특강

토픽 모델링 (Topic modeling)

2022/05/23 | 빅데이터 동아리

송지훈

공지사항

찾아가는 캠퍼스 특허 유니버시아드 (CPU) 설명회

- 웹사이트: https://www.kipa.org/cpu/1_info_01.jsp
- 일시: 2022년 5월 27일 14시~16시
- 장소: 법과대학 416 강의실
- 대회 신청기간: 2022년 6월 9일 (목요일) 까지
- 대회에 참가하길 희망하는 학생은 5월 27일 설명회에 꼭 참석하시기 바랍니다.

학습목표

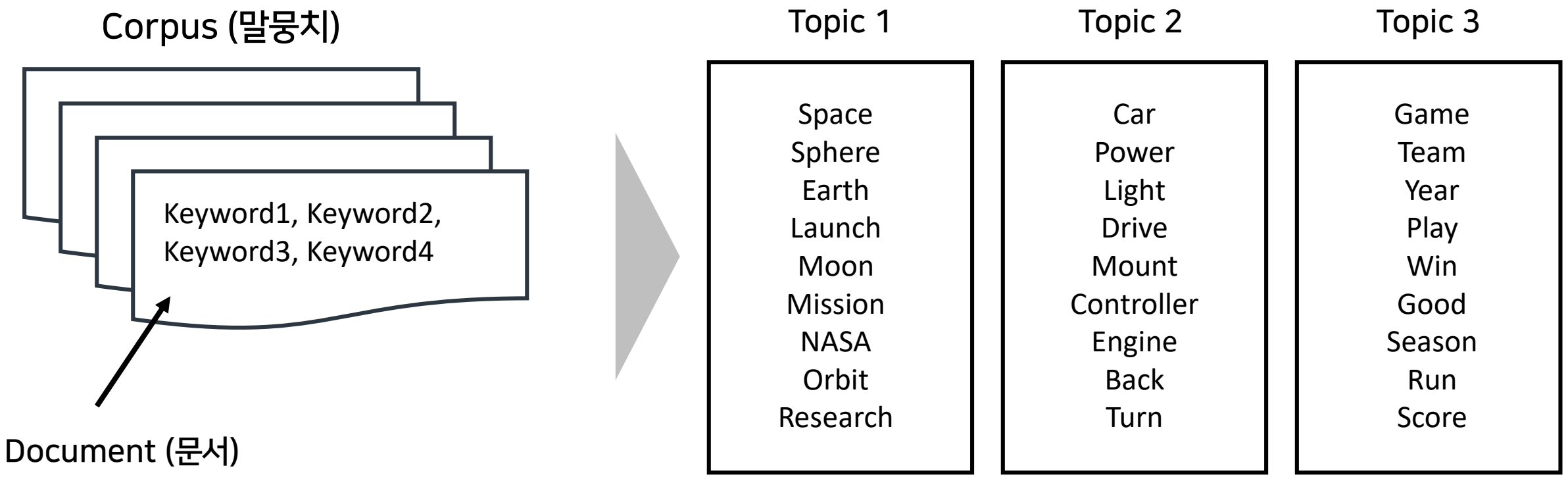
- 토픽 모델링이란?
- 토픽 모델링 실습 (구글 Colab 활용)

토픽 모델링이란?

토픽 모델링 (Topic modeling)

- 문서 집합에서 주제를 찾아내는 기법 → 문서내에 존재하는 다양한 키워드를 바탕으로 주제(topic)를 도출하는 통계적 분석 방법 (아이디어: "방대한 양의 문서가 존재할 때 누가 이걸 대신 읽고 주제를 파악해줄 수 있을까?")

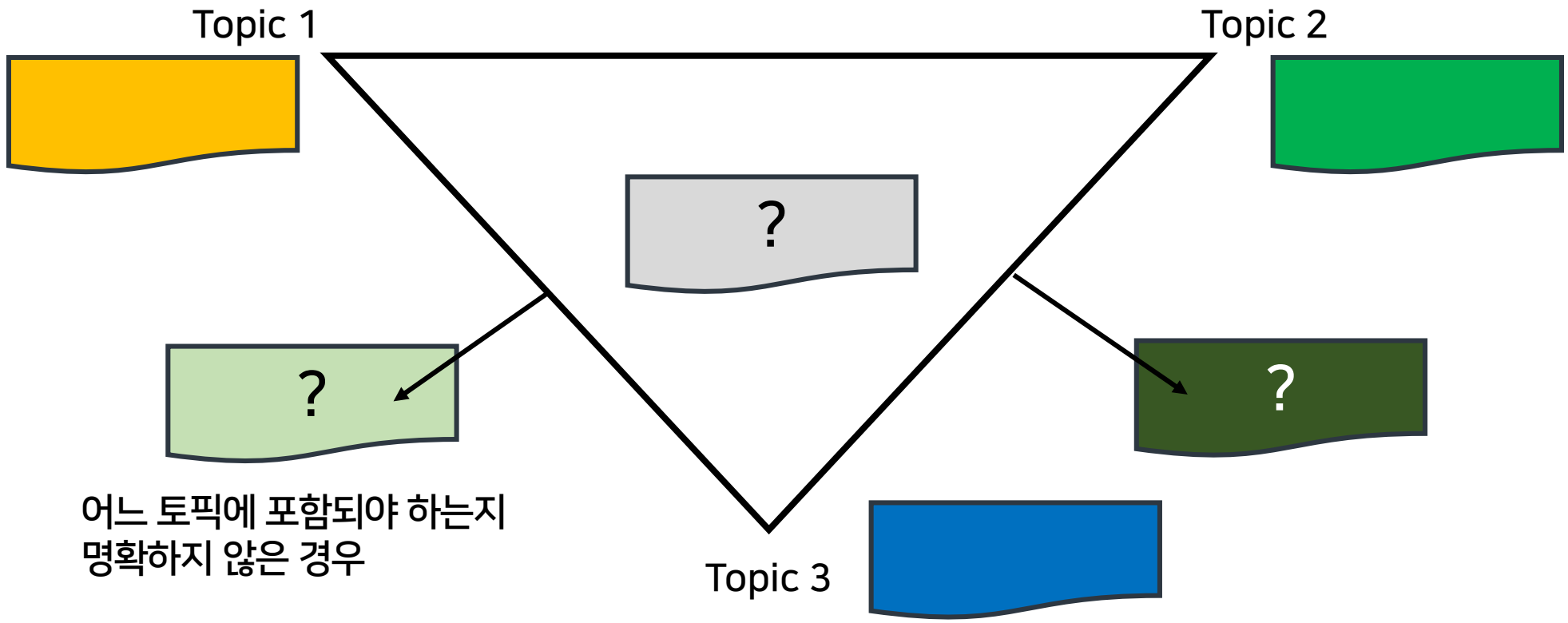
1) 토픽차원에서 설명: 개별 토픽을 기준으로, 어떤 단어들이 빈번하게 등장을 하는가



토픽 모델링이란?

토픽 모델링 (Topic modeling)

2) 문서차원에서 설명: 각각의 개별문서가 어떤 토픽을 더 많이 내포하고 있는지 파악
(특정 토픽이 해당하는 문서에서 차지하는 비중을 산출)



토픽 모델링이란?

토픽 모델링을 이용한 핀테크 기술 동향 분석

김태경, 최회련, 이홍철*
고려대학교 산업경영공학과

A Study on the Research Trends in Fintech using Topic Modeling

Table 3. Topic Modeling

no.	Topic Word
Topic 1	transfer, account, send, machine, ATM, automat, machine, device, money, bank
Topic 2	trade ,order, price, market, exchange, product, trader, match, rate, valu
Topic 3	modul, mobil, payment, card, nfc, chip, function. device, equip, phone
Topic 4	cpo, identifi, offer, provid, control, accept, financi, plan, custom, manag
Topic 5	network, secur, support, access, modul, embodi, comput, software, onlin, program
Topic 6	loan, amount, guarante, mortgage, lend, borrow, home, incom, benefit, rate
Topic 7	display, interfac, imag, machin, plural, graphic, screen, configur, type, view
Topic 8	custom, servic, provid, investment, option, allocation, analyze, monetary, bill, databas

수 있다. 그 결과 Topic 1 은 ATM(Automated banking), Topic 2는 거래 및 교환, Topic 3은 NFC(Near field communication), Topic 4는 금융 데이터 관리, Topic 5는 금융 소프트웨어, Topic 6은 주택 담보 대출, Topic 7은 디스플레이, Topic 8 자산관리, Topic 9는 보안, Topic 10은 인터넷 전문 은행, Topic 11은 경매 및 입찰, Topic 12는 재무 리스크 관리, Topic 13은 모바일 결제, Topic 14는 신용카드 결제, Topic 15는 금융데이터 분

"중요하지 않다"의 의미가 아니라, 상대적으로 특허 출원이 감소하고 있는 분야로 해석

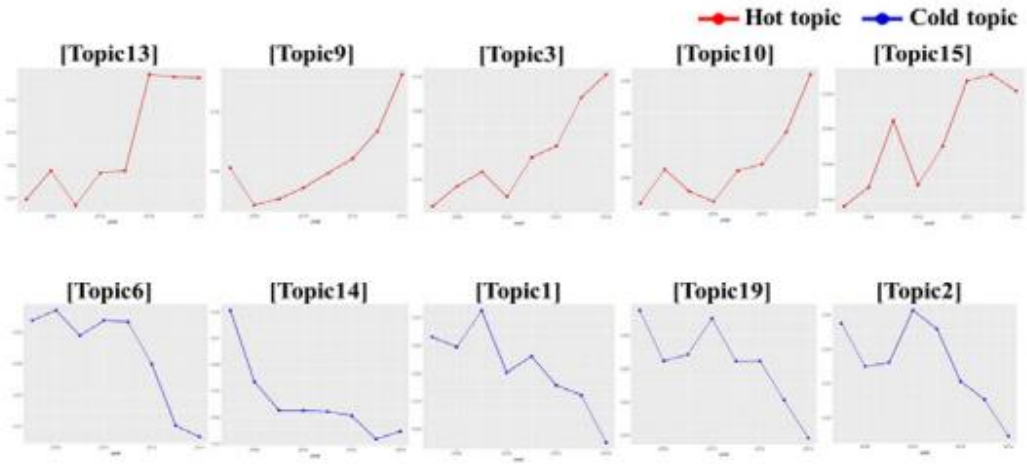


Fig. 4. Hot&Cold Topic of Total country

Source: 김태경, 최회련 and 이홍철. (2016). 토픽 모델링을 이용한 핀테크 기술 동향 분석. 한국산학기술학회논문지, 17(11), 670-681.

토픽 모델링이란?

토픽의 네이밍은 자동으로 주어지는 것이 아니라 연구자의 분석/판단에 따라 네이밍이 결정 ("단점")

토픽 네이밍을 잘 하려면 "domain knowledge"가 필요

Table 1 20 topics and keywords in topics

[T1] Robotics	[T2] Text/Word searching	[T3] Video	[T4] Programming	[T5] Computer management
sensor	user	display	code	product
power	search	frame	processor	task
control	context	motion	program	system
robot	input	screen	array	item
fluid	display	camera	element	tool
light	be	video	data	work
system	interest	color	address	computer
[T6] Programming syntax	[T7] SQL	[T8] Text preprocessing	[T9] Network speed adm.	[T10] General networking
first	data	document	time	network
second	system	text	level	node
portion	input	speech	real	neural
set	store	web	high	data
parameter	event	word	traffic	packet
section	be	file	line	protocol
end	set	search	speed	computer
[T11] VLSI	[T12] Network administration	[T13] Object recongition	[T14] Problem solving method	[T15] Multimedia
signal	server	object	model	content
input	agent	region	rule	media
unit	event	light	state	audio
output	client	subject	system	video
circuit	call	scene	process	user
control	request	interest	problem	music
pattern	session	detector	decision	stream

Source: 박주섭, 홍순구 and 김종원. (2017). 토픽모델링을 활용한 과학기술동향 및 예측에 관한 연구. 한국산업정보학회논문지, 22(4), 19-28.

토픽 모델링이란?

토픽 모델링 (Topic modeling)

- 방대한 양의 문서(비정형 데이터)에서 추상적인 토픽을 발견하기 위한 통계적 모델
- “특정 주제에 관한 문서에서는 특정 단어의 등장 빈도가 더 높을 것이다”라는 직관에 기반
- 문서에 숨겨진 의미구조를 발견하기 위한 방법 중 하나 → 여러 단어들의 동시 출현 횟수 및 확률이 중요
- 문서에서 발견되는 키워드 분포 분석을 통해, 문서들을 주제별로 분류가 가능 → 비지도 학습 알고리즘에 해당
- Clustering(군집화)의 경우 하나의 문서는 하나의 그룹에만 속할 수 있지만, 토픽 모델링 기법 중 LDA 방법을 적용하면, 하나의 주제가 여러 문서에 동시에 존재할 수 있는걸 나타낼 수 있음
- 주로 사용되는 토픽 모델링 기법으로는 LDA (latent dirichlet allocation, 잠재 디리클레 할당) 기법이 있음 (여러 기법이 있으나, 여기서는 LDA만 다룰 예정)

LDA

LDA (latent dirichlet allocation)은 대표적인 토픽 모델링 알고리즘 중 하나

- 주어진 문서에 대하여 각 문서에 어떤 토픽들이 어떤 분포로 존재하는지에 대한 확률 모형
- 문서내의 단어 (또는 키워드) 기반 토픽별 단어의 분포, 문서별 토픽의 분포 두 가지 모두 추정
- Latent → 문서내 여러가지 하위 주제들이 숨겨져 있다
- Dirichlet → 디리클레 분포 (단어가 한 주제에 속할 확률 값을 추정하는데 사용, 확률분포)
- Allocation → 할당 (문서에 주제를 할당, 단어를 주제에 할당)
- 하나의 문서는 여러 주제로 구성이 되어있고, 문서의 주제 분포에 따라 단어의 분포가 결정된다고 가정
  (Each document as a mixture of topics; each topic as a mixture of words)

LDA

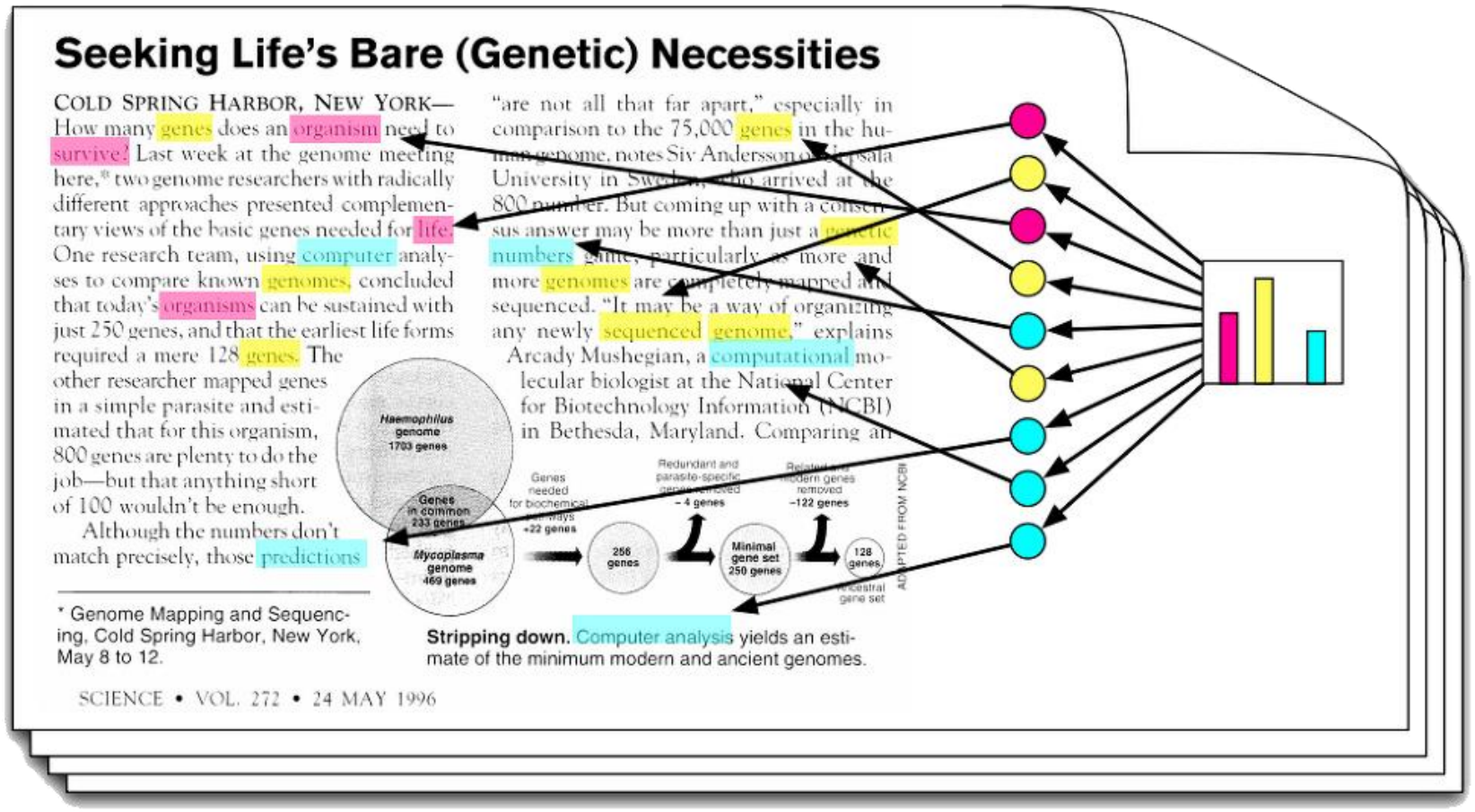
Topics *Documents* *Topic proportions & assignments*

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

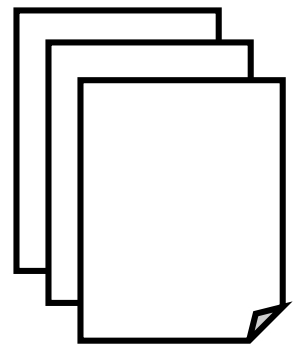
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

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

SCIENCE • VOL. 272 • 24 MAY 1996

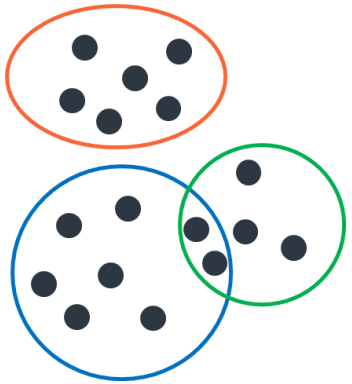
Source: http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf

LDA

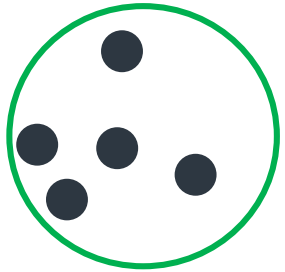


문서들의 집합

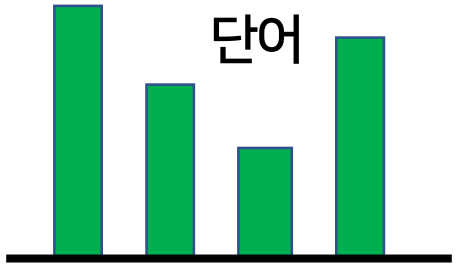
토픽 모델링



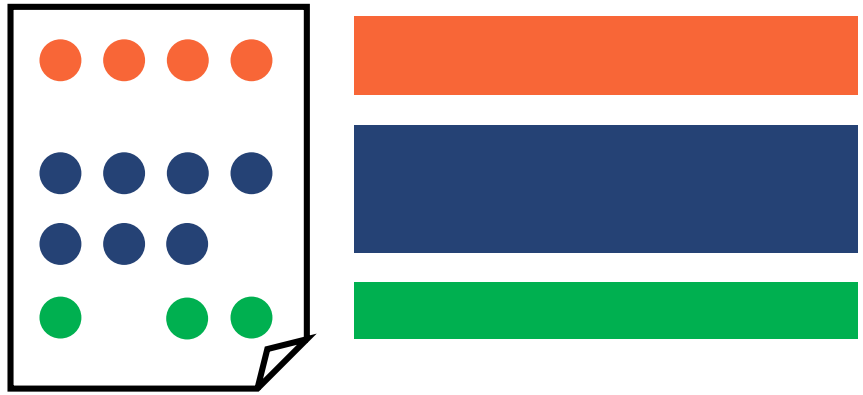
주제별 (k-topics)
단어들의 분포



주제



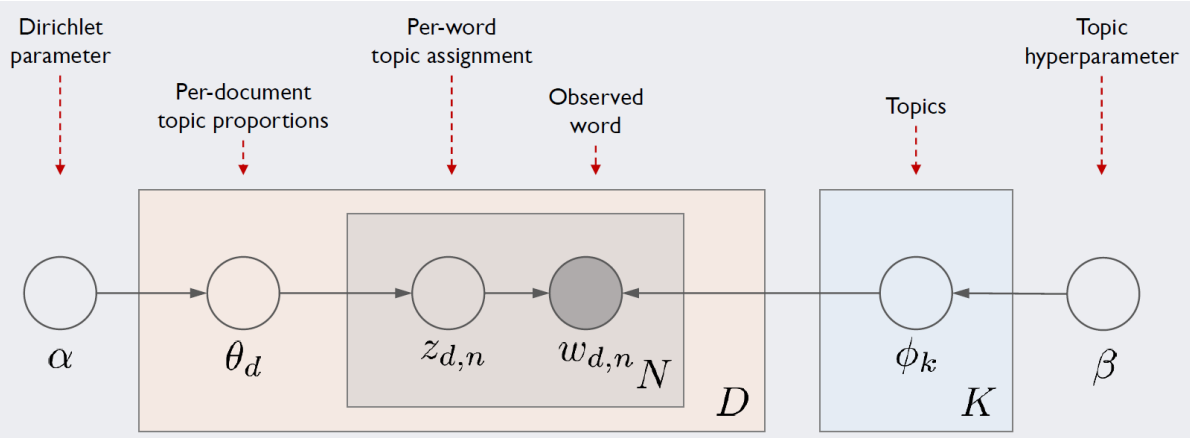
단어의 빈도



문서내 주제들의 분포

LDA의 원리

- 전체 문서들의 집합을 K 개의 주제로 표현이 가능하다고 가정할 때, 각각의 주제에 대한 단어 분포 (β_k)를 디리클레 분포 (β)로 정의
- 개별 주제의 비율을 디리클레 분포로 정의한 후, 각각의 단어에 대해 주제 비율로 주제를 배정하고, 개별 주제의 단어 분포에서 단어를 배정 (확률적으로 추론하는 기법, inference)



Graphical model representation of LDA

- Θ = the probability distribution of topics in documents (문서별 주제의 비율, 모두의 합 = 1)
- φ = the probability distribution of words in topics (각 토픽에서 어떤 단어들이 얼마나 분포하는지를 나타냄, 모두의 합 = 1)
- α, β = 하이퍼파라미터 (hyperparameter)
- α = 문서들의 토픽 분포를 얼마나 밀집되게 할 것인지에 대한 설정 값
- β = 주제내 단어들의 토픽 분포를 얼마나 밀집되게 할 것인지에 대한 설정 값

- D = 말뭉치 전체 문서 개수 (number of documents)
- N = d번째 문서의 단어 수
- K = 전체 토픽 수 (사용자가 토픽 개수를 지정, 통계적 기법 또는 직관에 의해 결정)
- $w_{d,n}$ = d번째 문서에 등장한 n번째 단어, 관찰 가능한 변수
- $z_{d,n}$ = d번째 문서에 n번째 단어가 어떤 토픽에 해당하는지를 설명하는 값

Source: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf; https://ratsgo.github.io/from%20frequency%20to%20semantics/2017/06/01/LDA/

LDA의 원리

- LDA는 토픽의 단어분포와 문서의 토픽분포의 결합으로 문서 내 단어들이 생성된다고 가정
- LDA의 inference(추론)는 실제 관찰가능한 문서 내 단어를 가지고 우리가 알고 싶은 토픽의 단어분포, 문서의 토픽분포를 추정하는 과정입니다.

1번 문서 : 문고리, 거래

2번 문서 : 가방, 나눔, 문고리, 드림

3번 문서 : 비대면, 거래, 택배

2개의 토픽으로 고정

	1번문서		2번문서				3번문서		
단어	문고리	거래	가방	나눔	문고리	드림	비대면	거래	택배
주제	topic1	topic2	topic1	topic1	topic2	topic2	topic3	topic2	topic3

표1. 문서별 명사추출 결과

토픽-문서	1번문서	2번문서	3번문서
topic1	1.01	2.01	0.01
topic2	1.01	2.01	1.01
topic3	0.01	0.01	2.01

표2. 토픽별 문서 단어 분포 계산

파라미터 값인 α 를 0.01로 설정 후 더해준 값

Source: http://bigdata.emforce.co.kr/index.php/2020072401/

토픽-단어	문고리	거래	가방	나눔	드림	비대면	택배
topic1	1.001	0.001	1.001	1.001	0.001	0.001	0.001
topic2	1.001	2.001	0.001	0.001	1.001	0.001	0.001
topic3	0.001	0.001	0.001	0.001	0.001	1.001	1.001

표3. 토픽별 단어 분포 계산

파라미터 값인 β 를 0.001로 설정 후 더해준 값

	1번문서		2번문서				3번문서		
단어	문고리	거래	가방	나눔	문고리	드림	비대면	거래	택배
주제	미분류	topic2	topic1	topic1	topic2	topic2	topic3	topic2	topic3

표4. "문고리"단어 토픽 선별하기

'문고리'의 키워드를 topic1 ~ topic3에 확률을 계산

1번 문서 내 topic1이 있을 확률 : 1.01/3.03 = 0.333

(이 때 분모의 경우 1번 토픽 내 문서 분포 합입니다.)

1번 토픽 내 단어가 '문고리'일 확률 : 1.001/3.007 = 0.332

(이 때 분모의 경우 1번 토픽 내 키워드 분포 합입니다.)

마지막으로 1번 문서의 '문고리'가 topic1일 확률 0.333*0.332=0.110이 됩니다.

경청해 주셔서 감사합니다

This PPT-slide is only for educational purposes.
(이 PPT 슬라이드는 교육용입니다.)