Vector Representation

언어를 컴퓨터가 알아듣도록 숫자로 표현하기

Vector space representation

- One hot representation (Bag of Words model)
  - (row, column) 은 (문서, 단어) 해당하는 값은 단어의 중요도 혹은 빈도수를 의미

	기계	학습	은	텍스트	마이닝	는
Doc 1	3	2	5	0	0	0
Doc 2	0	0	0	3	5	5
…	…	…	…	…	…	…

Doc 1 = [(0, 3), (1, 2), (2, 5)]
Doc 2 = [(3, 3), (4, 5), (5, 5)]

	0	1	2	3	4	5
Doc 1	3	2	5	0	0	0
Doc 2	0	0	0	3	5	5
…	…	…	…	…	…	…

Vector space representation

- One hot representation (Bag of Words model)

	0	1	2	3	4	5
Doc 1	3	2	5	0	0	0
Doc 2	0	0	0	3	5	5
…	…	…	…	…	…	…

- Column 개수 |V|는 문서 전체에서 등장한 단어 종류로, 매우 큽니다.
- 한 문서에 등장하는 단어의 개수는 적기 때문에, 대부분의 값이 0입니다 (Sparse vector).
- 문서에 등장한 단어를 쉽게 확인할 수 있어 해석이 쉽지만, 모든 단어는 다른 단어로 취급합니다. 단어간 유사성을 표현하기 어렵습니다.

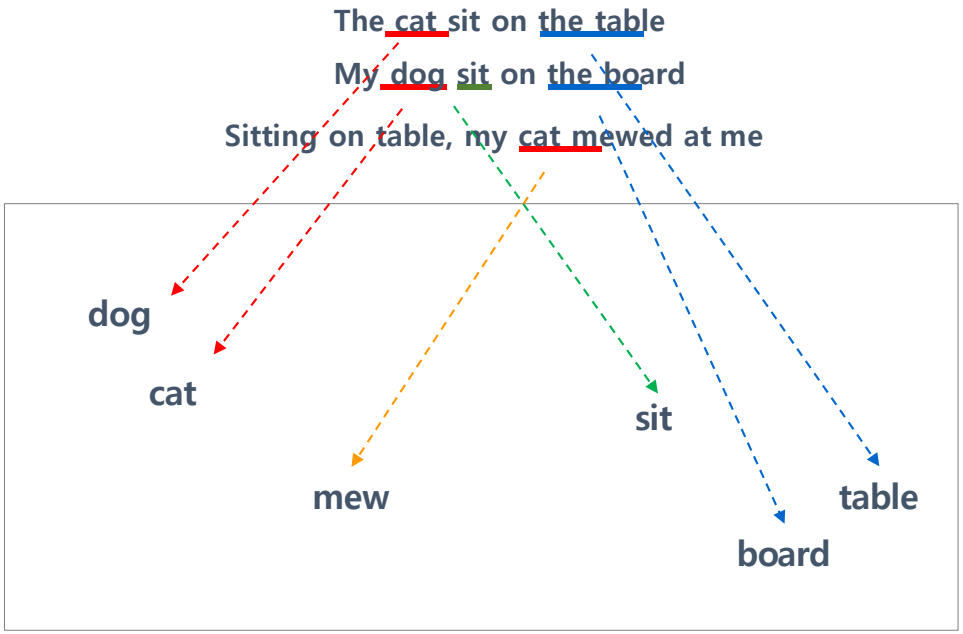
Vector space representation

- Distributed representation
  - 단어/문서를 Word2Vec 등으로 d차원 공간의 벡터로 표현
  - 단어의 "의미적 유사성"을 벡터 공간에 표현
    - 벡터 공간의 거리가 가까운 단어 또는 문서는 의미가 비슷

'dog' =	[**0.31, -0.21, 2.01, 0.58**, ...]
'cat' =	[**0.45, -0.17, 1.79, 0.61**, ...]
'topic modeling' =	[**-2.01, 0.03, 0.22, 0.54**, ...]
'dim. reduction' =	[**-1.88, 0.11, 0.19**, **0.45**, ...]

Vector space representation

- Distributed representation
  - 각 벡터는 "의미 공간"에서 좌표 역할



< 단어의 **의미 공간** >

Document clustering

- 군집화 (Clustering)는 비슷한 데이터를 하나의 집합으로 그룹화합니다.
  - 리뷰가 비슷한 영화들의 군집화 결과

cluster # 1	cluster # 2	cluster # 3
해무	인터스텔라	응답하라 1988
베를린	미스터 고	인턴
내가 살인범이다	다크 나이트	님아, 그 강을 건너지 마오
신세계	영웅: 샐러멘더의 비밀	카트
곡성(哭聲)	인셉션	인사이드 아웃
검은 사제들	트랜스포머 3	형
악마를 보았다	배틀쉽	비긴 어게인
용의자	스카이라인	두근두근 내 인생
감기	2012	라라랜드
감시자들	그래비티	반창꼬

Document clustering

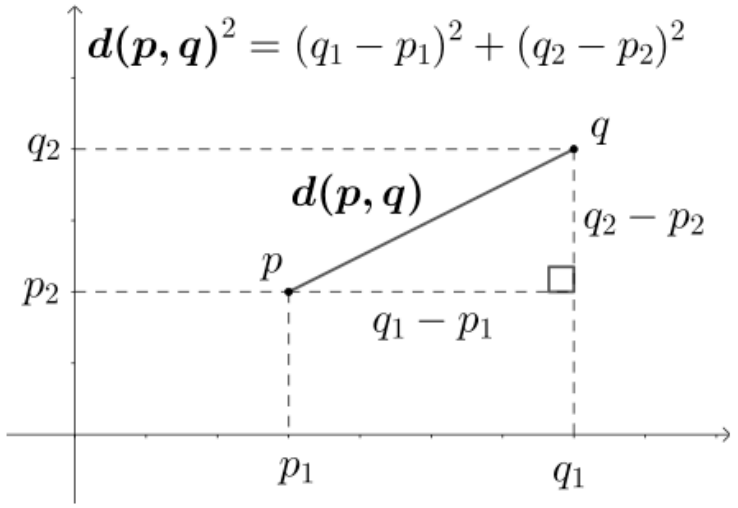
- 벡터 공간에서의 거리 척도로 Euclidean, Cosine 등이 이용됩니다.
- Euclidean distance 는 벡터의 크기(문서의 길이)에 영향을 받는다.
  - doc 1: 3 단어 / doc 2: 4 단어 / doc 3: 7 단어
  - 문서 1과 문서 3은 방향은 비슷하지만 단어 수가 달라서 거리가 멀다

	Term 1	Term 2	Term 3	Term 4	Term 5
Doc 1	1	1	1		
Doc 2			2	1	1
Doc 3	2	2	2		1

Euclidean Distance

- 두 점 사이에 줄을 긋고, 그 줄의 길이를 계산하는 것!

$$distance(A, B) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2}$$



2D에서 euclidean distance

Cosine Similarity

- Cosine similarity

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

  - $\cos((3, 0, 2), (1, 2, 0)) = \frac{3\times1 + 0\times2 + 2\times0}{\sqrt{3^2+2^2}\times\sqrt{1^2+2^2}} = \frac{3}{\sqrt{13\times5}}$

- Vector A 와 Vector B
  - 같은 방향(0) = 1
  - 완전히 반대 방향(180) = -1
  - 서로 독립적(90) = 0

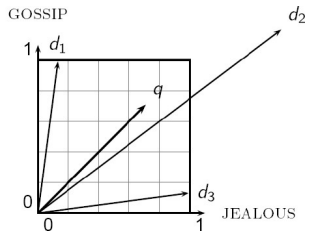
Document clustering

- Bag of words model 에는 Euclidean 보다 Cosine 이 적절합니다.
  - 두 문서에 공통으로 등장한 단어에 대해서만 유사성을 판단합니다.
  - Cosine은 문서 길이에(벡터의 크기,norm) 영향을 받지 않습니다
  - Sparse representation 에서는 벡터의 방향이 가장 중요합니다.

	Term 1	Term 2	Term 3	Term 4	Term 5
Doc 1	1	1	1		
Doc 2			2	1	1
Doc 3	2	2	2		1

Euclidean(d1, d2) = Euclidean(d1, d3)이지만, d1과 d3이 공통된 단어가 많기 때문에 더 비슷

$$Sim(D_1, D_2) = \frac{\sum_i x_{1i} x_{2i}}{\sqrt{\sum_j x_j^2}\sqrt{\sum_k x_k^2}}$$

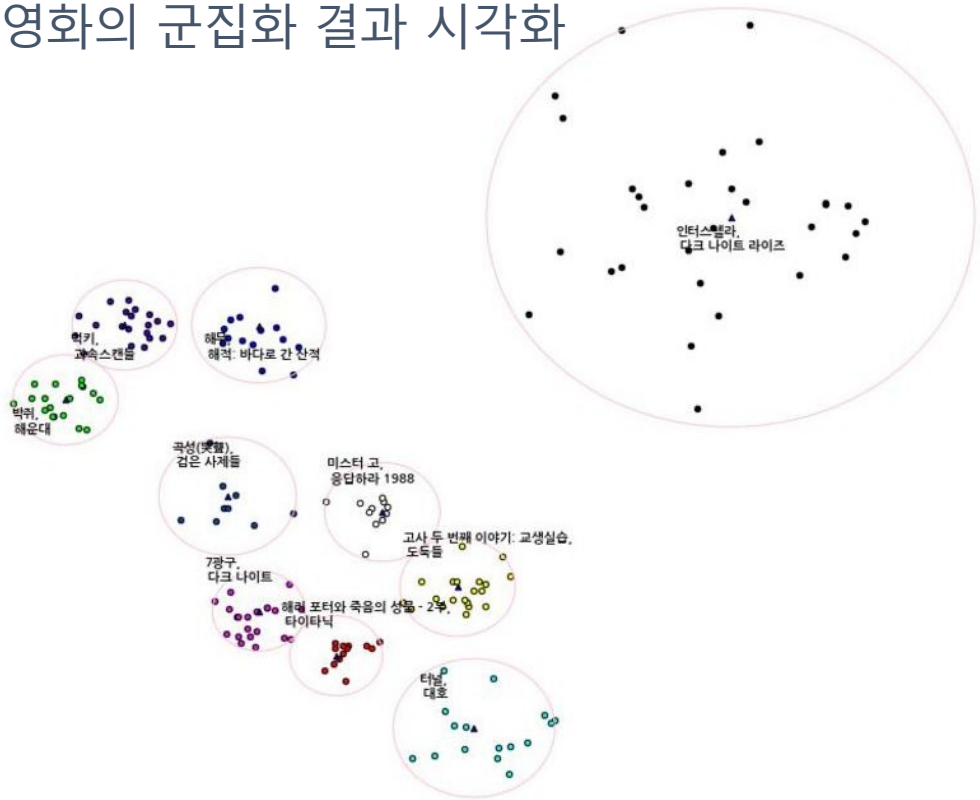


Document clustering

- Document distance/similarity 를 계산할 때에는 Cosine 이 적합합니다.
  - 문서 표현에 distributed representation 을 이용한다 하더라도 벡터의 방향이 가장 중요합니다.
  - Logistic regression, Neural network 등의 머신 러닝 알고리즘도 벡터 방향이 큰 영향을 미칩니다.

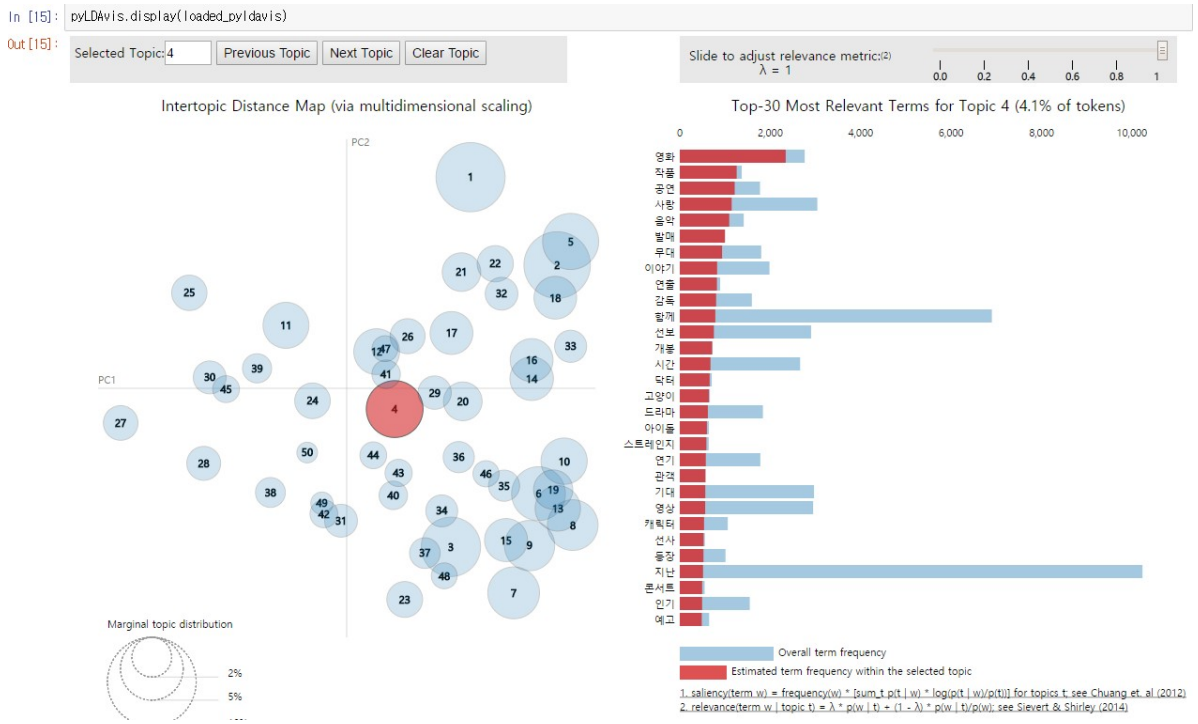
Word/Document Visualization

- 고차원의 벡터를 2차원으로 압축함으로써, 벡터 공간을 설명합니다.
  - 리뷰 기반 영화의 군집화 결과 시각화



Word/Document Visualization

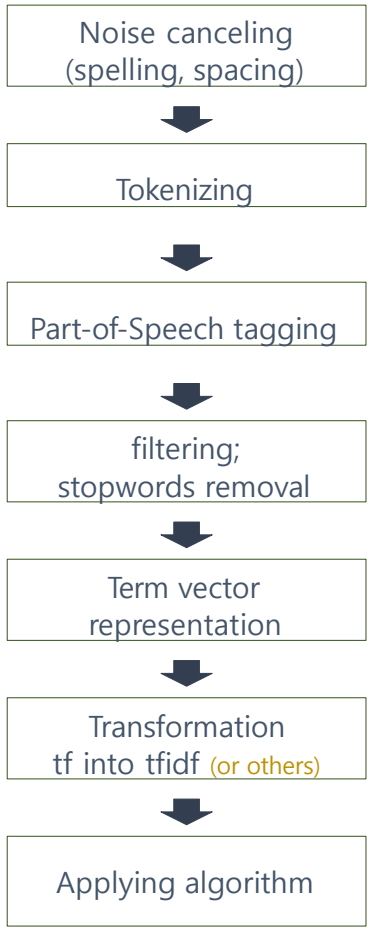
- 고차원의 벡터를 2차원으로 압축함으로써, 벡터 공간을 설명합니다.
  - pyLDAVis 를 이용한 토픽 모델링의 시각화



Text Processing

Feat. TF-IDF and some other techniques

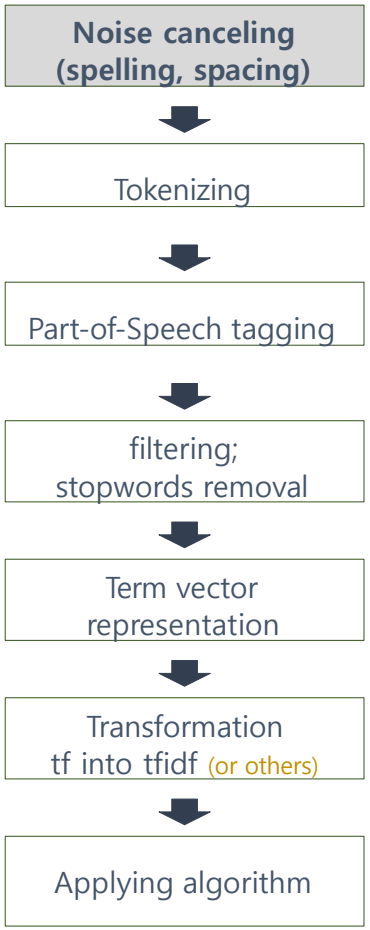
Framework


Noise canceling
(spelling, spacing)
Tokenizing
Part-of-Speech tagging
filtering;
stopwords removal
Term vector
representation
Transformation
tf into tfidf (or others)
Applying algorithm

Spelling

- 한 단어가 다르게 적힌다면,
  - 같은 의미로 여러개 다른 단어로 표현
  - Bag of words model가 더 큰 sparse representation
  - 미등록단어 (Out of vocabulary) 문제 발생

- 사전에 존재하는 올바른 단어로 수정합니다
  - Edit distance(String distance) 비교를 통해 수정



Spelling

- 오탈자는 수기로 입력된 데이터에서 주로 발생

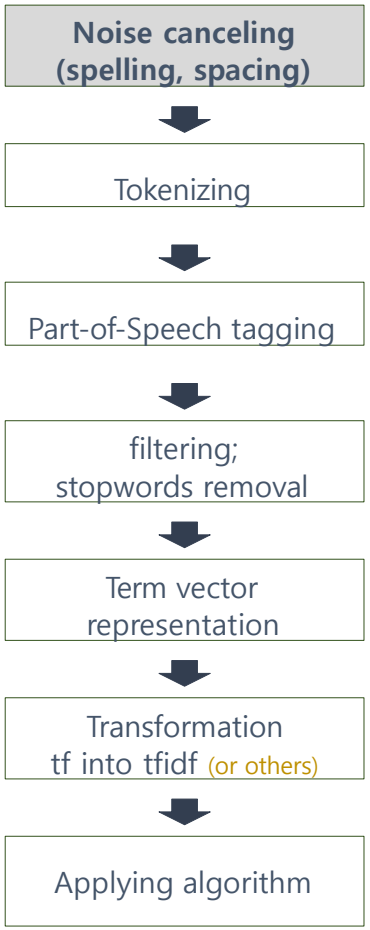
- 오탈자이외 은어, 비속어 등 올바른 표현으로 고쳐야하는 경우 종종 발생

데이터
제조외
제조, 도매, 부동산
건설업
조립금속제조, 기타화학제조
서빗스 도소매
편의점, **담배**
소매업.서비스업. 부동산업
식음료
제조 및 도소, 부동산업
제조업

사전		
가구내 고용활동	보험	운수
가스	부동산	원료재생
개인	사업시설관리	음식점
건설	사업지원	임대
과학	사회복지	임업
광업	서비스	자가소비생산활동
교육	소매	전기
국제	수도사업	전문
금융	수리	정보
기술	숙박	제조
기타	스포츠	증기
농업	어업	출판
단체	여가	폐기물처리
도매	영상	하수처리
방송통신	예술	협회
보건	외국기관	환경복원

spacing

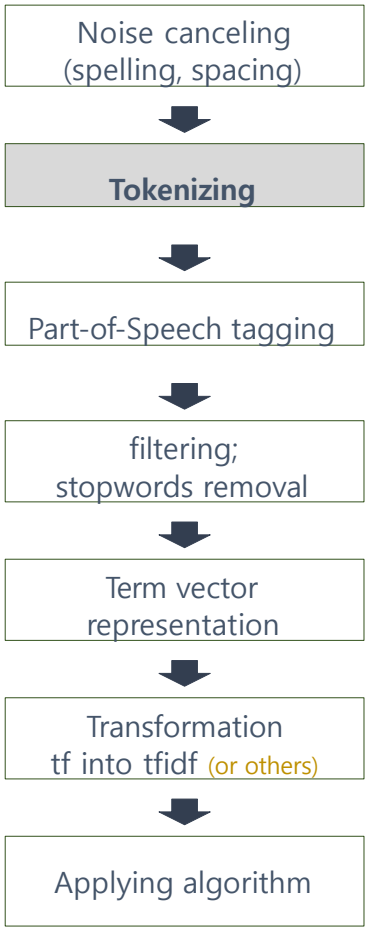
- 한국어의 어절은 띄어쓰기로 구분됩니다
  - 띄어쓰기 오류는 자연어처리기의 정확도와 계산 시간에 영향을 줍니다
  - 띄어쓰기 오류에 대응할 수 있는 토크나이저 혹은 띄어쓰기 오류 교정이 필요합니다



Tokenizing

- 토크나이징은 어절에서 단어를 나누는 것입니다
  - [토크나이징, 은, 어절, 에서, 단어,를, 나누는, 것, 입니다]

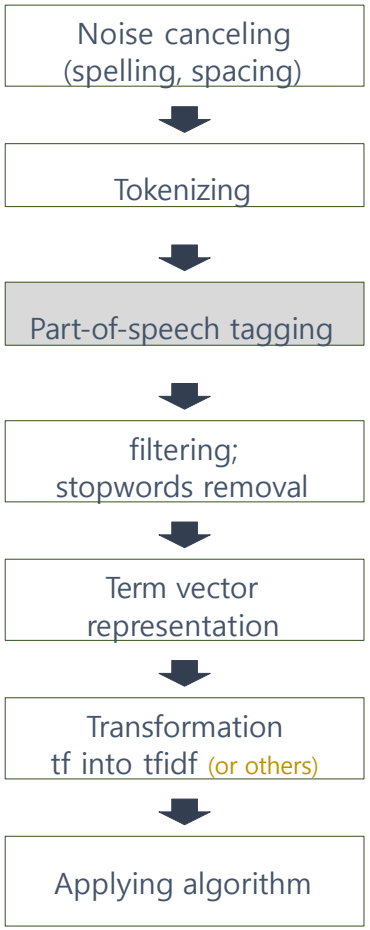
- 정확히는 "문장"을 "토큰"으로 나누는 것입니다
  - 토큰은 n-gram, 어절, 단어, phrase 등으로 목적에 따라 다르게 정의
  - 토큰을 나누는 다양한 방법이 학습에 영향을 줌



Part-of-Speech tagging

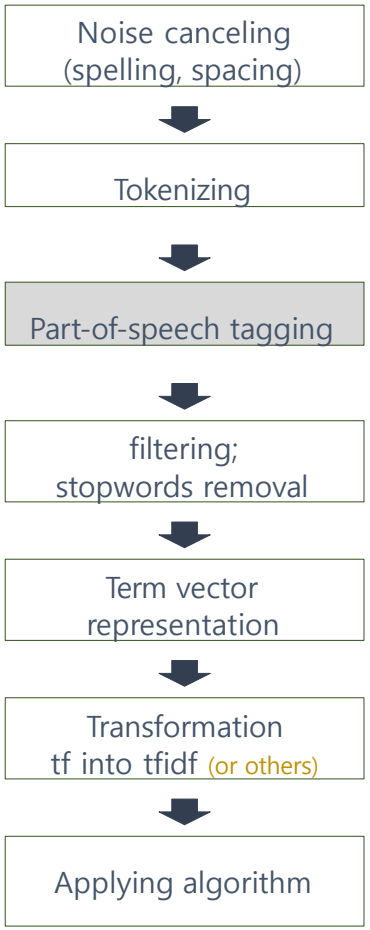
- 품사 판별은 주어진 단어의 품사를 구분합니다

  - [토크나이징, 은, 어절, 에서, 단어,를, 나누는, 것, 입니다] →
    [(토크나이징, 명사),
      (은, 조사),
      (어절, 명사),
      (에서, 조사),
      (단어, 명사),
      (를, 조사),
      (나누는, 동사),
      (것, 명사),
      (입니다, 형용사)]



Morphological analysis

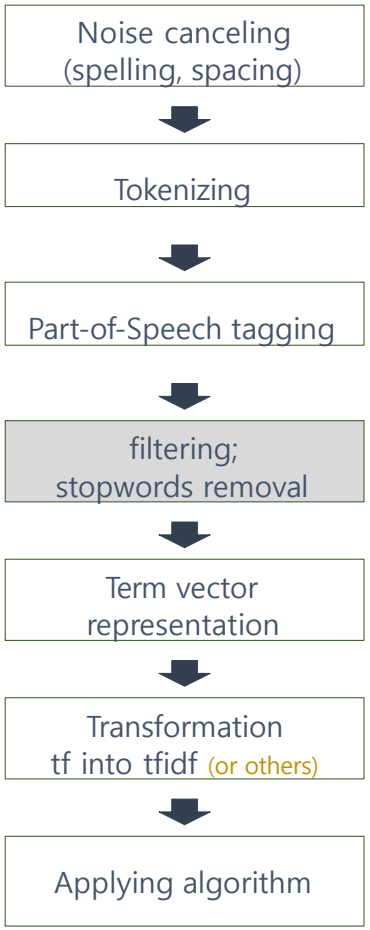
- 형태소 분석은 단어의 형태소를 인식
  - 형태소는 단어를 구성하는 최소단위
  - 품사 판별: "입니다" → 형용사
  - 형태소 분석: "입니다"
    → 이/형용사어근 + ㅂ니다/어미
- 형태소 분석을 바탕으로 단어의 품사를 추정



Stopwords removal

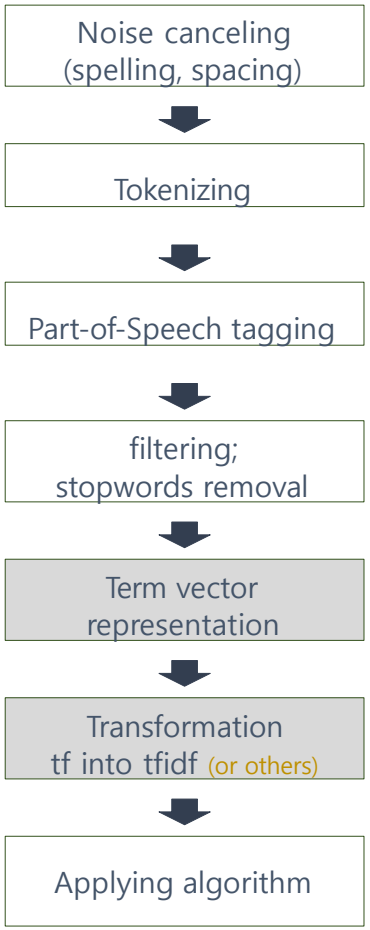
- Bag of words model 의 불필요한 단어를 제거합니다
  - 거의 등장하지 않는 단어
  - -은, -는과 같은 조사(영어에서는 a, the, am, are, ...)
  - 키워드 추출을 위한 명사 선택
  - 특수문자

"`, ′, ◇, `, ″, ′, ', ·, \", ·, △, ●, , ■, (,),
\", >>, `, /, -,~,=, • <,>, .,?, !,【,】, …, ◆,※"



Term vector representation

- Term weighting
  - $(i, j) = weight$
  - Term frequency vector 는 문서 i 에서의 단어 j 의 중요도를 단어의 빈도수로 표현
  - (i, j) 의 중요도는 정의하기 나름이며, 반드시 TF 혹은 TF-IDF 를 이용해야 하는 것도 아닙니다



Term vector representation

- TF-IDF는 Information Retrieval 을 위하여 제안된 term weighting 입니다
  - 흔하게 등장하는 단어의 영향력을 줄입니다.

$$\mathrm{TF-IDF}(w,d) = \mathrm{TF}(w) \times \log\left(\frac{N}{\mathrm{DF}(w)}\right)$$

TF: 단어 w가 문서 d에서 등장한 빈도 수

DF: 단어 w가 등장한 문서의 개수

N: 문서집합에서 문서의 총 개수.

- DF(w) = N 이면, 그 단어는 정보력이 없기 때문에 TF-IDF(w,d) = 0

Term vector representation

- Document frequency (df) 가 큰 단어는 정보력이 적습니다
  - 어디에나 등장하는 토큰은 정보력이 없음을 의미
  - 무의미하거나 문법적인 역할(조사)
  - 흔하게 등장하기 때문에 문서 간 거리를 계산할 때에도 무시

Word	Document frequency
1위	50
야구팀은	500
엘지 트윈스	1000
-은, -는	10,000

Term vector representation

흔한 단어의 영향력은 낮추고

	이번	트와이스	의	는	TT	1위	노래	음악중심	빅뱅
Doc 1	5	12	8	15	8	3	3	2	0
Doc 2	1	0	7	8	0	4	1	0	8
Doc 3	2	1	5	7	1	1	0	2	4

	이번	트와이스	의	는	TT	1위	노래	음악중심	빅뱅
Doc 1	0.3	2.5	0.2	0.1	3.2	3.6	0.3	0.8	0
Doc 2	0.06	0	0.175	0.57	0	4.8	1	0	3.3
Doc 3	0.12	0.08	0.125	0.48	0.4	1.2	0	0.8	1.65

문서 집합 전체에서 흔하게 등장하지 않고, 특정 문서에서 자주 나오는 단어의 영향력을 높임

article

박태환이 금지 약물 양성반응 통보를 받은 이후에 '도핑 파문'이 일어난 T 병원 김모 원장과 나눈 대화 내용을 녹음해 검찰에 제출한 것으로 알려졌다. 일부 매체는 이에 대해 "박태환이 김 원장에게 '아무 문제가 없는 주사약이라고 해놓고 이게 무슨 일이냐'라고 강하게 따진 것으로 전해졌 …

↓

토크나이징/ 품사판별 후

[박태환/N] [이/J] [금지/N] [약물/N] [양성반응/N] [통보/N] [를/J] …

↓

Stopword removal

Term	박태환/N	~~이/J~~	금지/N	약물/N	양성반응/N	통보/N	~~를/J~~	…
frequency	28	35	12	15	13	5	32	…

의미를 지닌 단어 선택: 명사, 동사, 형용사
문법 기능을 하는 단어 배제: 조사, 어미

↓

Vector representation

Term	1		55	21	3	27		…
frequency	28		12	15	13	5		…