

---

# *IR Basics: Retrieval Models*

# 목차

- + 불리언 모델
- + 벡터공간 모델
  - 코사인 유사도
  - 질의 벡터, 문서 벡터 성분 표현
    - ◆ 이진벡터표현
    - ◆ TF 벡터 표현
    - ◆ IDF 벡터 표현
    - ◆ TF\*IDF 벡터 표현
- + 질의 문서 유사도
- + Sublinear tf scaling
- + Maximum tf normalization
- + 피봇 기반 문서 길이 정규화

## 정보검색 모델(IR model)

- ✚ 불리언 모델 (Boolean model)
- ✚ 벡터공간 모델 (Vector space model)
- ✚ 확률 모델 (Probabilistic model)
- ✚ 언어 모델 (Language model)

# 불리언 모델 (Boolean model)

## Query

- 피연산자인 용어(term)들을 불리언 연산자(AND, OR, NOT)와 결합하여 표현
- e.g.) 한국 AND 인공위성

## Retrieval

- 각 query term  $T$ 에 대해  $T$ 를 포함하는 문서집합  $S_T$ 를 대응시키고,
- 불리언 수식으로 표현된 query를 만족하는 집합  $R$ 을 찾아,
- $R$ 에 포함된 문서들을 사용자에게 제시

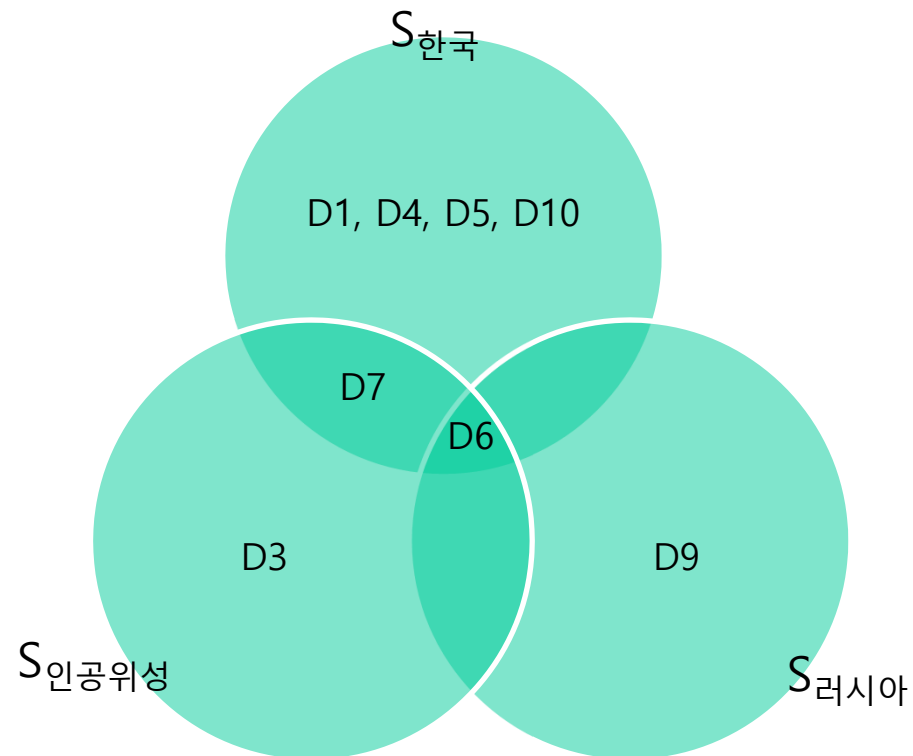
# 불리언 모델 (Boolean model)

Query = 한국 AND 인공위성 AND 러시아

검색결과 R

=  $S_{\text{한국}} \text{ AND } S_{\text{인공위성}} \text{ AND } S_{\text{러시아}}$   
=  $\{D1, D4, D5, D6, D7, D10\} \text{ AND } \{D3, D6, D7\} \text{ AND } \{D6, D9\}$   
=  $\{D6, D7\} \text{ AND } \{D6, D9\}$   
=  $\{D6\}$

D1=한국 대통령 선거로 ...  
D2=미국 금융 위기의 영향으로 ...  
D3=유럽의 인공위성 기술은 ...  
D4=한국의 국악 공연은 ...  
D5=한국에 대한 미국의 무역 적자가 ...  
D6=러시아는 한국의 인공위성 개발에 대해 ...  
D7=한국은 인공위성 발사를 위해 ...  
D8=미국 메이저리거 선수들이 ...  
D9=러시아는 중국 기업에 대한 ...  
D10=한국은 담수화 사업을 위해 ...



# 벡터공간 모델 (Vector space model)

## + 벡터공간

- 색인 용어(term)를 축(axis)에 대응시켜 얻어지는 벡터공간 정의
  - ◆ e.g.) 크기  $n$ 의 색인 용어 집합에 대해  $n$ -차원 벡터공간이 정의됨

## + Query

- 벡터공간 내의 한 벡터 (질의벡터)

## + Document

- 벡터공간 내의 한 벡터 (문서벡터)

## + Retrieval

- 각 문서벡터에 대해 질의벡터와 유사한 정도를 계산하여 유사도 (similarity) 순으로 사용자에게 제시
  - ◆ 질의-문서 유사도(query-document similarity)

# 벡터공간 모델 (Vector space model)

## 📌 질의-문서 유사도

### ● 코사인 유사도 (cosine similarity)

- ◆ 질의벡터와 문서벡터의 사잇각이 적을수록 1에 가까운 값을, 사잇각이 클수록 0에 가까운 값을 부여하는 수식

$$\text{sim}(Q, D) = \frac{Q \cdot D}{\|Q\| \|D\|} = \frac{\sum_{i=1}^N q_i \times d_i}{\sqrt{\sum_{i=1}^N q_i^2} \sqrt{\sum_{i=1}^N d_i^2}}$$

내적(inner product)  
스칼라곱(scalar product) →  
점곱(dot product)

$$Q = (1, 0, 1, 1, 0, 1)$$

$$D = (1, 1, 0, 1, 1, 0)$$

$$Q \cdot D = 1 \times 1 + 0 \times 1 + 1 \times 0 + 1 \times 1 + 0 \times 1 + 1 \times 0 = 2$$

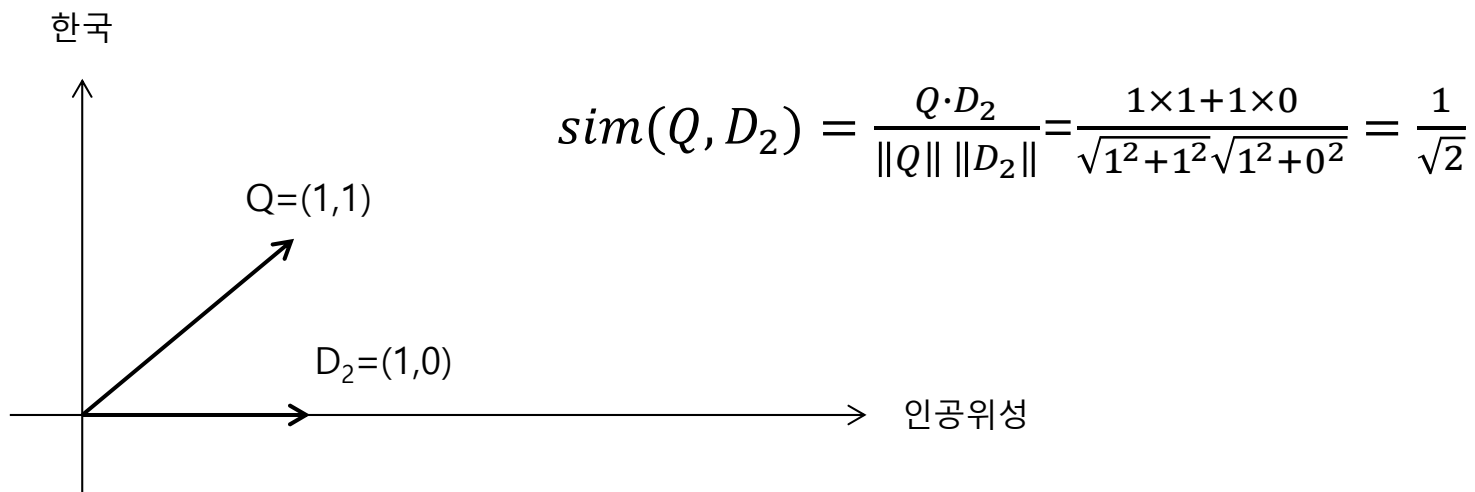
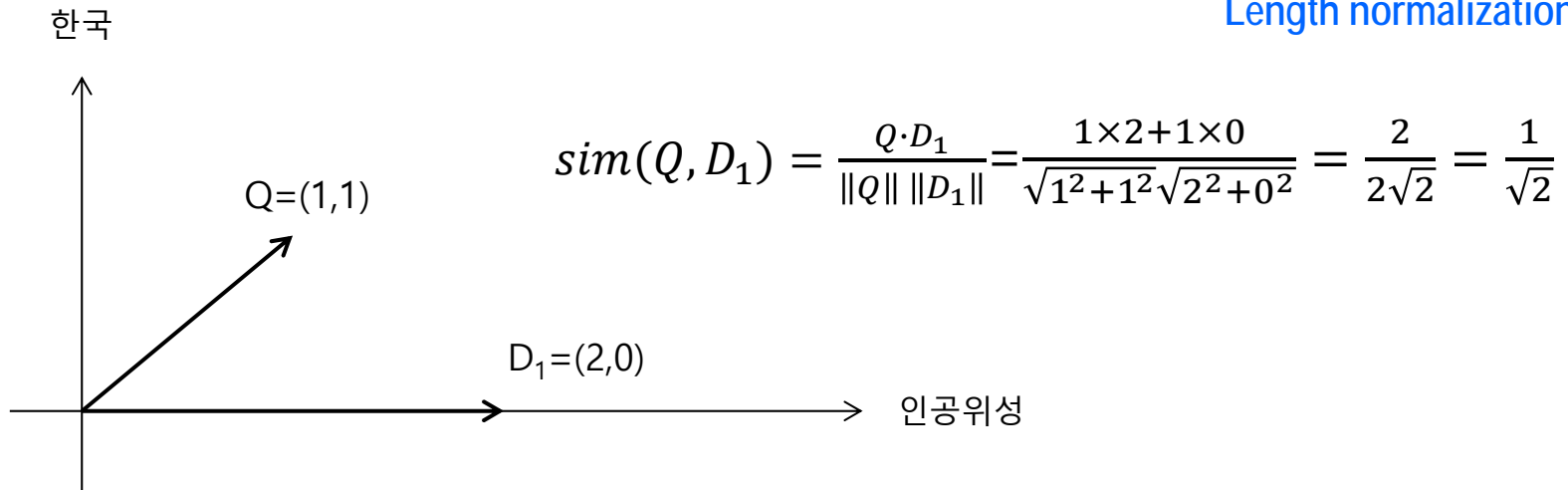
$$\|Q\| = \sqrt{1^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2} = \sqrt{4} = 2$$

$$\|D\| = \sqrt{1^2 + 1^2 + 0^2 + 1^2 + 1^2 + 0^2} = \sqrt{4} = 2$$

$$\text{sim}(Q, D) = 2 / (2 \times 2) = 0.5$$

# Query-Document distance: cosine similarity

Length normalization





# 벡터공간 모델 (Vector space model)

## ✚ 질의/문서벡터 성분 표현

### ● 이진(binary) 벡터 표현

- ◆ 용어의 질의/문서 내 출현 여부를 고려하여 출현 용어에 대응하는 벡터성분에 1을, 미출현 용어의 벡터성분에 0을 할당하는 방식
- ◆ 문서집합
  - Doc-123 = [한국, 한국, 마스크, 품질]
  - Doc-124 = [한국, 마스크, 마스크, 마스크, 부족]
  - Doc-125 = [한국, 코로나, 방역]
- ◆ 용어집합  $T = \{\text{마스크, 방역, 부족, 코로나, 품질, 한국}\}$
- ◆ 문서벡터
  - Doc-123 = (1, 0, 0, 0, 1, 1)
  - Doc-124 = (1, 0, 1, 0, 0, 1)
  - Doc-125 = (0, 1, 0, 1, 0, 1)

	마스크	방역	부족	코로나	품질	한국
Doc-123	1	0	0	0	1	1
Doc-124	1	0	1	0	0	1
Doc-125	0	1	0	1	0	1

# 벡터공간 모델 (Vector space model)

## 📌 질의/문서벡터 성분 표현

### ● TF 벡터 표현 (raw tf)

- ◆ 용어의 질의/문서 내 출현 회수(term frequency, TF)를 고려하여 용어에 대응하는 벡터성분에 해당 용어의 출현 회수를 할당하는 방식
- ◆ 문서집합
  - Doc-123 = [한국, 한국, 마스크, 품질]
  - Doc-124 = [한국, 마스크, 마스크, 마스크, 부족]
  - Doc-125 = [한국, 코로나, 방역]
- ◆ 용어집합  $T = \{\text{마스크, 방역, 부족, 코로나, 품질, 한국}\}$
- ◆ 문서벡터
  - Doc-123 = (1, 0, 0, 0, 1, 2)
  - Doc-124 = (3, 0, 1, 0, 0, 1)
  - Doc-125 = (0, 1, 0, 1, 0, 1)

	마스크	방역	부족	코로나	품질	한국
Doc-123	1	0	0	0	1	2
Doc-124	3	0	1	0	0	1
Doc-125	0	1	0	1	0	1

# 벡터공간 모델 (Vector space model)

## 🌈 질의/문서벡터 성분 표현

### ● IDF 벡터 표현

- ◆ 각 용어에 대해 문서집합에서 용어가 출현한 문서 수(document frequency, DF)에 반비례하는 값을 해당 용어에 대응하는 벡터성분에 할당하는 방식
- ◆ 문서집합 (총 색인 문서 수  $N=3$ )
  - Doc-123 = [한국, 한국, 마스크, 품질]
  - Doc-124 = [한국, 마스크, 마스크, 마스크, 부족]
  - Doc-125 = [한국, 코로나, 방역]
- ◆ 용어집합  $T = \{\text{마스크, 방역, 부족, 코로나, 품질, 한국}\}$
- ◆ 문서벡터 ( $IDF=N/DF$ 로 가정)
  - Doc-123 = (3/2, 0, 0, 0, 3/1, 3/3)
  - Doc-124 = (3/2, 0, 3/1, 0, 0, 3/3)
  - Doc-125 = (0, 3/1, 0, 3/1, 0, 3/3)

	마스크	방역	부족	코로나	품질	한국
Doc-123	3/2	0	0	0	3/1	3/3
Doc-124	3/2	0	3/1	0	0	3/3
Doc-125	0	3/1	0	3/1	0	3/3

# 벡터공간 모델 (Vector space model)

## ✚ 질의/문서벡터 성분 표현

### ● TF-IDF 벡터 표현

- ◆ 용어의 문서 내 출현 빈도인 TF와 문서집합 내 역문헌빈도인 IDF를 동시에 고려하는 벡터성분 표현
- ◆ 벡터공간모델의 대표적 용어 가중치 부여 방식

# TF\*IDF 벡터 표현

Q=[한국 위성 발사 한국]

D=[한국 위성 발사 한국 한국 위성 발사 한국]

		발사	위성	한국
DF (총 문서 수 $N=2^{10}$ )	df	$2^4$	$2^6$	$2^8$
	$N/df$	$2^{10}/2^4=2^6$	$2^{10}/2^6=2^4$	$2^{10}/2^8=2^2$
Q	tf	1	1	2
	TF-IDF	$1*2^6$	$1*2^4$	$2*2^2$
D	tf	2	2	4
	TF-IDF	$2*2^6$	$2*2^4$	$4*2^2$

Q=[ 64, 16, 8]

D=[128, 32, 16]

# TF\*IDF 벡터 표현

Q=[한국 위성 발사 한국]

D=[한국 위성 발사 한국 한국 위성 발사 한국]

		발사	위성	한국
DF (총 문서 수 $N=2^{10}$ )	df	$2^4$	$2^6$	$2^8$
	$\log(N/df)$	$\log(2^{10}/2^4)=6$	$\log(2^{10}/2^6)=4$	$\log(2^{10}/2^8)=2$
Q	tf	1	1	2
	$1+\log(tf)$	$1+\log 1=1$	$1+\log 1=1$	$1+\log 2=2$
	TF-IDF	$1*6=6$	$1*4=4$	$2*2=4$
D	tf	2	2	4
	$1+\log(tf)$	$1+\log 2=2$	$1+\log 2=2$	$1+\log 4=3$
	TF-IDF	$2*6=12$	$2*4=8$	$3*2=6$

raw tf

Q=[ 6, 4, 4]

D=[12, 8, 6]

$$\begin{aligned}
 Q \cdot D &= 6 \times 12 + 4 \times 8 + 4 \times 6 = 128 \\
 \|Q\| &= \sqrt{6^2 + 4^2 + 4^2} = \sqrt{68} = 8.25 \\
 \|D\| &= \sqrt{12^2 + 8^2 + 6^2} = \sqrt{244} = 15.62 \\
 \cos(Q, D) &= 128 / (8.25 \times 15.62) = 0.99
 \end{aligned}$$

# log(tf)

tf	$1+\log_2(\text{tf})$	차이
1	1.00	
2	2.00	1.00
3	2.58	0.58
4	3.00	0.42
5	3.32	0.32
6	3.58	0.26
7	3.81	0.22
8	4.00	0.19
9	4.17	0.17
10	4.32	0.15
11	4.46	0.14
12	4.58	0.13
13	4.70	0.12

Q=[한국,경제]

D1=[한국,한국,경제,경제,경제]

D2=[한국,경제,경제,경제,경제]

## tf 기반 유사도

$\text{sim}(Q,D1)=2+3=5$

$\text{sim}(Q,D2)=1+4=5$

## $1+\log(\text{tf})$ 기반 유사도

$\text{sim}(Q,D1)=2+2.58=4.58$

$\text{sim}(Q,D2)=1+3=4$

# 벡터공간 모델 (Vector space model)

## 질의/문서벡터 성분 표현 SMART 표기법

[modified from Figure 6.15 in (Manning et al., 2008)]

Term frequency		Document frequency		Normalization	
n (natural)	$tf(t, d)$	n (no)	1	n (none)	1
l (logarithm)	$1 + \log(tf(t, d))$	t (idf)	$\log \frac{N}{df(t)}$	c (cosine)	$\frac{1}{\sqrt{w_1^2 + w_2^2 + \dots + w_M^2}}$
a (augmented)	$0.5 + 0.5 \frac{tf(t, d)}{\max_{t'}(tf(t', d))}$			p (pivoted)	$\frac{1}{(1-s) + s \frac{dl}{avdl}}$
b (Boolean)	$\begin{cases} 1 & \text{if } tf(t, d) > 0 \\ 0 & \text{otherwise} \end{cases}$				
d (double)	$1 + \log(1 + \log(tf(t, d)))$				

↑  
용어가중치(term weighting)  
부여 기법으로 볼 수 있음

$dl \rightarrow$  문서길이  
 $avdl \rightarrow$  평균문서길이  
 $s \rightarrow$  일반적으로 0.2

**ddd.qqq**

문서벡터에서의 tf 요소      질의벡터에서의 정규화  
문서벡터에서의 df 요소      질의벡터에서의 df 요소  
문서벡터에서의 정규화      질의벡터에서의 tf 요소

**bnn.bnn**  
**Inc.Itc**



# 벡터 성분 표현

Term frequency		Document frequency		Normalization	
n (natural)	$tf(t, d)$	n (no)	1	n (none)	1
l (logarithm)	$1 + \log(tf(t, d))$	t (idf)	$\log \frac{N}{df(t)}$	c (cosine)	$\frac{1}{\sqrt{w_1^2 + w_2^2 + \dots + w_M^2}}$
a (augmented)	$0.5 + 0.5 \frac{tf(t, d)}{\max_{t'}(tf(t', d))}$			p (pivoted)	$\frac{1}{(1-s) + s \frac{dl}{avdl}}$
b (Boolean)	$\begin{cases} 1 & \text{if } tf(t, d) > 0 \\ 0 & \text{otherwise} \end{cases}$				
d (double)	$1 + \log(1 + \log(tf(t, d)))$				

전체문서집합  $C=\{D1,D2,D3,D4\}$ , 질의  $Q=[\text{한국, 마스크}]$

$D1=[\text{한국, 한국}], D2=[\text{한국, 방역, 방역}], D3=[\text{코로나, 방역}], D4=[\text{코로나}]$

bnn.bnn 기반 벡터표현

	방역	코로나	한국
D2	$1 \times 1 \times 1$	0	$1 \times 1 \times 1$
Q	0	0	$1 \times 1 \times 1$

bnn.bnn  
ddd.qqq

# 벡터 성분 표현

Term frequency		Document frequency		Normalization	
n (natural)	$tf(t, d)$	n (no)	1	n (none)	1
l (logarithm)	$1 + \log(tf(t, d))$	t (idf)	$\log \frac{N}{df(t)}$	c (cosine)	$\frac{1}{\sqrt{w_1^2 + w_2^2 + \dots + w_M^2}}$
a (augmented)	$0.5 + 0.5 \frac{tf(t, d)}{\max_{t'}(tf(t', d))}$			p (pivoted)	$\frac{1}{(1-s) + s \frac{dl}{avdl}}$
b (Boolean)	$\begin{cases} 1 & \text{if } tf(t, d) > 0 \\ 0 & \text{otherwise} \end{cases}$				
d (double)	$1 + \log(1 + \log(tf(t, d)))$				

전체문서집합  $C=\{D1,D2,D3,D4\}$ , 질의  $Q=[\text{한국, 마스크}]$

$D1=[\text{한국, 한국}], D2=[\text{한국, 방역, 방역}], D3=[\text{코로나, 방역}], D4=[\text{코로나}]$

Inn.ltn 기반 벡터표현

	방역	코로나	한국
D2	$(1 + \log(2)) \times 1 \times 1$	0	$(1 + \log(1)) \times 1 \times 1$
Q	0	0	$(1 + \log(1)) \times \log(\frac{4}{2}) \times 1$

Inn.ltn  
ddd.qqq

# 벡터 성분 표현

Term frequency		Document frequency		Normalization	
n (natural)	$tf(t, d)$	n (no)	1	n (none)	1
l (logarithm)	$1 + \log(tf(t, d))$	t (idf)	$\log \frac{N}{df(t)}$	c (cosine)	$\frac{1}{\sqrt{w_1^2 + w_2^2 + \dots + w_M^2}}$

전체문서집합  $C=\{D1,D2,D3,D4\}$ , 질의  $Q=[\text{한국, 마스크}]$   
 $D1=[\text{한국, 한국}]$ ,  $D2=[\text{한국, 방역, 방역}]$ ,  $D3=[\text{코로나, 방역}]$ ,  $D4=[\text{코로나}]$   
 Inc.ltc 기반 벡터표현

	방역	코로나	한국
D2	$(1 + \log(2)) \times 1$	0	$(1 + \log(1)) \times 1$
Q	0	0	$(1 + \log(1)) \times \log(\frac{4}{2})$

	방역	코로나	한국
D2	$(1 + \log(2)) \times 1 \times \frac{1}{\sqrt{5}}$	0	$(1 + \log(1)) \times 1 \times \frac{1}{\sqrt{5}}$
Q	0	0	$(1 + \log(1)) \times \log(\frac{4}{2}) \times \frac{1}{1}$

$$D2=[2,0,1], \|D2\|=\sqrt{2^2+0^2+1^2}=\sqrt{5}$$

$$Q=[0,0,1], \|Q\|=\sqrt{0^2+0^2+1^2}=1$$

# 질의-문서 유사도 (Query-Document Similarity)

✚ 문서 D 혹은 질의 Q에서의 용어 t의 가중치(중요도, term weight)

●  $w_{t,D} \propto \frac{TF(t,D) \times IDF(t)}{Length(D)}$

●  $w_{t,Q} \propto \frac{TF(t,Q) \times IDF(t)}{Length(Q)}$

## $Q \cap D$

- Set of **matching terms**
- Q, D를 각각 질의 용어 집합, 문서 용어 집합이라고 할 때, 질의와 문서에 공통으로 발견되는 용어들의 집합

✚ 질의-문서 유사도

●  $sim(Q, D) = \sum_{t \in Q \cap D} (w_{t,Q} \times w_{t,D})$

✚ *Inc./tc* 질의-문서 유사도

● 
$$sim(Q, D) = \frac{\sum_{t \in Q \cap D} \left( (1 + \log(tf(t, D))) \times (1 + \log(tf(t, Q))) \times \log\left(\frac{N}{df(t)}\right) \right)}{\sqrt{\sum_{t \in D} (1 + \log(tf(t, D)))^2} \times \sqrt{\sum_{t \in Q} \left( (1 + \log(tf(t, Q))) \times \log\left(\frac{N}{df(t)}\right) \right)^2}}$$

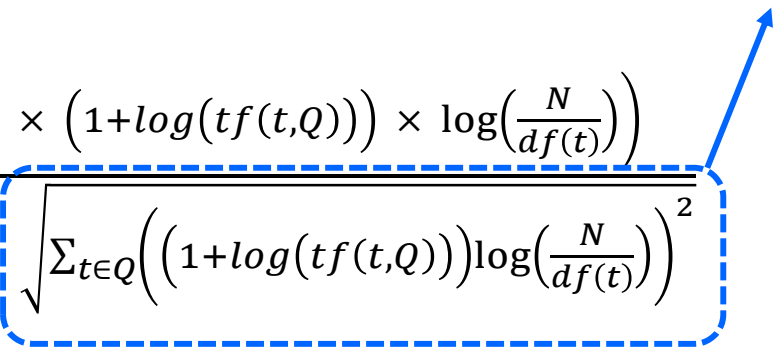
# 질의-문서 유사도 (Query-Document Similarity)

## 질의-문서 유사도

- $sim(Q, D) = \sum_{t \in Q \cap D} (w(t, Q) \times w(t, D))$

## Inc./tc 질의-문서 유사도

- $sim(Q, D) =$

$$\frac{\sum_{t \in Q \cap D} \left( (1 + \log(tf(t, D))) \times (1 + \log(tf(t, Q))) \times \log\left(\frac{N}{df(t)}\right) \right)}{\sqrt{\sum_{t \in D} (1 + \log(tf(t, D)))^2} \times \sqrt{\sum_{t \in Q} \left( (1 + \log(tf(t, Q))) \log\left(\frac{N}{df(t)}\right) \right)^2}}$$


- 서로 다른 D에 대해 상수
- 문서 유사도 순위에 영향 없음

- $sim(Q, D) \approx \frac{\sum_{t \in Q \cap D} \left( (1 + \log(tf(t, D))) \times (1 + \log(tf(t, Q))) \times \log\left(\frac{N}{df(t)}\right) \right)}{\sqrt{\sum_{t \in D} (1 + \log(tf(t, D)))^2}}$

# Sublinear tf scaling

$tf(\text{미국}, \text{doc-111})=1$   
 $tf(\text{미국}, \text{doc-222})=2$

$tf(\text{미국}, \text{doc-777})=10$   
 $tf(\text{미국}, \text{doc-888})=11$

$$tf'(t, d) = \begin{cases} 1 + \log(tf(t, d)) \\ 0 \end{cases} \quad \begin{matrix} tf(t, d) > 0 \\ otherwise \end{matrix}$$

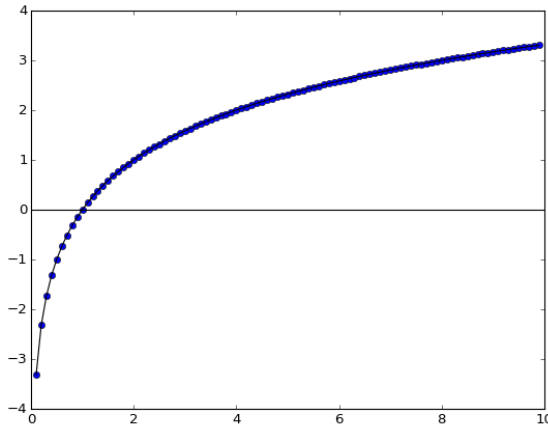


Image created by matplotlib, numpy, python

# Maximum tf normalization

Doc-123    미국, 한국, 수출, 수입

Doc-456    미국, 한국, 수출, 수입, 미국, 한국, 수출, 수입

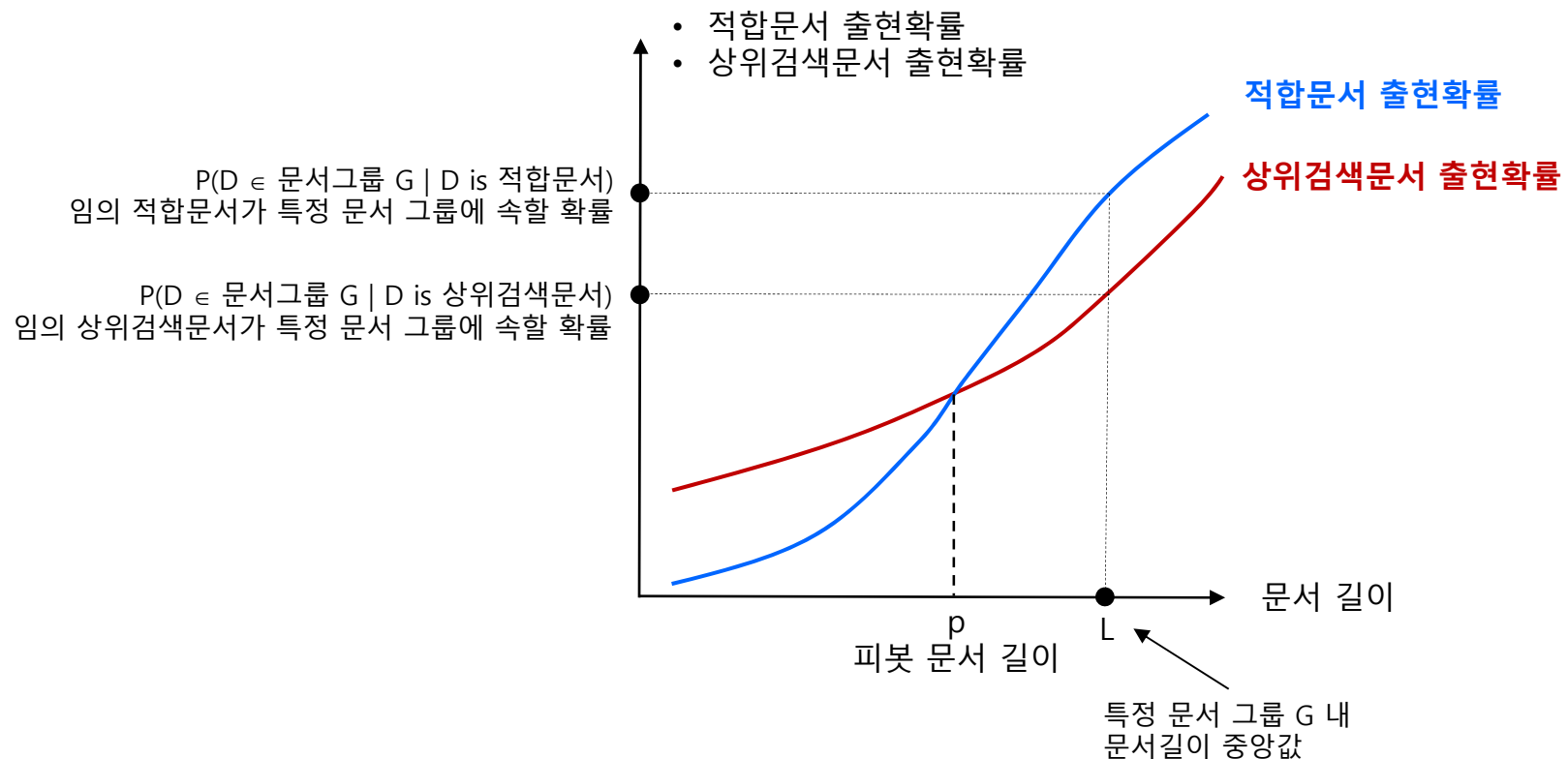
$$tf'(t, d) = \alpha + (1 - \alpha) \frac{tf(t, d)}{\max_{t \in d} tf(t, d)}$$

$\alpha$ 는 0~1 사이의 값  
(일반적으로 0.4, 초기 연구에서는 0.5 사용)

# Pivoted Document Length Normalization (피벗 기반 문서 길이 정규화)

## Pivoted Document Length Normalization (Singhal et al., 1996)

- TREC 질의 50개, 문서 741,856건, 질의-적합문서 쌍 9,805개
- 바이트 길이 오름차순 기준 전체 문서 정렬
- 정렬된 문서들을 앞에서부터 1000개 문서 씩 하나의 그룹으로 구분 (총 742개 그룹)
- 각 질의에 대해 상위 1000개 검색 문서 생성 (Inc.Itc 벡터검색적용)

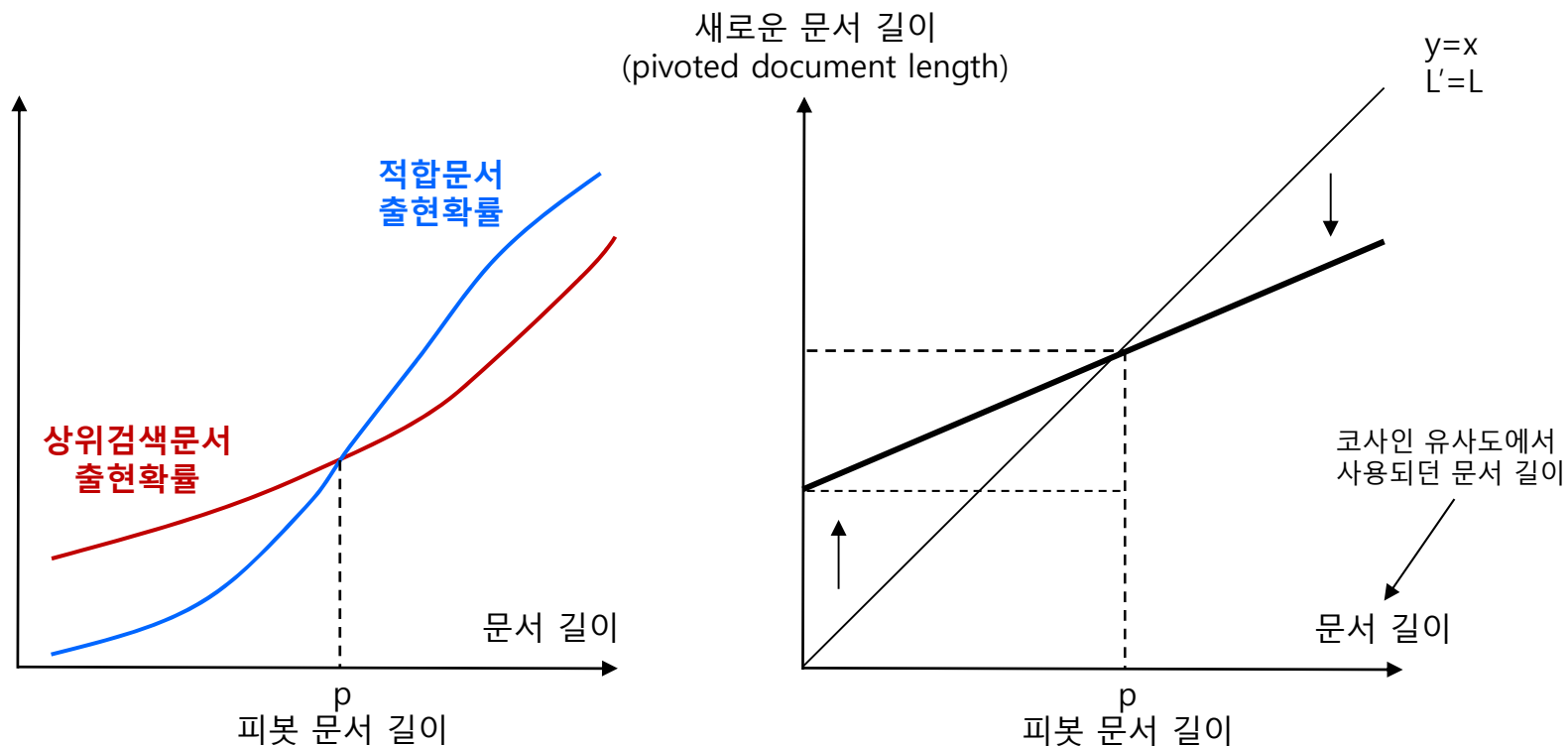




# Pivoted Document Length Normalization (피벗 기반 문서 길이 정규화)

## Pivoted Document Length Normalization (Singhal et al., 1996)

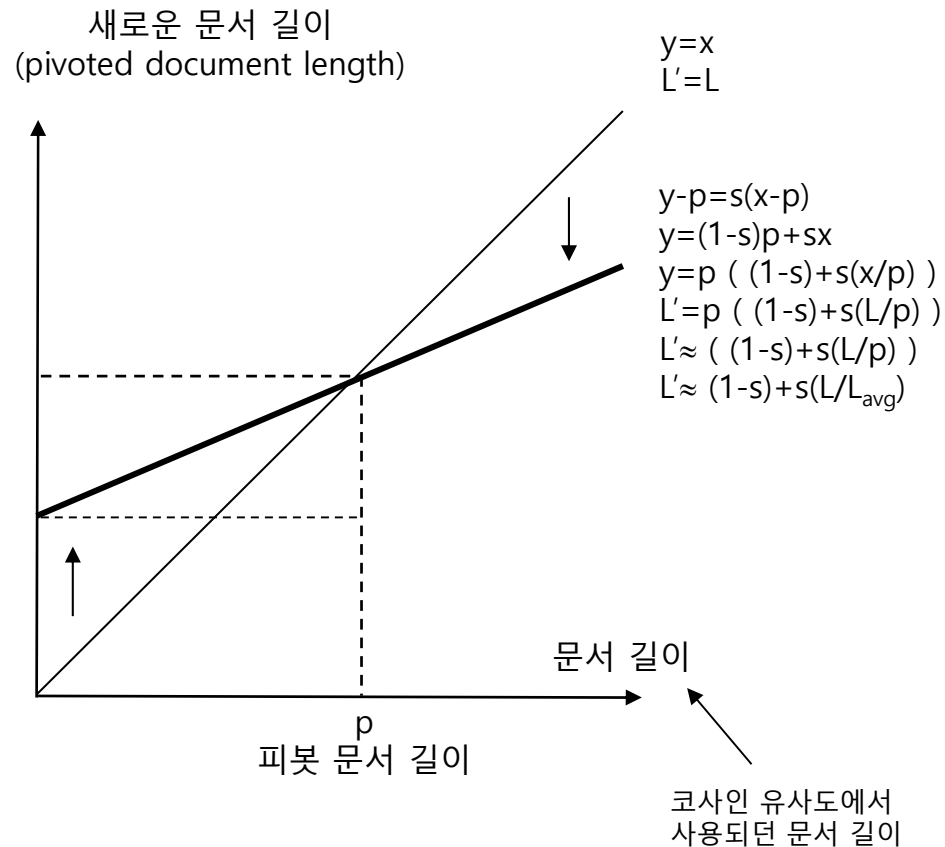
- 짧은 길이 문서의 경우 해당 길이 적합문서의 출현 확률보다 큰 확률로 검색되는 경향이 있음
- 긴 길이 문서의 경우 해당 길이 적합문서의 출현 확률보다 낮은 확률로 검색되는 경향이 있음
- 코사인 길이 정규화는 짧은 길이 문서를 선호하는 경향이 있음
- 피벗 문서 길이보다 짧은 문서는 해당 문서의 길이보다 더 큰 길이 값을 부여할 필요 있음
- 피벗 문서 길이보다 긴 문서는 해당 문서의 길이보다 더 작은 길이 값을 부여할 필요 있음



# Pivoted Document Length Normalization (피벗 기반 문서 길이 정규화)

참고:

- Amit Singhal, Chris Buckley, Mandar Mitra. (1996). Pivoted Document Length Normalization. SIGIR-1996.
- Amit Singhal, John Choi, Donald Hindle, David D. Lewis, Fernando C. N. Pereira. (1998). AT&T at TREC-7. TREC-1998.
- Amit Singhal. (2001). Modern Information Retrieval: A Brief Overview. IEEE Data Eng. Bull. 24(4):35-43.



피벗 길이 정규화 기반 새로운 문서 길이

$$\frac{1}{(1-s) + s \times \frac{dl}{avdl}}$$

dl	<p>문서길이</p> <ul style="list-style-type: none"> <li>• 벡터길이: <math>\sqrt{w_1^2 + w_2^2 + \dots + w_M^2}</math></li> <li>• 문서 내 서로 다른 단어 수 (# of unique terms)</li> <li>• 문서 byte 크기 (byte size)</li> </ul>
avdl	평균 문서 길이
s	일반적으로 0.2

# Pivoted Document Length Normalization (피벗 기반 문서 길이 정규화)

## **Pivoted cosine normalization (피벗 코사인 정규화) (Singhal et al., 1996)**

- 문서 길이로 코사인 유사도에서의 문서 벡터 길이 사용  $\sqrt{w_1^2 + w_2^2 + \dots + w_M^2}$
- 문서 벡터의 각 성분 값으로  $1 + \log(tf(t, d))$  사용 (lnc.ltc)
- pivot 값으로 문서 벡터 길이의 평균 값 사용
- slope=0.7

## **Pivoted unique normalization (피벗 용어수 정규화) (Singhal et al., 1996)**

- 문서 길이로 문서 내 서로 다른 용어 개수(# of unique terms) 사용
- pivot 값으로 문서 길이 평균 사용
- slope=0.2

## **Pivoted byte size normalization (피벗 바이트 크기 정규화) (Singhal et al., 1996)**

- OCR 스캔 등을 통해 생성된 문서 텍스트의 경우 용어 인식 오류 위험
- 문서 길이로 (용어 기반 방법 대신) 문서의 바이트 크기(byte size) 사용

# 확률모델 (Probabilistic model)

## 확률모델

- 확률이론(probability theory)에 기반한 검색모델
- 예)
  - ◆  $P(\text{Relevant} \mid D3) = 0.8$
  - ◆  $P(\text{Relevant} \mid D1) = 0.7$
  - ◆  $P(\text{Relevant} \mid D2) = 0.4$

# References

- ✚ Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze (2008). Introduction to Information Retrieval. Cambridge University Press.
- ✚ Bruce Croft, Donald Metzler, Trevor Strohman. (2009). Search Engines: Information Retrieval in Practice. Addison-Wesley Publishing Company.
- ✚ Ricardo Baeza-Yates, Berthier Ribeiro-Neto. (1999). Modern Information Retrieval. Addison-Wesley Publishing Company.
- ✚ Amit Singhal, Chris Buckley, Mandar Mitra. (1996). Pivoted Document Length Normalization. SIGIR-1996.
- ✚ Amit Singhal, John Choi, Donald Hindle, David D. Lewis, Fernando C. N. Pereira. (1998). AT&T at TREC-7. TREC-1998.
- ✚ Amit Singhal. (2001). Modern Information Retrieval: A Brief Overview. IEEE Data Eng. Bull. 24(4):35-43.