web pages



Choosing words to select the right document

She pet the dog

…

The dog is happy

…

She is happy

…

Roadmap

1. Karen Spärk Jones
2. Finding relevant web pages
3. Term frequency - inverse document frequency (tf-idf)
    a. **Term frequency**
    b. Inverse document frequency
4. Information theoretic interpretation of tf-idf
    a. Entropy
    b. Mutual information

Term frequency - inverse document frequency (tf-idf)

Is a product of two terms:

1. Term frequency
2. Inverse document frequency

Term frequency

Intuitively: measures how common a word is in a specific document

$$\mathrm{tf}(w_i, d_j) = \frac{f_{w_i,d_j}}{\sum_{w_{i'} \in d_j} f_{w_{i'},d_j}}$$

where f_{w_i,d_j} is the frequency of word w_i in document d_j

Term frequency

She pet the dog

The dog is happy

She is happy

D = { d_1, ..., d_3 }

W= { w_1, w_2, ..., w_6 } = { She, dog, pet, the, is, happy }

$$\mathrm{tf}(w_i, d_j) = \frac{f_{w_i,d_j}}{\sum_{w_{i'} \in d_j} f_{w_{i'},d_j}}$$

$$\mathrm{tf}(\text{‘dog’}, d_1) = \mathrm{tf}(w_2, d_1) = \frac{f_{w_2,d_1}}{\sum_{w_{i'} \in d_1} f_{w_{i'},d_1}} = \frac{1}{1+1+1+1} = \frac{1}{4}$$

Term frequency

Range = [0, 1]

$$\mathrm{tf}(w_i, d_j) = \frac{f_{w_i,d_j}}{\sum_{w_{i'} \in d_j} f_{w_{i'},d_j}}$$

Interpretation:

1 if it is the only distinct word that appears in the document

0 if the word does not appear in the document at all

Chicken Chicken Chicken: Chicken Chicken

Doug Zongker

University of Washington

Chicken

Chicken chicken chicken chicken chicken chicken chicken chicken chicken chicken chicken *chicken* chicken chicken chicken chicken chicken. Chicken chicken chicken chicken chicken chicken chicken. Chicken, chicken chicken chicken, chicken chicken, chicken chicken chicken “chicken chicken” chicken “chicken chicken” chicken. Chicken, chicken chicken chicken chicken chicken chicken (chicken chicken) chicken chicken chicken chicken chicken, chicken chicken chicken chicken chicken chicken chicken chicken chicken chicken chicken chicken chicken. Chicken chicken chicken chicken chicken chicken chicken, chicken chicken chicken, chicken chicken chicken chicken chicken chicken chicken.

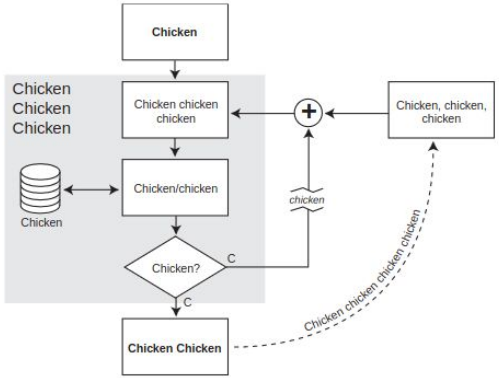
CC Chickens: C.3.2 [Chickens]: Chicken Chickens—chicken/chicken chicken; C.3.4 [Chicken chicken]: Chicken chicken chicken—chickens; C.2.4 [Chicken-chicken chickens]: Chicken/Chicken, chicken chickens

Chickens: chicken, chicken chicken, chicken, chicken

1 Chicken

Chicken chicken, chicken, chicken chicken chicken chicken, chicken chicken chicken chicken chicken chicken [4]. Chicken chicken chicken chicken, chicken chicken chicken: chicken chicken chicken chicken chicken chicken chicken chicken chicken. (chicken chicken chicken chicken chicken chicken, chicken chicken chicken chicken chicken chicken chicken!)

Chicken, chicken-chicken chicken chicken—chicken chicken, chicken chicken 95% chicken chicken-chicken chicken, chicken chicken chicken chicken—chicken chicken chicken chicken. Chicken, chicken chicken, chicken chicken chicken 1987. Chicken chicken



Chicken 1 *Chicken chicken chicken. Chicken chicken, chicken chicken (chicken chicken chicken) chicken chicken-chicken.*

Chicken chicken chicken chicken chicken. Chicken-chicken chicken chicken chicken chicken chicken, chicken chicken chicken chicken, chicken, chicken chicken chicken “chicken” chicken. Chicken chicken chicken chicken chicken chicken chicken, chicken, chicken chicken. Chicken chicken, chicken chicken chicken chicken chicken, chicken chicken chicken chicken chicken chicken (chicken, chicken-chicken chicken chicken chicken-chicken chicken chicken-chicken chicken). Chicken, chicken chicken chicken chicken chicken chicken chicken chicken chicken chicken. Chicken chicken chicken chicken chicken chicken, chicken chicken chicken chicken chicken chicken chicken chicken (chicken). Chicken chicken chicken chicken chicken chicken chicken chicken chicken chicken chicken chicken chicken, chicken chicken, chicken chicken chicken chicken.

Roadmap

1. Karen Spärk Jones
2. Finding relevant web pages
3. Term frequency - inverse document frequency (tf-idf)
    a. Term frequency
    b. Inverse document frequency
4. Information theoretic interpretation of tf-idf
    a. Entropy
    b. Mutual information

Inverse document frequency

Intuitively: measures how common a word is across all documents

Given a collection of documents D, and a word w_i

$$\text{idf}(w_i, D) = \log\left(\frac{|D|}{|\{d \in D \text{ where } w_i \in d\}|}\right)$$

Inverse document frequency

She pet the dog

The dog is happy

She is happy

D = { d_1, ..., d_3 }

W= { w_1, w_2, ..., w_6 } = { She, dog, pet, the, is, happy }

$$\text{idf}(w_i, D) = \log\left(\frac{|D|}{|\{d \in D \text{ where } w_i \in d\}|}\right)$$

$$\text{idf}(\text{‘dog’}, D) = \text{idf}(w_2, D) = \log\left(\frac{|D|}{|\{d \in D \text{ where } w_2 \in d\}|}\right) = \log\frac{3}{2}$$

Inverse document frequency

Range of inner expression = [1, |D|]

Range of log = [0, log |D|]

$$\operatorname{idf}(w_i, D) = \log\left(\frac{|D|}{|\{d \in D \text{ where } w_i \in d\}|}\right)$$

Interpretation of the range:

idf ~= log|D| if the word does not appear often (rare) in the set of documents D

Idf ~= 0 if the word appears in all documents (common) in the set of documents D

Roadmap

1. Karen Spärk Jones
2. Finding relevant web pages
3. **Term frequency - inverse document frequency (tf-idf)**
    a. Term frequency
    b. Inverse document frequency
4. Information theoretic interpretation of tf-idf
    a. Entropy
    b. Mutual information

Term frequency–inverse document frequency (tf-idf)

Intuitively: How rare a word is in a collection of documents but specific to one particular doc

Given a word w_i, a document d_j and a set of documents D.

$$\text{tf-idf}(w_i, d_j, D) = \text{tf}(w_i, d_j) \cdot \text{idf}(w_i, D)$$

Combines both term frequency and inverse document frequency

Term frequency–inverse document frequency (tf-idf)

Range

Range of tf(t, d) = [0,1]

~0 is rare, ~1 means word is common in that document

Range of idf(t, d, D) = [0, log|D|]

~log|D| is if the word is rare, and ~0 if its common

Range of tf-idf(t, d, D) = [0, log|D|]

Where high tf-idf means the word is common in a particular doc (high tf) but rare in the set of docs D (high idf).

Term frequency–inverse document frequency (tf-idf)

She pet the dog

The dog is happy

She is happy

D = { d_1, ..., d_3 }

W= { w_1, w_2, ..., w_6 } = { She, dog, pet, the, is, happy }

$$\text{tf-idf}(w_i, d_j, D) = \text{tf}(w_i, d_j) \cdot \text{idf}(w_i, D)$$

$$\text{tf-idf}(\text{‘she’}, d_1, D) = \frac{1}{4} \log \frac{3}{2} = 0.101$$

$$\text{tf-idf}(\text{‘she’}, d_2, D) = \frac{0}{4} \log \frac{3}{2} = 0$$

$$\text{tf-idf}(\text{‘she’}, d_1, D) = \frac{1}{3} \log \frac{3}{2} = 0.135$$

Roadmap

1. Karen Spärk Jones
2. Finding relevant web pages
3. Term frequency - inverse document frequency (tf-idf)
    a. Term frequency
    b. Inverse document frequency
4. Information theoretic interpretation of tf-idf
    a. Entropy
    b. Mutual information

Information theoretic interpretation

Based on “An information-theoretic perspective of tf–idf measures” [Aizawa, 2003]

Self-information of an event x of random variable X

Intuition:

If an event is certain (probability=1), then it contains no information

If an event is unlikely (low probability), then it contains more information

$$\mathrm{I}_X(x) := -\log\left[p_X(x)\right] = \log\left(\frac{1}{p_X(x)}\right).$$

Self-information of an event x of random variable X

Example:

Let X be 1 if coin is heads, and 0 if tails

Let x=1 i.e the event that the coin is heads

$$\mathrm{I}_X(x) := -\log\left[p_X(x)\right] = \log\left(\frac{1}{p_X(x)}\right)$$

If the coin is fair, then I(x) = log 2 = 1

If the probability of the coin being heads is 0.25, then I(x) = log 4 = 2

Entropy of a random variable X

Intuitively: How certain are we about the outcome of X

a fair coin is more uncertain than an unfair coin.

Maximum value is 1 when things are uncertain

$$\mathrm{H}(X) := -\sum_{x \in \mathcal{X}} p_X(x) \log p_X(x) = \sum_{x \in \mathcal{X}} p_X(x) \log \frac{1}{p_X(x)}$$

Entropy of a random variable X

Fair coin example:

$$\mathrm{H}(X) := -\sum_{x \in \mathcal{X}} p_X(x) \log p_X(x) = -(0.5 \log 0.5 + 0.5 \log 0.5) = -\log 0.5 = 1$$

Unfair coin example:

$$\mathrm{H}(X) := -\sum_{x \in \mathcal{X}} p_X(x) \log p_X(x) = -(0.7 \log 0.7 + 0.3 \log 0.3) \ = 0.88$$

Entropy of a random variable X given another random variable Y takes value y

$$\mathrm{H}(X) := -\sum_{x \in \mathcal{X}} p_X(x) \log p_X(x)$$

$$\mathrm{H}(X|Y=y) = -\sum_{x \in \mathcal{X}} P_{X|Y}(x|y) \log P_{X|Y}(x|y)$$

Entropy of a random variable X given another random variable Y takes value y

$$\mathrm{H}(X|Y=y) = -\sum_{x\in\mathcal{X}} P_{X|Y}(x|y)\log P_{X|Y}(x|y)$$

Let Y be 1 if the dice roll is even, and 0 if it is odd
Let X be 1 if the dice roll is 2, and 0 otherwise

$$\mathrm{H}(X|Y=1) = -\frac{1}{3}\log\frac{1}{3} - \frac{1}{3}\log\frac{1}{3} = 0.366$$

Let X be 1 if the dice roll is 2 **or 4**, and 0 otherwise

$$\mathrm{H}(X|Y=1) = -\frac{2}{3}\log\frac{2}{3} - \frac{2}{3}\log\frac{2}{3} = 0.27$$

Conditional entropy of random variable X given random variable Y

$$\mathrm{H}(X|Y=y) = -\sum_{x\in\mathcal{X}} P_{X|Y}(x|y)\log P_{X|Y}(x|y)$$

$$\mathrm{H}(X|Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_Y(y)}$$

Roadmap

1. Karen Spärk Jones
2. Finding relevant web pages
3. Term frequency - inverse document frequency (tf-idf)
    a. Term frequency
    b. Inverse document frequency
4. Information theoretic interpretation of tf-idf
    a. Entropy
    b. Mutual information

Mutual information between two random variables

Given two random variables X and Y

$$I(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} P_{(X,Y)}(x,y) \log \left(\frac{P_{(X,Y)}(x,y)}{P_X(x)\, P_Y(y)} \right)$$

Aka information gain: how much knowing about X tells us about Y (and vice versa, as it is symmetric). Is 0 when X and Y are independent.

Roadmap

1. Karen Spärk Jones
2. Finding relevant web pages
3. Term frequency - inverse document frequency (tf-idf)
    a. Term frequency
    b. Inverse document frequency
4. **Information theoretic interpretation of tf-idf**
    a. Entropy
    b. Mutual information

Tf-idf as the portion of mutual information contributed by the specific word w_i to identify a document d_j in a given set of documents D

Tf-idf as the portion of mutual information contributed by the specific word w_i to identify a document d_j in a given set of documents D

- $D = \{ d_1, \ldots, d_N \}$ be the set of N documents
- $W = \{ w_1, \ldots, w_M \}$ be the set of M distinct words in D

Let random variable $\mathcal{D}$ and $\mathcal{W}$ take values in D and W respectively.
Let the probability of picking any particular document d_j be equally likely:

$$P(d_j) = \frac{1}{N} \quad \forall d_j \in D$$

Information theoretic interpretation of tf-idf

The entropy of $\mathcal{D}$ is given by:

$$H(\mathcal{D}) = -\sum_{d_j \in D} P(d_j) \log P(d_j) = -N \left(\frac{1}{N} \log \frac{1}{N} \right) = -\log \frac{1}{N}$$

The entropy of $\mathcal{D}$ given some word w_i is given by:

$$H(\mathcal{D}|w_i) = -\sum_{d_j \in D} P(d_j \mid w_i) \log P(d_j \mid w_i) = -N_i \left(\frac{1}{N_i} \log \frac{1}{N_i} \right) = -\log \frac{1}{N_i}$$

Information theoretic interpretation of tf-idf

We can rewrite the mutual information between $\mathcal{D}$ and $\mathcal{W}$ as:

$$I(\mathcal{D};\mathcal{W}) = \mathrm{H}(\mathcal{D}) - \mathrm{H}(\mathcal{D}|\mathcal{W}) = \sum_{w_i \in \mathcal{W}} P(w_i)(\mathrm{H}(\mathcal{D}) - \mathrm{H}(\mathcal{D}|w_i))$$

$$= \sum_{w_i \in \mathcal{W}} P(w_i) \left(-\log \frac{1}{N} + \log \frac{1}{N_i} \right) = \sum_{w_i \in \mathcal{W}} P(w_i) \log \frac{N}{N_i}$$

$$= \sum_{w_i \in \mathcal{W}} \sum_{d_j \in \mathcal{D}} P(w_i|d_j) P(d_j) \log \frac{N}{N_i} = \sum_{w_i \in \mathcal{W}} \sum_{d_j \in \mathcal{D}} \frac{f_{ij}}{f_{\_j}} P(d_j) \log \frac{N}{N_i}$$

$$= \frac{1}{N} \sum_{w_i \in \mathcal{W}} \sum_{d_j \in \mathcal{D}} \frac{f_{ij}}{f_{\_j}} \log \frac{N}{N_i}$$

N is the total number of documents

N_i is the number of documents which contain word w_i

$f_{\_j}$ is the frequency of all distinct words w_i in some document d_j

f_{ij} is the frequency of word w_i in some document d_j

Information theoretic interpretation of tf-idf

We can rewrite the mutual information between $\mathcal{D}$ and $\mathcal{W}$ as:

$$\mathrm{tf}(w_i, d_j) = \frac{f_{w_i,d_j}}{\sum_{w_{i'} \in d_j} f_{w_{i'},d_j}}$$

$$I(\mathcal{D};\mathcal{W}) = \frac{1}{N} \sum_{w_i \in \mathcal{W}} \sum_{d_j \in \mathcal{D}} \frac{f_{ij}}{f_{\_j}} \log \frac{N}{N_i}$$

$$\mathrm{idf}(w_i, D) = \log \left(\frac{|D|}{|\{d \in D \text{ where } w_i \in d\}|} \right)$$

$$= \frac{1}{N} \sum_{w_i \in \mathcal{W}} \sum_{d_j \in \mathcal{D}} \mathrm{tf}(i,j) \cdot \mathrm{idf}(i, D)$$

N is the total number of documents

N_i is the number of documents which contain word w_i

$f_{\_j}$ is the frequency of all distinct words w_i in some document d_j

f_{ij} is the frequency of word w_i in some document d_j

Summary

1. Karen Spärk Jones
2. Finding relevant web pages
3. Term frequency - inverse document frequency (tf-idf)
    a. Term frequency
    b. Inverse document frequency
4. Information theoretic interpretation of tf-idf
    a. Entropy
    b. Mutual information