Machine Learning :: Text feature extraction (tf-idf) – Part I & Part I
Text similarity 직접 구현해보기!

http://blog.christianperone.com/2011/09/machine-learning-text-feature-extraction-tf-idf-part-i/
http://blog.christianperone.com/2011/10/machine-learning-text-feature-extraction-tf-idf-part-ii/

Short introduction to Vector Space Model(VSM)

Term frequency – inverse document frequency(tf-idf)는 택스트 정보를 Vectror Space Model(VSM)이나 Sparse features로 바꾸어 준다.

VSM은 텍스트 정보를 벡터로 나타내는 대수적 모델이며 Vector의 구성은 용어의 중요성이나 해당용어의 유무를 나타낸다.

Going to vector space

```
Train Document Set:

d1: The sky is blue.
d2: The sun is bright.

Test Document Set:

d3: The sun in the sky is bright.
d4: We can see the shining sun, the bright sun.
```

"The, is, at, on" 등은 다른 문서에서도 사용하는 것들이기 때문에 무시하도록 한다.

$$\mathrm{E}(t) = \begin{cases} 1, & \text{if } t \text{ is "blue"} \\ 2, & \text{if } t \text{ is "sun"} \\ 3, & \text{if } t \text{ is "bright"} \\ 4, & \text{if } t \text{ is "sky"} \end{cases}$$

Train document set의 단어들을 인덱싱 해준다.

$$\mathrm{tf}(t,d) = \sum_{x \in d} \mathrm{fr}(x,t) \;\; \mathrm{fr}(x,t) = \begin{cases} 1, & \text{if } x = t \\ 0, & \text{otherwise} \end{cases}$$

Test document set을 vector space로 바꿔준다. 각각의 벡터들은 우리가 indexing한 단어들이다.
Test document set에 우리가 indexing한 단어들이 얼마나 많이 등장하는지 확인하고 이것은 term-frequency라고 한다.

$$\vec{v_{d_n}} = (\mathrm{tf}(t_1, d_n), \mathrm{tf}(t_2, d_n), \mathrm{tf}(t_3, d_n), \ldots, \mathrm{tf}(t_n, d_n))$$

예를 들어 tf(t1,d2)는 d2 document에 t1(blue)가 들어있는 횟수이다.

$$\vec{v_{d_3}} = (\mathrm{tf}(t_1, d_3), \mathrm{tf}(t_2, d_3), \mathrm{tf}(t_3, d_3), \ldots, \mathrm{tf}(t_n, d_3)) \quad \vec{v_{d_3}} = (0, 1, 1, 1)$$
$$\vec{v_{d_4}} = (\mathrm{tf}(t_1, d_4), \mathrm{tf}(t_2, d_4), \mathrm{tf}(t_3, d_4), \ldots, \mathrm{tf}(t_n, d_4)) \quad \vec{v_{d_4}} = (0, 2, 1, 0)$$

$$M_{|D| \times F} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 2 & 1 & 0 \end{bmatrix}$$ D는 문서의 개수, F의 feature의 수

Tf-idf weight : 자주 사용하는 term에는 scale down을 하고 자주 사용하지 않는 term에는 scale up을 한다.

Term frequency같은 단순한 방법을 사용하면 Keyword spamming 문제가 발생 할 수 있다.
-정보검색 시스템에서 높은 빈도와 순위를 가져 편향이 발생한다.-
문제를 해결하기 위해서는 vectror를 정규화 시켜줘야 한다.

Cf) Back to Vector normalization

$\vec{v_{d_4}} = (0, 2, 1, 0)$ 를 정규화하자 $\hat{v} = \frac{\vec{v}}{\|\vec{v}\|_p}$

L2-norm(Euclidean norm)을 이용하여 norm(크기, 길이)을 구해준다.

$$\hat{v} = \frac{\vec{v}}{\|\vec{v}\|_p}$$
$$\hat{v_{d_4}} = \frac{\vec{v_{d_4}}}{\|\vec{v_{d_4}}\|_2}$$

$$\hat{v_{d_4}} = \frac{(0,2,1,0)}{\sqrt{0^2+2^2+1^2+0^2}}$$
$$\hat{v_{d_4}} = \frac{(0,2,1,0)}{\sqrt{5}}$$
$$\hat{v_{d_4}} = (0.0, 0.89442719, 0.4472136, 0.0)$$ Ud4를 정규화하였다.

The term frequency – inverse document frequency (tf - idf) weight

Idf(inverse document frequency)

$$\mathrm{idf}(t) = \log \frac{|D|}{1 + |\{d : t \in d\}|}$$

$|\{d : t \in d\}$(문서에서 t라는 용어가 나타날 수) $|D|$ 전체 문서의 수

$$\text{tf-idf}(t) = \text{tf}(t, d) \times \text{idf}(t)$$

Tf-idf값은 단어 빈도(term frequency)가 큰 값이고, 문서빈도가 낮을 때 높은 값을 가진다.

$$M_{train} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 2 & 1 & 0 \end{bmatrix}$$

$$\text{idf}(t_1) = \log \frac{|D|}{1+|\{d:t_1 \in d\}|} = \log \frac{2}{1} = 0.69314718$$

$$\text{idf}(t_2) = \log \frac{|D|}{1+|\{d:t_2 \in d\}|} = \log \frac{2}{3} = -0.40546511$$

$$\text{idf}(t_3) = \log \frac{|D|}{1+|\{d:t_3 \in d\}|} = \log \frac{2}{3} = -0.40546511$$

$$\text{idf}(t_4) = \log \frac{|D|}{1+|\{d:t_4 \in d\}|} = \log \frac{2}{2} = 0.0$$

$$\vec{idf_{train}} = (0.69314718, -0.40546511, -0.40546511, 0.0) \qquad M_{idf} = \begin{bmatrix} 0.69314718 & 0 & 0 & 0 \\ 0 & -0.40546511 & 0 & 0 \\ 0 & 0 & -0.40546511 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

역 문서빈도의 값을 계산을 위해서 대각행렬 값으로 바꾸어 준다.

$$M_{tf\text{-}idf} = M_{train} \times M_{idf}$$

행렬들간의 관계는 독립적이지 않기 때문에 A*B 나 B*A의 값이 다르다.

$$M_{tf\text{-}idf} = M_{train} \times M_{idf} =$$
$$\begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 2 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} 0.69314718 & 0 & 0 & 0 \\ 0 & -0.40546511 & 0 & 0 \\ 0 & 0 & -0.40546511 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$
$$= \begin{bmatrix} 0 & -0.40546511 & -0.40546511 & 0 \\ 0 & -0.81093022 & -0.40546511 & 0 \end{bmatrix}$$

TF-IDF의 값을 구해준다. 구한 값을 정규화하면 마무리한다.

$$M_{tf\text{-}idf} = \frac{M_{tf\text{-}idf}}{\|M_{tf\text{-}idf}\|_2} = \begin{bmatrix} 0 & -0.70710678 & -0.70710678 & 0 \\ 0 & -0.89442719 & -0.4472136 & 0 \end{bmatrix}$$