세종사이버대학교

10

[AI활용텍스트분석]

토픽 모델링

학습 내용

1. 이번 강의를 시작하기에 앞서
2. 텍스트 분석 기법
3. LSA(Latent Semantic Analysis)
4. LDA(Latent Dirichlet Allocation)

학습 목표

- 토픽 모델링의 원리를 설명할 수 있다.
- 토픽 모델링의 대표적인 분석 방법인 LSA와 LDA에 대해 설명할 수 있다.
- LSA와 LDA의 파이썬 코드를 통해 실습할 수 있다.

오늘의 사전 학습

이번 강의부터 필요한 것

- 지난 강의에서 전처리한 헌법 텍스트 파일
- 구글 드라이브 계정(구글 colab 사용 예정)

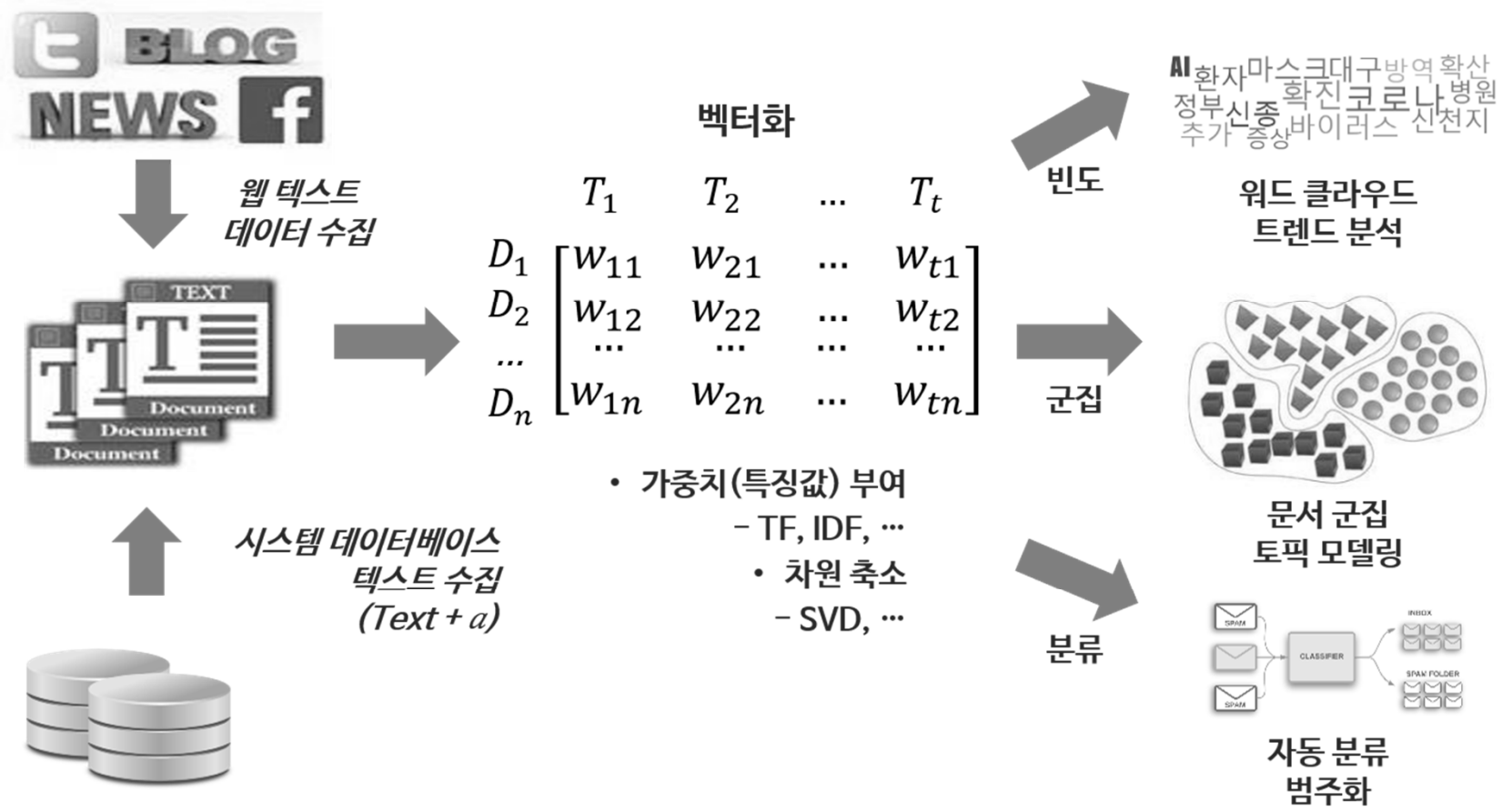
오늘의 사전 학습

이번 강의에서 얻을 수 있는 것

- 토픽 모델링의 원리를 이해한다.
- 토픽 모델링의 대표적인 분석 방법인 LSA와 LDA에 대해 공부한다.
- LSA와 LDA의 파이썬 코드를 통해 실습하여 이해한다.

복습 하기

| 텍스트 분석의 대략적인 절차



복습 하기

| 텍스트 분석 절차

1

이번 강의를 시작하기에 앞서

1) 정량화가 왜 필요한 것일까?
2) 문서 유사도
3) 흐름에 대해

1) 정량화가 왜 필요한 것일까?

세종사이버대학교

문서, 혹은 단어가 유사하다는 것의 의미

문서, 혹은 단어가 유사하다는 것은 어떤 의미인가?

문서 간 유사도를 측정하는 지표는 대체로 단어(Word, Term) 수준의 방법론들을 의미

두 문서에 겹치는 단어가 많을수록 유사도가 높다?

단어 수준의 유사도 측정은 ① 문서 길이 ② 동시 등장 단어 ③ 흔한 / 희귀한 단어 ④ 출현 빈도 등을 어떻게 처리하는지에 따라 다양한 방법론이 존재

2) 문서 유사도

세종사이버대학교

코사인 유사도

벡터 공간에서의 두 문서A와 B의 사이각에 대한 코사인 값

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

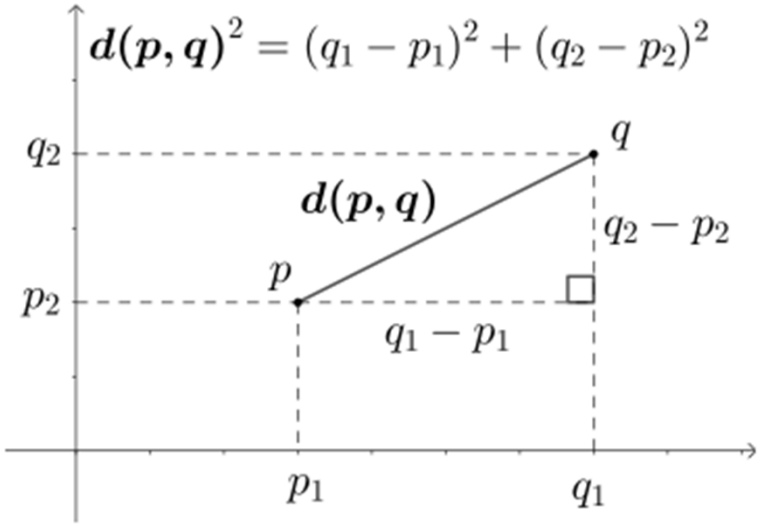
$$\cos\boldsymbol{\theta} = \boldsymbol{Similarity}(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} = \frac{\boldsymbol{x_1y_1}+\cdots+\boldsymbol{x_ny_n}}{\sqrt{\boldsymbol{x_1^2}+\cdots+\boldsymbol{x_n^2}}\sqrt{\boldsymbol{y_1^2}+\cdots+\boldsymbol{y_n^2}}}$$

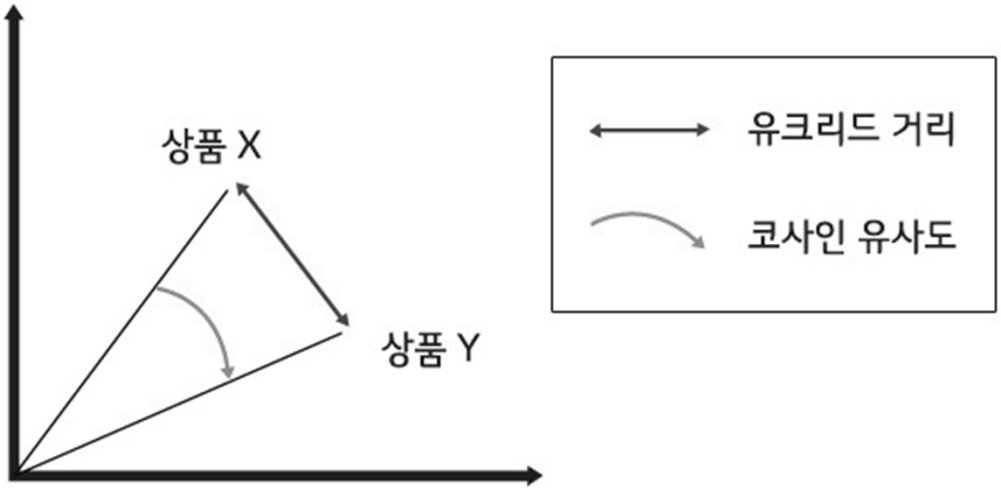
2) 문서 유사도

세종사이버대학교

거리의 유사함

벡터 공간에서의 두 문서 사이의 유클리드 거리





3) 흐름에 대해

세종사이버대학교

텍스트 분석, 검색엔진, 빅데이터, 인공지능

자동 색인
집합 모델
웹 수집
자연어 처리
벡터 모델
저장 구조
확률 모델
저장(Hadoop)
의도 예측
의도 분석
Deep Learning
키워드 광고
SNA
알파고

3) 흐름에 대해

세종사이버대학교

텍스트 분석, 검색엔진, 빅데이터, 인공지능

문헌정보학

- 자동 색인
- 자연어 처리

언어학

수학 / 통계학

- 집합 모델
- 벡터 모델
- 확률 모델

경영

- 웹 수집
- 저장 구조
- 저장(Hadoop)
- 의도 분석

비지니스

- 키워드 광고

사회, 기타

- SNA

경영

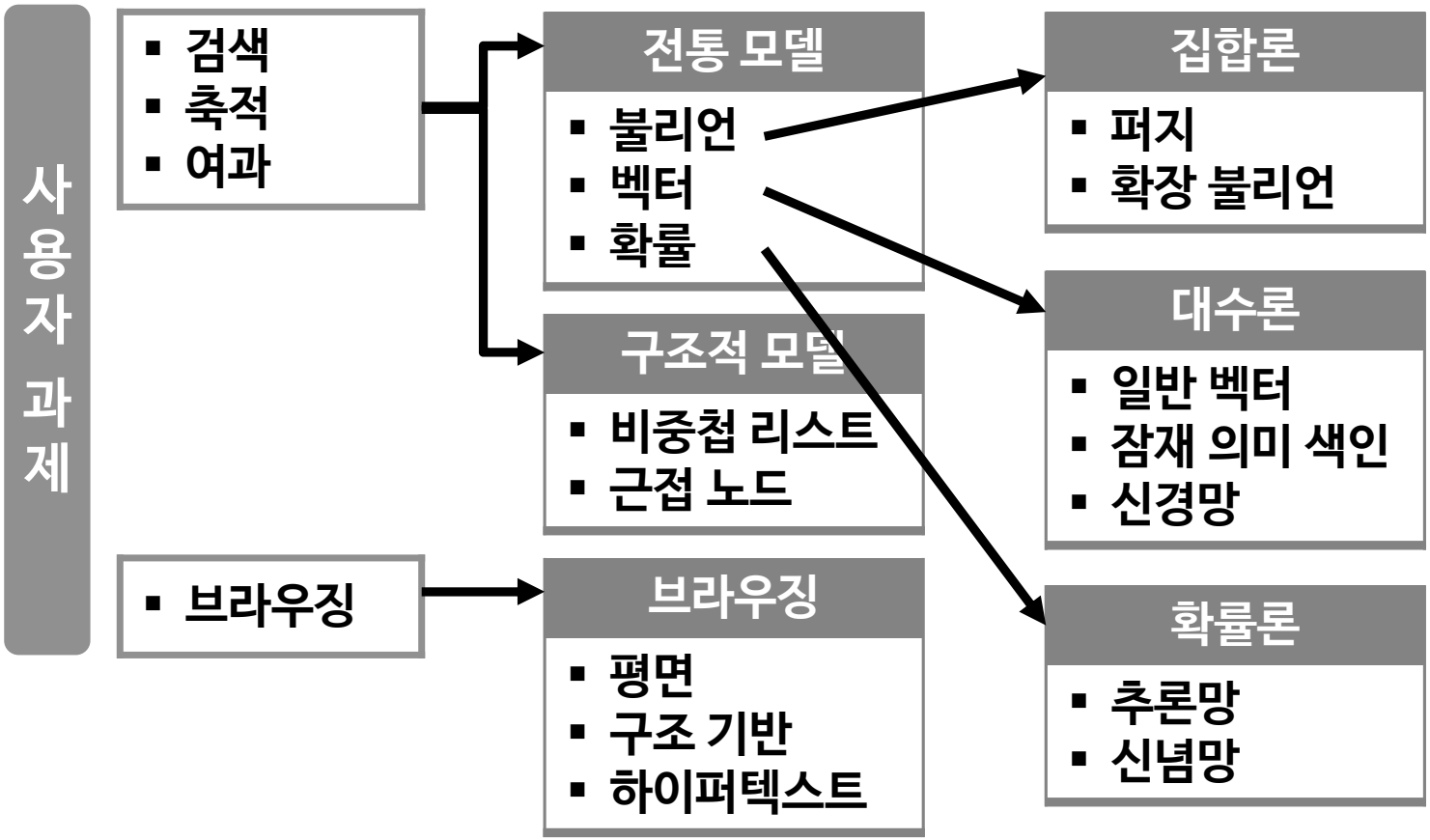
- 의도 예측
- Deep Learning
- 알파고

2) 문서 유사도

세종사이버대학교

흐름에 대해

▶ 텍스트 분석, 검색엔진, 빅데이터, 인공지능



2

텍스트 분석 기법

1) 기초적인 텍스트 분석 기법

1) 기초적인 텍스트 분석 기법

텍스트 분석은 아직까지도 명쾌한 영역을 가지고 있지 않은 것처럼 보임

어떤 분야에서는 기존의 데이터마이닝의 기법을 활용하는 것을 텍스트 분석이라 하고 또 다른 분야에서는 자연어 처리 및 텍스트 임베딩을 텍스트 분석이라고 함

본 교육에서는 두 가지 관점 모두에 대해 대략적인 내용을 살펴 봄

1) 기초적인 텍스트 분석 기법

세종사이버대학교

〈1/2〉

구분	기법 명칭	특징	비고
토픽 모델	LSA	▪ 차원 축소 기법인 SVD를 적용하여 키워드의 차원을 축소하여 연산	▪ 실질적인 효용은 미미하나 이후 관련 모델에 영향
	LDA	▪ 미리 알고 있는 주제별 단어수 분포를 바탕으로, 주어진 문서에서 발견된 단어수 분포를 분석 ▪ 확률값으로 문서의 주제 유추	▪ 인문사회 학술 논문 작성 등에서 많이 활용 ▪ 상대적으로 접근이 용이

1) 기초적인 텍스트 분석 기법

세종사이버대학교

〈2/2〉

구분	기법 명칭	특징	비고
신경망 모델	Word2vec	▪ 개별 키워드 자체를 벡터로 취급하지 않고 사용자가 정의하는 임의의 벡터 공간을 생성	▪ 단어의 벡터값 만 산출하므로 이후 추가적인 작업이 필요 ▪ 12강에서 상세하게 학습
데이터 마이닝	의사결정나무	▪ 정량화 한 텍스트 정보를 독립변수로 활용	▪ 비즈니스 응용 영역에 보조 데이터로 활용 ▪ 언어 자체에 대한 이해는 한계 ▪ 본 과정에서는 다루지 않음

3

LSA(Latent Semantic Analysis)

1) 개념
2) SVD(Singular Value Decomposition – 특이값 분해)
3) 참고 – 행렬 분해
4) 특이값 분해의 직관적 개념

1) 개념

세종사이버대학교

LSA란?

- ✓ 기존의 DTM은 단어의 의미를 전혀 고려하지 못한다는 단점이 있음
- ✓ DTM이나 TF-IDF 행렬에서 차원 축소를 통해 단어들의 잠재적인 의미를 끌어낸다는 아이디어
- ✓ 단어-문서행렬(Word-Document Matrix) 등 입력 데이터에 특이값 분해를 수행해 데이터의 차원수를 줄여 계산 효율성을 향상 및 숨어있는(Latent) 의미를 이끌어내기 위한 방법론

1) 개념

세종사이버대학교

LSA란?

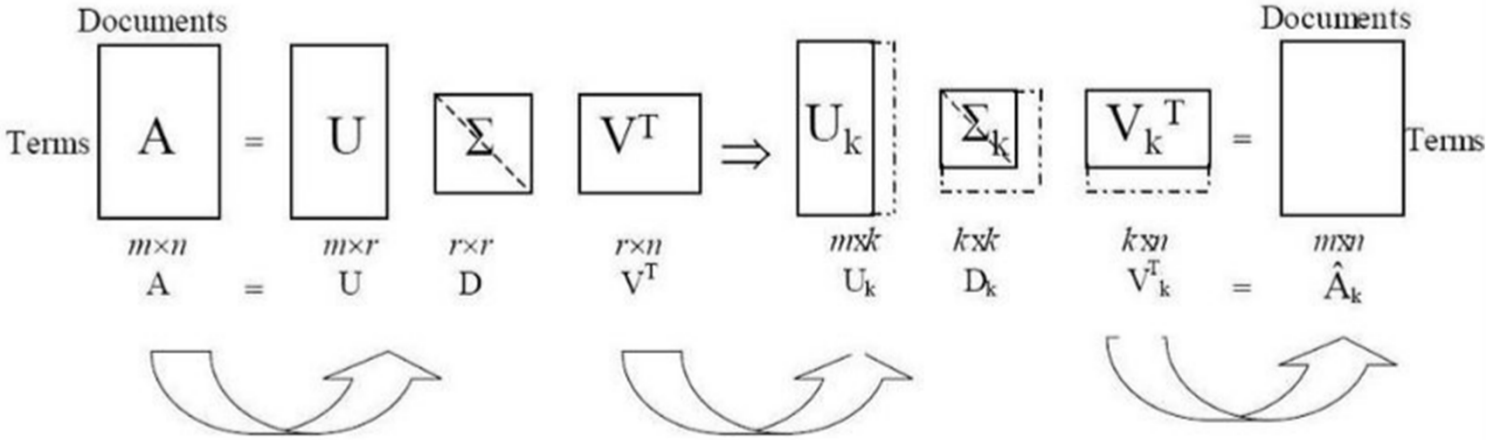
- ✓ 토픽 모델링을 위해 최적화된 알고리즘은 아니지만, 토픽 모델링이라는 분야에 아이디어를 제공한 알고리즘
- ✓ 차원 축소 기법은 SVD(Singular Value Decomposition-특이값 분해)를 사용함
- ✓ LSI(Latent Semantic Indexing)라고도 함
- ✓ 아이디어는 제공했으나 실질적인 상황에서 유용하지는 않음

2) SVD(Singular Value Decomposition – 특이값 분해)

A가 m × n 행렬일 때, 다음과 같이 3개의 행렬의 곱으로 표현

$$A = U\Sigma V^T$$

- 임의의 행렬 X를 세 행렬 U, Σ, V의 곱으로 분해
- U, V는 각각 직교행렬(각 열벡터가 서로 직교)
- Σ는 대각 행렬(대각 성분 이외에는 모두 0)


Documents
Terms
A
=
U
Σ
V^T
⇒
U_k
Σ_k
V_k^T
=
Documents
Terms
m×n
m×r
r×r
r×n
m×k
k×k
k×n
m×n
A
=
U
D
V^T
U_k
D_k
V_k^T
=
Â_k

3) 참고 – 행렬 분해

행렬 분해(行列分解, Matrix Decomposition)

- 행렬을 특정한 구조를 가진 다른 행렬의 곱으로 나타내는 것을 의미함
- 행렬분해는 선형 방정식의 해를 구하거나, 행렬 계산을 효율적으로 하거나, 행렬의 특정 구조를 밝히는 등의 목적으로 사용됨

3) 참고 – 행렬 분해

세종사이버대학교

선형 방정식과 관련한 분해	▪ LU 분해 ▪ QR 분해 ▪ 계수 인수분해 ▪ 숄레스키 분해
고유값에 근거한 분해	▪ 고유값 분해 ▪ 조르단 분해 ▪ 슈어 분해 ▪ QZ 분해 ▪ 특이값 분해 ▪ 다카기 분해
다른 분해 방법들	▪ 극분해 ▪ 모스토우 분해 ▪ 싱크혼 일반 형식 ▪ 윌리엄슨 일반 형식

3) 참고 – 행렬 분해

세종사이버대학교

$A = U \quad \Sigma' \quad \Sigma \quad V^T$

문서(6) × 단어(7)

1	1	0	0	0	0	0
0	0	1	1	1	0	0
0	1	0	0	0	1	0
0	0	1	0	0	0	1
0	0	0	1	1	0	1
1	0	1	0	1	1	1

U (문서(6) × 분류(차원))

0.12	0.57	−0.32	0.00	−0.71	−0.24
0.44	−0.36	−0.41	0.71	0.00	−0.08
0.12	0.57	−0.32	0.00	0.71	−0.24
0.33	−0.07	0.56	0.00	0.00	−0.75
0.44	−0.36	−0.41	−0.71	0.00	−0.08
0.69	0.30	0.37	0.00	0.00	0.55

Σ' (분류(차원) × 분류(차원))

2.98	0.00	0.00	0.00	0.00	0.00	0.00
0.00	1.88	0.00	0.00	0.00	0.00	0.00
0.00	0.00	1.36	0.00	0.00	0.00	0.00
0.00	0.00	0.00	1.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	1.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.87	0.00

Σ (Σ' 와 동일)

V^T (분류(차원) × 단어(7))

0.27	0.08	0.49	0.30	0.53	0.27	0.49
0.46	0.61	−0.07	−0.38	−0.22	0.46	−0.07
0.04	−0.47	0.38	−0.61	−0.34	0.04	0.38
0.00	0.00	0.71	0.00	0.00	0.00	−0.71
−0.71	0.00	0.00	0.00	0.00	0.71	0.00
0.35	−0.56	−0.33	−0.18	0.44	0.35	−0.33
0.30	−0.30	0.00	0.60	−0.60	0.30	0.00

$U\Sigma' =$ (문서(6) × 분류(차원))

분류(차원)	
0.3573	1.0707
1.31	−0.6762
0.3573	1.0707
0.9825	−0.1315
1.31	−0.6762
2.0543	0.5635

문서(6)

$\Sigma' V^T =$ (분류(차원) × 단어(7))

단어(7)						
0.8038	0.2382	1.4588	0.8932	1.5779	0.8038	1.4588
0.8641	1.1458	−0.1315	−0.7138	−0.4132	0.8641	−0.1315

4

LDA(Latent Dirichlet Allocation)

1) 개념
2) LDA 동작
3) 결론

1) 개념

LDA(Latent Dirichlet Allocation)

관찰된 변수(Observed Variable)를 통해 각각의 확률을
계산하여 토픽을 생성하는 사후 추론 방법

선형 대수적인 도구(Singular Value Decomposition)를
사용해서 단어와 개념 사이의 관계를 파악하는 기법인
LSA(Latent Semantic Index)에서 발전됨

각 문서는 주제가 무작위로 혼합 되어 있으며 각 단어는 해당 주제 중 하나에서 나옴

Topics

Documents

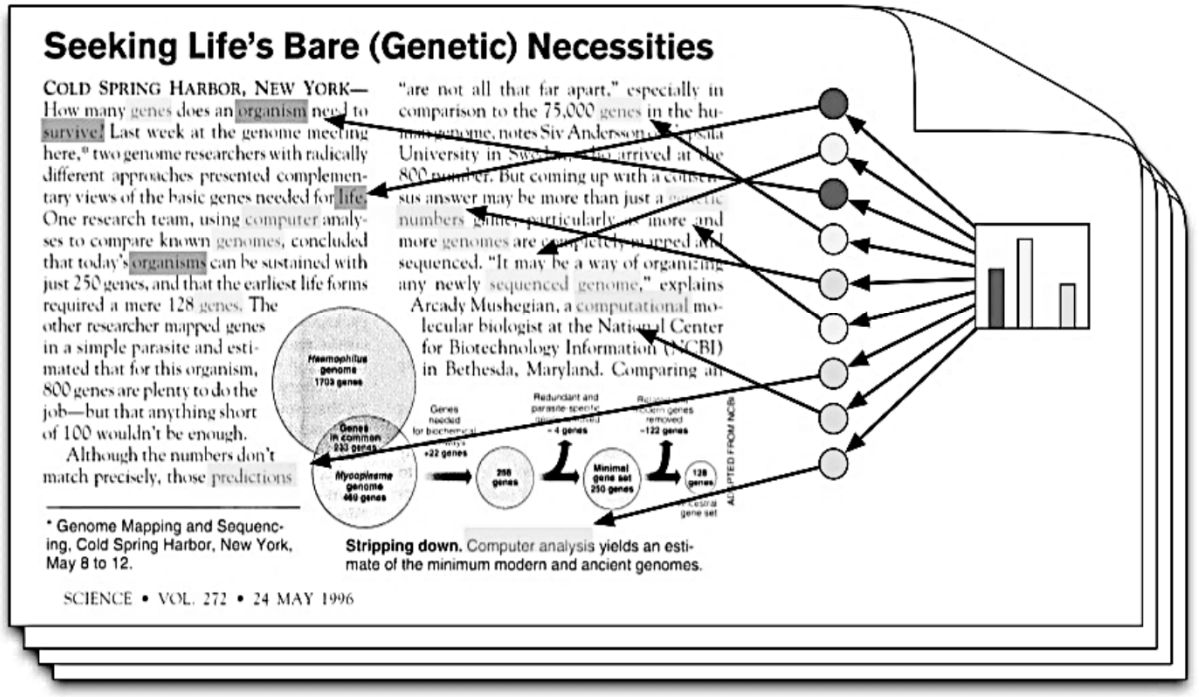
Topic proportions and assignments

```
gene      0.04
dna       0.02
genetic   0.01
...

life      0.02
evolve    0.01
organism  0.01
...

brain     0.04
neuron    0.02
nerve     0.01
...

data      0.02
number    0.02
computer  0.01
...
```



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

2) LDA 동작

세종사이버대학교

순서

1 사용자는 토픽의 개수 k 지정

- LDA가 나누게 될 토픽 개수는 사용자가 지정
- LDA는 토픽 개수 k를 입력 받으면, k개의 토픽이 D개의 전체 문서에 걸쳐 분포되어 있다고 가정
- 최적의 토픽 개수를 찾기 위한 방법
  - Perplexity, Coherence

2) LDA 동작

세종사이버대학교

순서

2 모든 단어를 k개의 토픽 중 하나의 토픽에 임의 할당

- 모든 문서의 모든 단어에 대해서 k개의 토픽 중 하나의 토픽을 랜덤으로 할당
- 할당 후 각 문서는 토픽을 가지며, 토픽은 단어 분포를 가지는 상태

2) LDA 동작

세종사이버대학교

순서

3 모든 문서의 모든 단어에 대해서 아래의 사항을 반복 진행

- 어떤 문서의 각 단어 w는 자신은 잘못된 토픽에 할당되어져 있지만, 다른 단어들은 전부 올바른 토픽에 할당되어져 있는 상태라고 가정
- 이에 따라 단어 w는 아래의 두 가지 기준에 따라서 토픽 재할당
  - P(topic t | document d) : 문서 d의 단어들 중 토픽 t에 해당하는 단어들의 비율
  - P(word w | topic t) : 단어 w를 갖고 있는 모든 문서들 중 토픽 t가 할당된 비율

4 이를 반복하면, 모든 할당이 완료된 수렴 상태 완료

2) LDA 동작

세종사이버대학교

예시

- 문서 1 : 대한민국, 민주공화국
- 문서 2 : 대한민국, 주권, 국민, 권력, 국민
- 문서 3 : 대한민국, 국민, 요건, 법률

	국민	권력	대한민국	민주공화국	법률	요건	주권
문서 1	0	0	1	1	0	0	0
문서 2	2	1	1	0	0	0	1
문서 3	1	0	1	0	1	1	0

	국민	권력	대한민국	민주공화국	법률	요건	주권
문서 1	0	0	B	A	0	0	0
문서 2	B	A	A	0	0	0	A
문서 3	???	0	B	0	A	B	0

토픽A	토픽B
50%	50%
75%	25%
33%	66%

	토픽A	토픽B
국민	0%	100%
권력	100%	0%
대한민국	33%	66%
민주공화국	100%	0%
법률	0%	100%
요건	0%	100%
주권	100%	0%

문서3의 '국민'에 대한 토픽의 결정은,

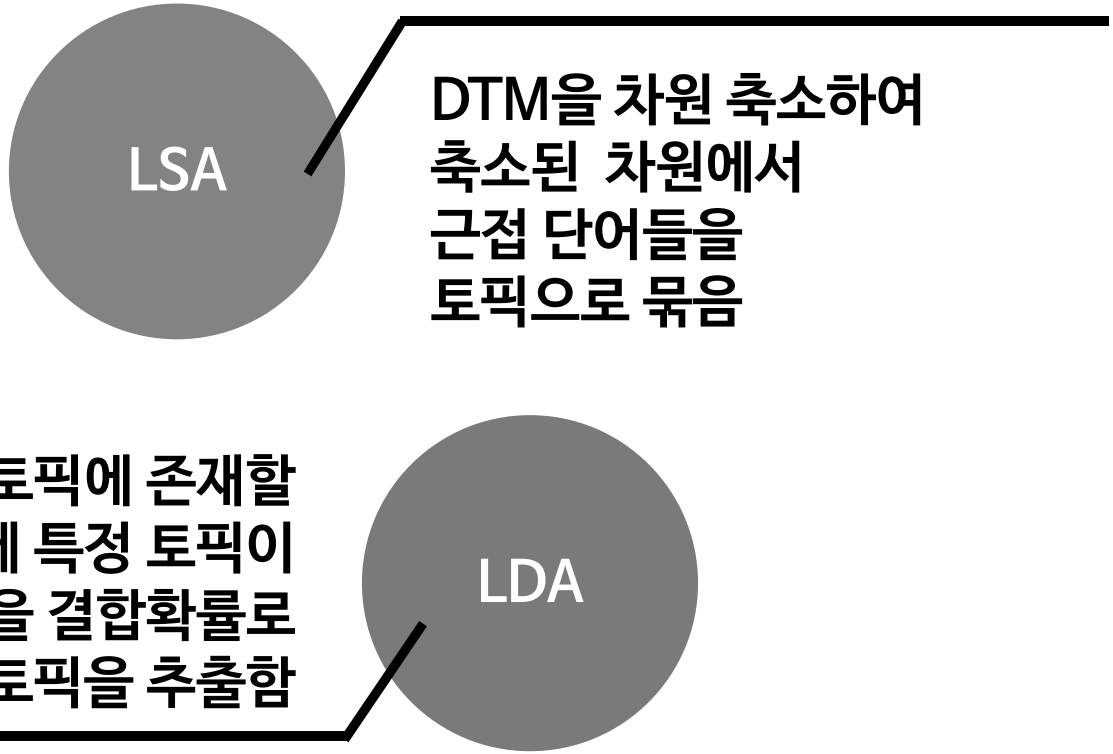
1) 문서3 자체는 토픽 B일 확률이 더 높음
2) 단어-토픽 행렬에서 토픽 B일 확률이 높음

잠재 토픽에 대한 두 확률의 곱으로 결정

모든 단어에 대해 반복 → 학습의 과정

3) 결론

세종사이버대학교



실습하기 : LDA, LSA 실습(파이썬 실습)

학습 정리

LSA

- 기존의 DTM은 단어의 의미를 전혀 고려하지 못한다는 단점
- DTM에서 차원 축소를 통해 근접 단어들을 토픽으로 묶어 단어들의 잠재적인 의미를 끌어낸다는 아이디어
- 차원 축소 기법은 SVD(Singular Value Decomposition : 특이값 분해)를 사용
- 토픽 모델링이라는 분야에 아이디어를 제공한 알고리즘이지만 실제 상황에서 유용하지는 않음

학습 정리

LDA

- 관찰된 변수(Observed Variable)를 통해 각각의 확률을 계산하여 토픽을 생성하는 사후 추론 방법
- 단어가 특정 토픽에 존재할 확률과 문서에 특정 토픽이 존재할 확률을 결합확률로 추정하여 토픽 추출
- LSA(Latent Semantic Index)에서 발전됨
- 인문, 사회 분야의 연구에서 많이 활용