Factory Lecture

TF-IDF and Cosine Similarity

Incheon Paik

Contents

- **Term Weighting**
- **Term Frequency (TF)**
- **IDF (Inverse Document Frequency)**
- **TF-IDF Term Weight**

Term Frequency

- Consider the number of occurrences of a term in a document
  - Bag of words model
  - Document is a vector in N a column below
- Let's consider the following document set.

D1	Today weather is sunny and cloudy. Rainy and cloudy tomorrow
D2	The soccer game is interesting. I like basketball game.
D3	Yesterday weather was cloudy and sunny. I like sunny day.
D4	The baseball game is not interesting. I win the tennis game.

Calculating Term Frequency (TF)

D1 Today weather is sunny and cloudy. Rainy and cloudy tomorrow

D2 The soccer game is interesting. I like basketball game.

D3 Yesterday weather was cloudy and sunny. I like sunny day.

D4 The baseball game is not interesting. I win the tennis game.

Term Frequency

Words / **Doc #**	Today	weather	is	sunny	and	Cloudy	rainy	tomorrow	The	soccer	game	interesting	I	like	basketball	Yesterday	day	baseball	not	win	tennis
D1	1	1	1	1	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
D2	0	0	1	0	0	0	0	0	1	1	2	1	1	1	1	0	0	0	0	0	0
D3	0	1	1	2	1	1	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0
D4	0	0	1	0	0	0	0	0	1	0	2	1	1	0	0	0	0	1	1	1	1

Problem of Term Frequency

- Which of these tells you more about a medical document?
  - 10 occurrences of hernia(脱腸)?
  - 10 occurrences of the ?
- Why is it?
  - Is the term a common word that exists in every document or a meaningful word that give feature to the document?
- How can we get the information?
  - If a term is found in more documents, it will have less meaning for feature of the document.
  - Document Frequency

Document Frequency (DF)

D1	Today weather is sunny and cloudy. Rainy and cloudy tomorrow
D2	The soccer game is interesting. I like basketball game.
D3	Yesterday weather was cloudy and sunny. I like sunny day.
D4	The baseball game is not interesting. I win the tennis game.

Document Frequency

Words / **Doc #**	Today	weather	is	sunny	and	Cloudy	rainy	tomorrow	The	soccer	game	interesting	I	like	basketball	Yesterday	day	baseball	not	win	tennis
DF	1	2	4	2	2	2	1	1	2	1	2	2	3	2	1	1	1	1	1	1	1

- **Usually, we use Inverse Document Frequency (IDF), and it can be calculated in the form of 1/DF.**
- **But by far the most commonly used version is: IDF = log (n/DF)**

Summary : TF × IDF (or tf.idf)

- Assign a tf.idf weight to each term i in each document d

$$w_{i,d} = tf_{i,d} \times \log(n / df_i)$$

What is the wt of a term that occurs in all of the docs?

$tf_{i,d}$ = frequency of term i in document j
n = total number of documents
df_i = the number of documents that contain term i

- Increases with the number of occurrences within a doc
- Increases with the rarity of the term across the whole corpus

Calculating TF-IDF

D1 Today weather is sunny and cloudy. Rainy and cloudy tomorrow

D2 The soccer game is interesting. I like basketball game.

D3 Yesterday weather was cloudy and sunny. I like sunny day.

D4 The baseball game is not interesting. I win the tennis game.

$$w_{i,d} = tf_{i,d} \times \log(n / df_i)$$

What is the wt of a term that occurs in all of the docs?

$tf_{i,d}$ = frequency of term i in document j

n = total number of documents

df_i = the number of documents that contain term i

TF-IDF

Words / **Doc #**	Today	weather	is	sunny	and	Cloudy	rainy	tomorrow	The	soccer	game	Interesting	I	Like	basketball	Yesterday	day	baseball	not	Win	Tennis
D1	1* log(4/1)	1* log(4/2)	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D2	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D3	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D4	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Example (Calculating TF-IDF)

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

(1) キーワード抽出対象テキスト中の代表キーワード候補出現数 (TF)
(2) 全てのドキュメント数 (N)
(3) 代表キーワード候補が含まれるドキュメントの数 (DF)

Example (Calculating TF-IDF)

- Example text : “a.txt”

本棚が届きました。さっそく組み立て。しかし、一部の部品に不良品があり一段だけ固定できません。本棚への道は険しいです。今週中に部品交換に行ってきます。

- Morphological Analysis using “chasen”

```
% chasen a.txt|grep '名詞'|sort|uniq -c|sort -nr
  2 本棚　　ホンダナ　　本棚　名詞-一般
  2 部品　　ブヒン　部品　名詞-一般
  1 不良　　フリョウ　　不良　名詞-形容動詞語幹
  1 品 ヒン　品　　名詞-接尾-一般
  1 道 ミチ　道　　名詞-一般
  1 中 チュウ　中　　名詞-接尾-副詞可能
  1 組み立て　クミタテ　　組み立て　　名詞-一般
  1 今週　　コンシュウ　　今週　名詞-副詞可能
  1 交換　　コウカン　　交換　名詞-サ変接続
  1 固定　　コテイ　固定　名詞-サ変接続
  1 一部　　イチブ　一部　名詞-副詞可能
  1 一段　　イチダン　　一段　名詞-一般
```

Example (Calculating TF-IDF)

- **(2)の「全ドキュメント数 N」。対象となるドキュメント群は、ここでは、Yahoo! で検索できるすべての Web ページとする。Yahoo! でインデックスされているページは 192 億ページと言われているので、N = 19200000000 。**
- **(3) の DF（代表キーワード候補が含まれるドキュメントの数。対象ドキュメント群は Yahoo! で検索できる全 Web ページなので、Yahoo! 検索でのヒットした数が DF。ヒット数は Yahoo! APIで得ることができる。**

```
use LWP::Simple;
sub get_num { # 検索ヒット数獲得 by Yahoo! API
    my ($key) = @_; # UTF-8
    $key =~ s/([^0-9A-Za-z_ ])/'%'.unpack('H2',$1)/ge;
    my $url = "http://api.search.yahoo.com/WebSearchService/V1/".
        "webSearch?appid=YahooDemo&query=$key&results=1";
    my $c;
    ($c = get($url)) or die "Can't get $url¥n";
    my ($num) = ($c =~ /totalResultsAvailable="(¥d+)"/);
    return $num;
}
```

Example (Calculating TF-IDF)

- **TF-IDF の計算。**
  試しに「本棚」で計算。
  TF = 2, DF = 2771, N = 19200000000 なので、
  TFIDF ≒ 31.5 。

```
% perl -e 'print 2*log(19200000000/2771),"¥n"'
31.5024251422343
```

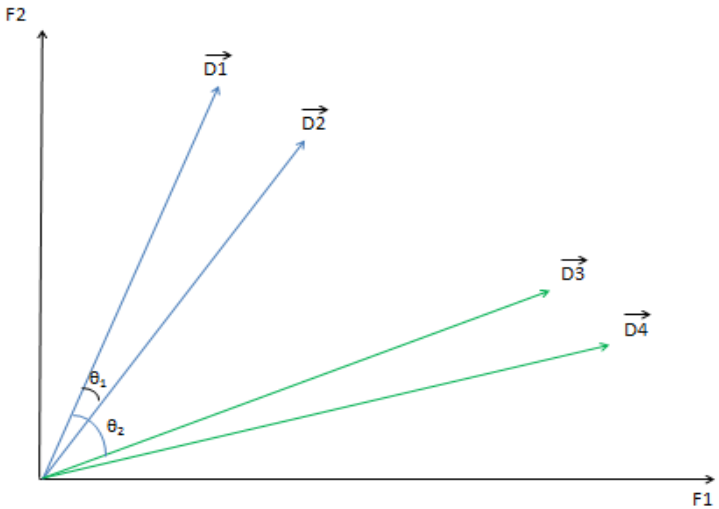
Feature Vector of A Document

D1 Today weather is sunny and cloudy. Rainy and cloudy tomorrow

D2 The soccer game is interesting. I like basketball game.

D3 Yesterday weather was cloudy and sunny. I like sunny day.

D4 The baseball game is not interesting. I win the tennis game.

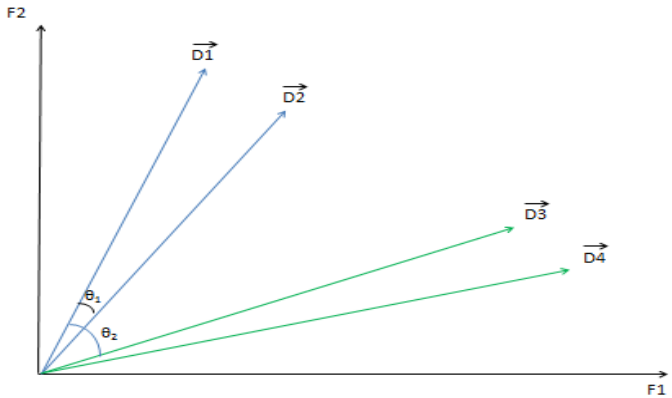


Words **Doc #**	W1: Today	W2: weather	W3: is	W4: Sunny	W5: and	W6: Cloudy	W7: rainy	W8: tomorrow	W9: The	W10: soccer	W11: game	W12: Interesting	W13: I	W14: Like	W15: basketball	W16: Yesterday	W17: day	W18: baseball	W19: not	W20: Win	W21: Tennis
D1	1* log(4/1)	1* log(4/2)	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D2	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D3	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D4	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

TF-IDF

Cosine Similarity

D1	Today weather is sunny and cloudy. Rainy and cloudy tomorrow
D2	The soccer game is interesting. I like basketball game.
D3	Yesterday weather was cloudy and sunny. I like sunny day.
D4	The baseball game is not interesting. I win the tennis game.



$$\overrightarrow{D_1} = (e_{11}, e_{12}, e_{13}, e_{14}, e_{15})$$

$$\overrightarrow{D_2} = (e_{21}, e_{22}, e_{23}, e_{24}, e_{25})$$

$$\overrightarrow{D_1} \bullet \overrightarrow{D_2} = \left|\overrightarrow{D_1}\right| \left|\overrightarrow{D_2}\right| \cos\theta$$

$$\cos\theta = \frac{\overrightarrow{D_1} \bullet \overrightarrow{D_2}}{\left|\overrightarrow{D_1}\right| \left|\overrightarrow{D_2}\right|}$$

$$= \frac{e_{11}e_{21}+e_{12}e_{22}+e_{13}e_{23}+e_{14}e_{24}+e_{15}e_{25}}{\mathrm{SQRT}(e_{11}^2+e_{12}^2+e_{13}^2+e_{14}^2+e_{15}^2) \times \mathrm{SQRT}(e_{21}^2+e_{22}^2+e_{23}^2+e_{24}^2+e_{25}^2)}$$