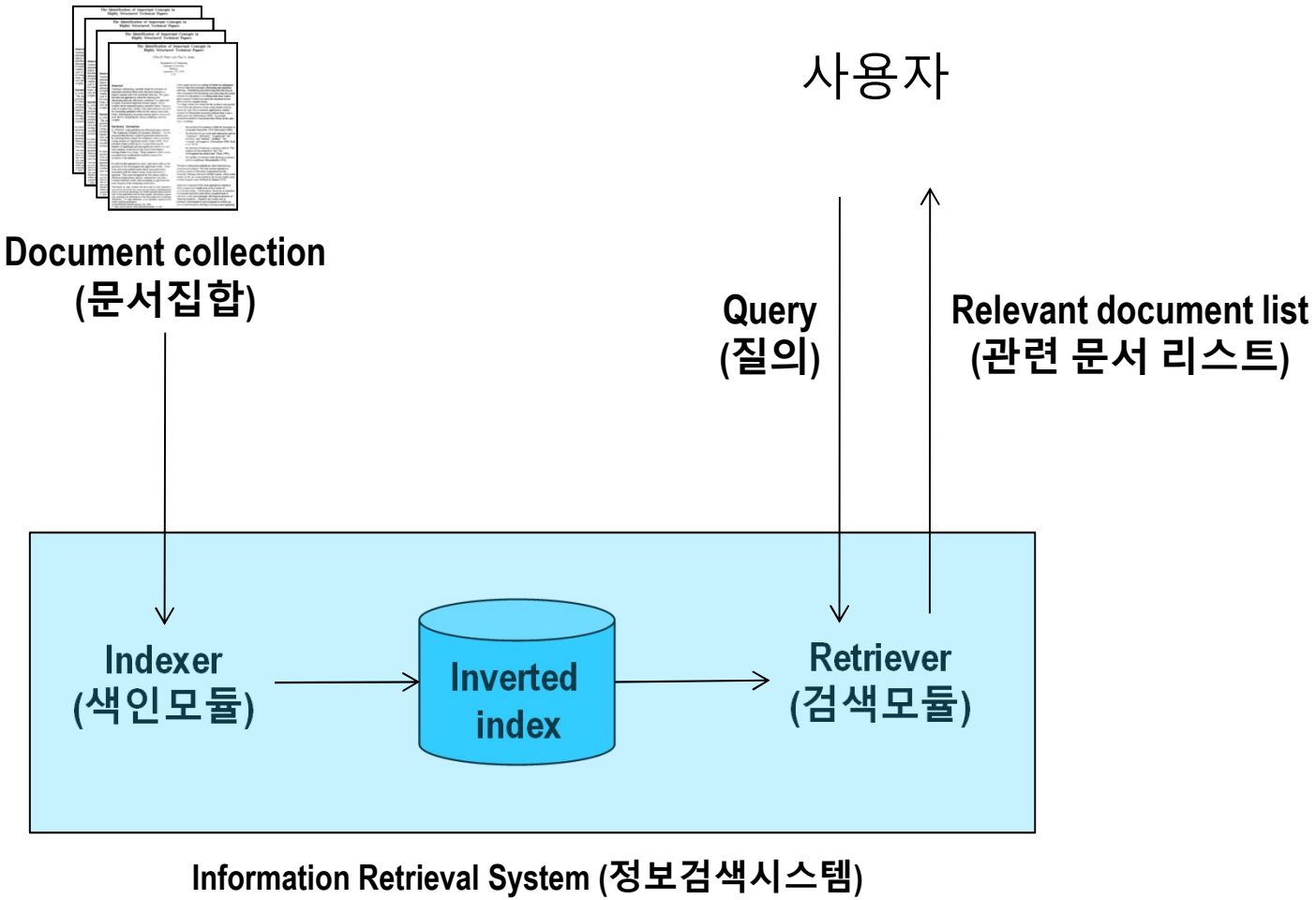
Indexing

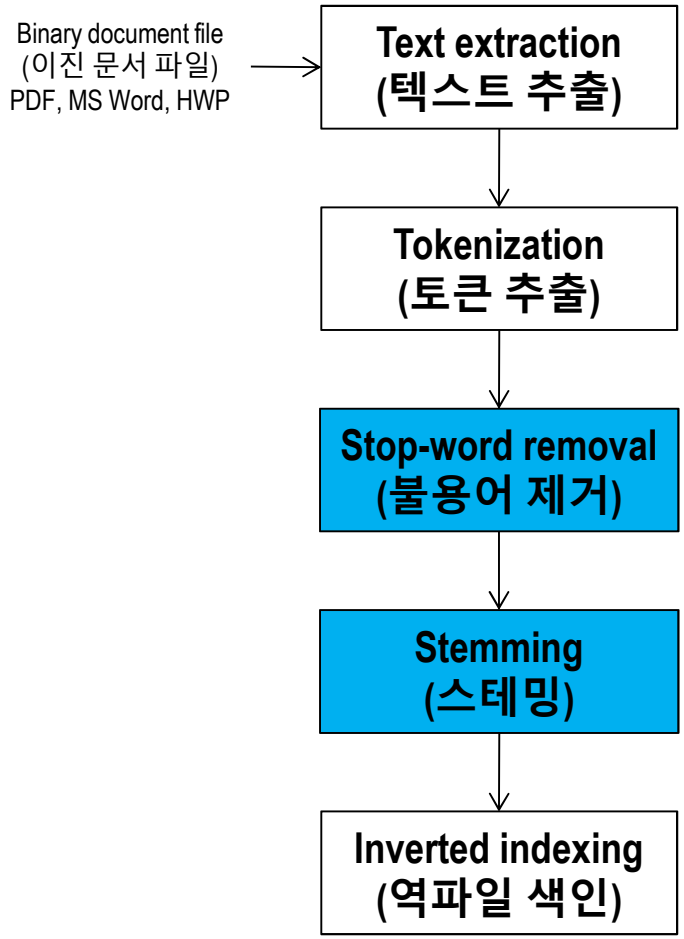
Information retrieval:

Information retrieval system (IR system)



Information Retrieval System (정보검색시스템)

Indexing (색인)



The sentences were properly revised by a student of this school

the, sentences, were, properly, revised, by, a, student, of, this, school

관사, 전치사, 대명사, be 동사를 불용어로 가정

sentences, properly, revised, student, school

sentenc, properli, revis, student, school

Indexing:

Stopword removal (불용어 제거)

English stopword list: 421
(영어 불용어 목록)

a about above across after again against all almost alone along already also although always among an and another any anybody anyone anything anywhere are area areas around as ask asked asking asks at away b back backed backing backs be because become becomes became been before began behind being beings best better between big both but by c came can cannot case cases certain certainly clear clearly come could d did differ different differently do does done down downed downing downs during e each early either end ended ending ends enough even evenly ever every everybody everyone everything everywhere f face faces fact facts far felt few find finds first for four from full fully further furthered furthering furthers g gave general generally get gets give given gives go going good goods got great greater greatest group grouped grouping groups h had has have having he her herself here high higher highest him himself his how however i if important in interest interested interesting interests into is it its itself j just k keep keeps kind knew know known knows l large largely last later latest least less let lets like likely long longer longest m made make making man many may me member members men might more most mostly mr mrs much must my myself n necessary need needed needing needs never new newer newest next no non not nobody noone nothing now nowhere number numbers o of off often old older oldest on once one only open opened opening opens or order ordered ordering orders other others our out over p part parted parting parts per perhaps place places point pointed pointing points possible present presented presenting presents problem problems put puts q quite r rather really right room rooms s said same saw say says second seconds see sees seem seemed seeming seems several shall she should show showed showing shows side sides since small smaller smallest so some somebody someone something somewhere state states still such sure t take taken than that the their them then there therefore these they thing things think thinks this those though thought thoughts three through thus to today together too took toward turn turned turning turns two u under until up upon us use uses used v very w want wanted wanting wants was way ways we well wells went were what when where whether which while who whole whose why will with within without work worked working works would y year years yet you young younger youngest your yours

출처: Fox, Christopher. "A stop list for general text." *ACM SIGIR Forum*. Vol. 24. No. 1-2. ACM, 1989.

Indexing:

Stemming (어간 생성)

Stemming increases recall(재현율) while harming precision(정확률) (Manning et al., 2008)

automate
automated
automates ⟶ autom
automating
automation

automatic
automatical ⟶ automat
automatically

operate
operating
operates
operation ⟶ oper
operative
operatives
operational

Operational research → 경영효율성향상연구
Operating system → 운영체제
Operative dentistry → 수술치과학

Python 설치 후
C:\Temp\> pip install nltk
C:\Temp\> python
>>> import nltk
>>> stemmer=nltk.stem.porter.PorterStemmer()
>>> print(stemmer.stem('automation'))
autom

Indexing: Stemming:

Porter's stemmer (Porter, 1980)

Consonant (자음) → a, e, i, o, u 이외 문자 및 자음 뒤의 y 이외 문자
Vowel (모음) → 자음이 아닌 문자

automation
↓
VCVC
↓
$(VC)^2$

measure:
한 단어나 그 일부분에서 VC가 반복되는 횟수
autom의 measure는 2임

일반형: $[C](VC)^m[V]$

Porter stemmer 규칙 일부

단계	규칙	예
1a	s →	girls → girl
1b	(m>0) eed → ee	agreed → agree feed → feed
1b	(*v*) ing →	monitoring → monitor sing → sing
2	(m>0) ation → ate	automation → automate
3	(m>0) ical → ic	practical → practic
4	(m>1) ate →	automate → autom
5a	(m>1) e →	probate → probat

ing 앞 부분에서 모음이 하나라도 있으면

ation으로 끝나는 단어에서 ation 앞부분의 **m**easure값이 0보다 크면

한 단어에 대해 1~5단계의 규칙들(1a, 1b, 1c, 2, 3, 4, 5a, 5b)을 순차 적용한다. 각 단계에서는 최장접미사가 일치되는 하나의 규칙을 적용한다

automation → automate → autom

접미사 (Suffix) 제거 규칙

(조건) suffix1 → suffix2

접미사 suffix1으로 끝나는 단어의 앞 부분이 주어진 조건을 만족하면 suffix1을 suffix2로 교체하라

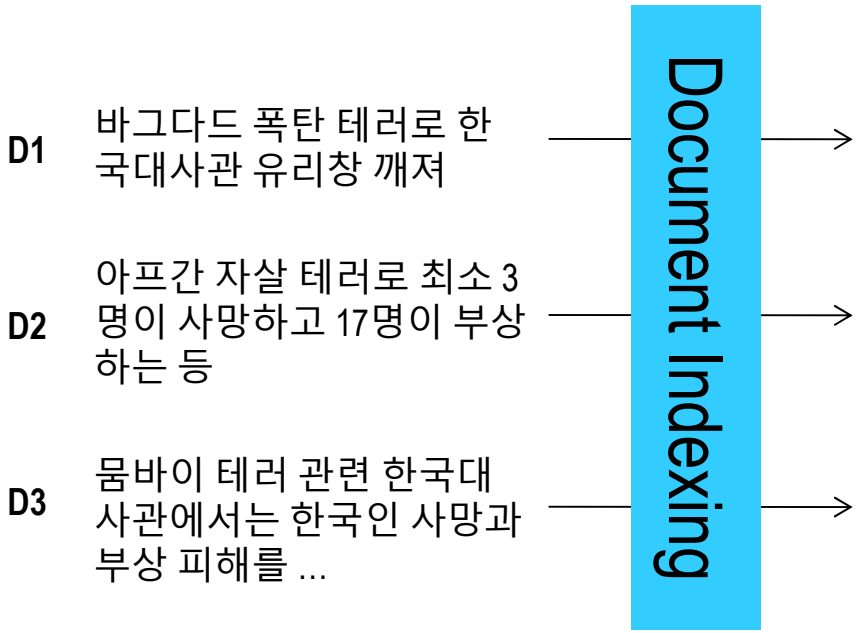
http://tartarus.org/~martin/PorterStemmer/
http://snowball.tartarus.org/algorithms/porter/stemmer.html

Stemming, n-gram 효과 (Hollink et al., 2004)

- CLEF datasets, MAP used
- Word-based vs. stemmed
  - Finnish +30.0%, Spanish +10.5%
- Word-based vs. character-level n-grams
  - 4-gram
    - Swedish +27.4%
  - 5-gram
    - Finnish +47.8%, German +20.9%

Indexing:

Inverted indexing (역파일 색인)



Inverted index (역파일 색인)

색인 용어	용어 출현 정보
바그다드	D1
폭탄	D1
테러	D1, D2, D3
한국대사관	D1, D3
유리창	D1
아프간	D2
자살	D2
사망	D2, D3
부상	D2, D3
뭄바이	D3
한국인	D3

Indexing:

Inverted indexing (역파일 색인)

D1 바그다드 폭탄 테러로 한 국대사관 유리창 깨져

D2 아프간 자살 테러로 최소 3 명이 사망하고 17명이 부상 하는 등

D3 뭄바이 테러 관련 한국대 사관에서는 한국인 사망과 부상 피해를 …

문서번호	출현 용어
D1	바그다드
D1	폭탄
D1	테러
D1	한국대사관
D1	유리창
D2	아프간
D2	자살
D2	테러
D2	사망
D2	부상
D3	뭄바이
D3	테러
D3	한국대사관
D3	한국인
D3	사망
D3	부상
D3	피해

Indexing:

Inverted indexing (역파일 색인)

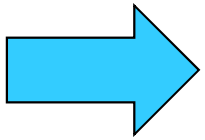
D1 바그다드 폭탄 테러로 한국대사관 유리창 깨져

D2 아프간 자살 테러로 최소 3명이 사망하고 17명이 부상하는 등

D3 뭄바이 테러 관련 한국대사관에서는 한국인 사망과 부상 피해를 ...

문서번호	출현 용어
D1	바그다드
D1	폭탄
D1	테러
D1	한국대사관
D1	유리창
D2	아프간
D2	자살
D2	테러
D2	사망
D2	부상
D3	뭄바이
D3	테러
D3	한국대사관
D3	한국인
D3	사망
D3	부상
D3	피해

Invert



출현 용어	문서번호
바그다드	D1
폭탄	D1
테러	D1
한국대사관	D1
유리창	D1
아프간	D2
자살	D2
테러	D2
사망	D2
부상	D2
뭄바이	D3
테러	D3
한국대사관	D3
한국인	D3
사망	D3
부상	D3
피해	D3

Indexing:

Inverted indexing (역파일색인)

문서번호	출현 용어
D1	바그다드
D1	폭탄
D1	테러
D1	한국대사관
D1	유리창
D2	아프간
D2	자살
D2	테러
D2	사망
D2	부상
D3	뭄바이
D3	테러
D3	한국대사관
D3	한국인
D3	사망
D3	부상
D3	피해

Invert

출현 용어	문서번호
바그다드	D1
폭탄	D1
테러	D1
한국대사관	D1
유리창	D1
아프간	D2
자살	D2
테러	D2
사망	D2
부상	D2
뭄바이	D3
테러	D3
한국대사관	D3
한국인	D3
사망	D3
부상	D3
피해	D3

Sort

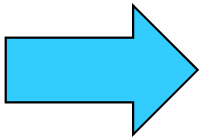
출현 용어	문서번호
바그다드	D1
뭄바이	D3
부상	D2
부상	D3
사망	D2
사망	D3
아프간	D2
유리창	D1
자살	D2
테러	D1
테러	D2
테러	D3
폭탄	D1
피해	D3
한국대사관	D1
한국대사관	D3
한국인	D3

Indexing:

Inverted indexing (역파일 색인)

출현 용어	문서번호
바그다드	D1
뭄바이	D3
부상	D2
부상	D3
사망	D2
사망	D3
아프간	D2
유리창	D1
자살	D2
테러	D1
테러	D2
테러	D3
폭탄	D1
피해	D3
한국대사관	D1
한국대사관	D3
한국인	D3

Merge



Inverted index

출현 용어	문서번호
바그다드	D1
뭄바이	D3
부상	D2, D3
사망	D2, D3
아프간	D2
유리창	D1
자살	D2
테러	D1, D2, D3
폭탄	D1
피해	D3
한국대사관	D1, D3
한국인	D3

Postings list

Posting

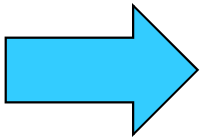
Dictionary **Postings lists**

Indexing:

Inverted indexing (역파일 색인)

출현 용어	문서번호	빈도
바그다드	D1	2
뭄바이	D3	3
부상	D2	1
부상	D3	2
사망	D2	3
사망	D3	1
아프간	D2	2
유리창	D1	1
자살	D2	1
테러	D1	2
테러	D2	3
테러	D3	2
폭탄	D1	4
피해	D3	1
한국대사관	D1	1
한국대사관	D3	1
한국인	D3	1

Merge



Dictionary

출현 용어	DF
바그다드	1
뭄바이	1
부상	2
사망	2
아프간	1
유리창	1
자살	1
테러	3
폭탄	1
피해	1
한국대사관	2
한국인	1

Postings lists

출현 용어	Postings list
바그다드	D1:2
뭄바이	D3:3
부상	D2:1, D3:2
사망	D2:3, D3:1
아프간	D2:2
유리창	D1:1
자살	D2:1
테러	D1:2, D2:3, D3:2
폭탄	D1:4
피해	D3:1
한국대사관	D1:1, D3:1
한국인	D3:1

Document frequency (df)

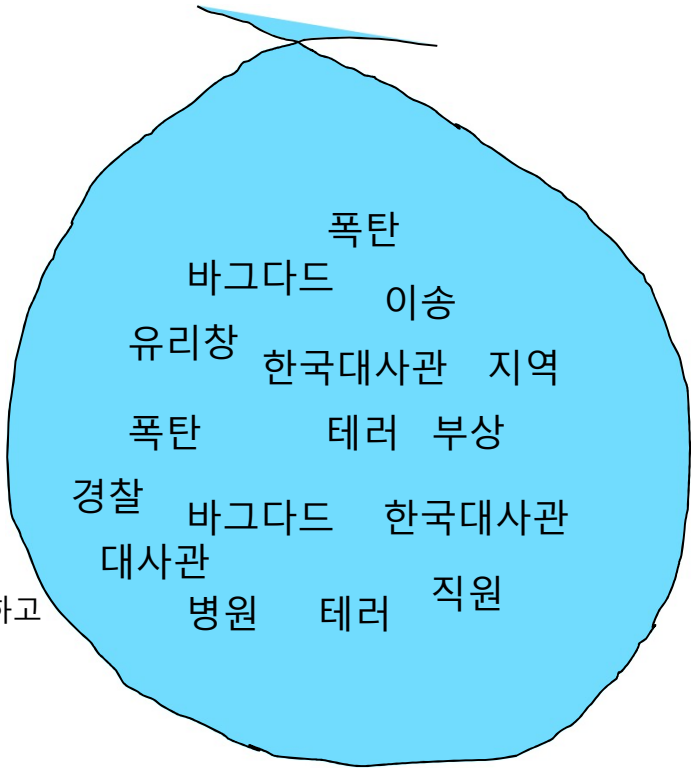
Term frequency (tf)

Document representation (문서 표현)

Document

바그다드 폭탄 테러로 한국대사관 유리창이 깨지고 대사관 직원이 부상을 입고 지역 병원으로 이송되었다. 이 폭탄 테러와 관련하여 한국대사관은 바그다드 경찰에 ...

"a bag of words (단어보자기)



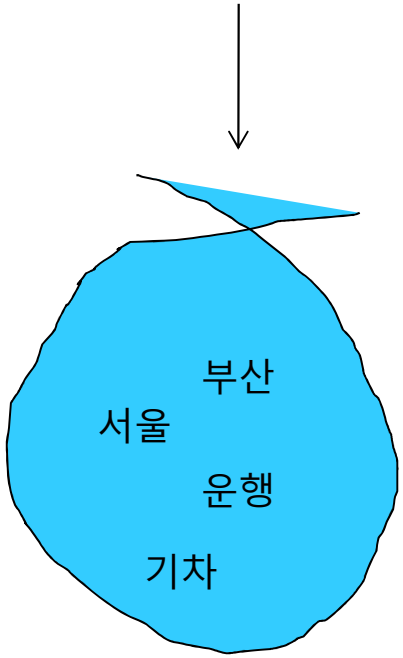
√ 실제로 한 단어의 출현은 다른 단어의 출현과 관련되어 있음에도 불구하고 문서 내에서 각 단어의 출현이 상호 독립적이라고 가정

*"**a bag of words**" document representation*

Document representation (문서 표현)

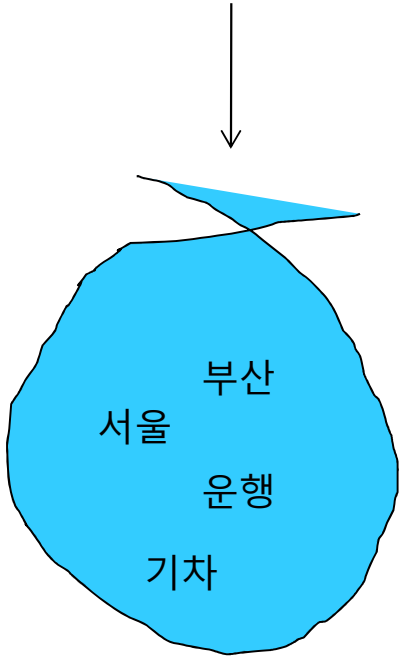
Document 1

부산에서 서울로 운행하는 기차



Document 2

서울에서 부산으로 운행하는 기차



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.
- Hollink, V., Kamps, J., Monz, C. et al. Information Retrieval (2004) 7: 33. https://doi.org/10.1023/B:INRT.0000009439.19151.4c