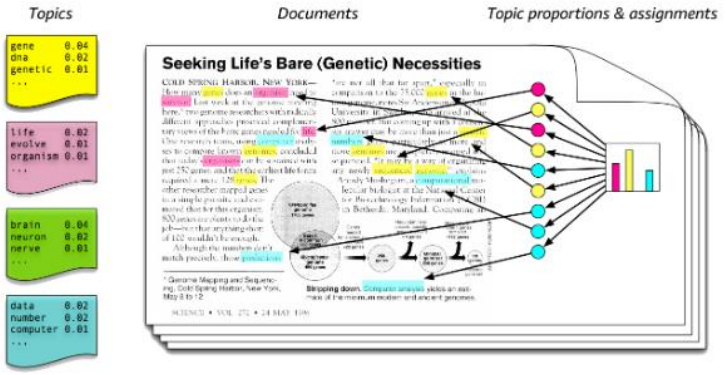
1. LDA?
2. 준비물
    - 베이지안
    - 깁스샘플링
3. 논문리뷰
4. 응용
    - 웹툰 댓글 리뷰
    - 반도체 불량 분석

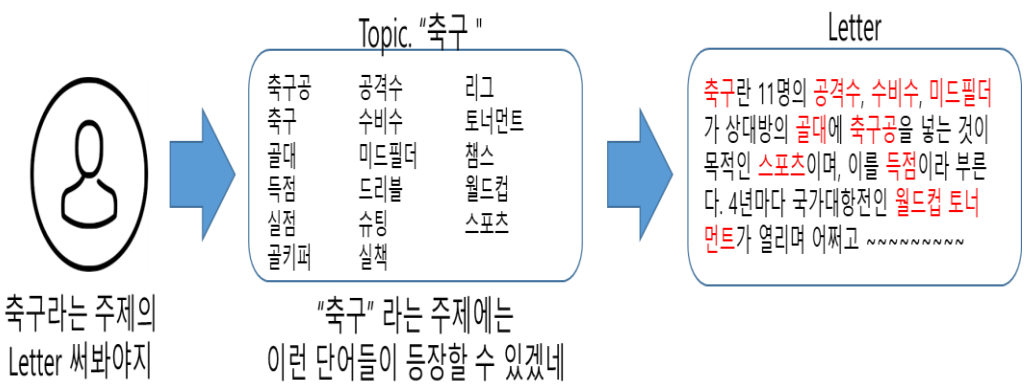
1. LDA



가정 1. 모든 문서는 다양한 주제에 대해 기술하고 있다
가정 2. 모든 주제는 고유의 단어 등장 확률을 가지고 있다

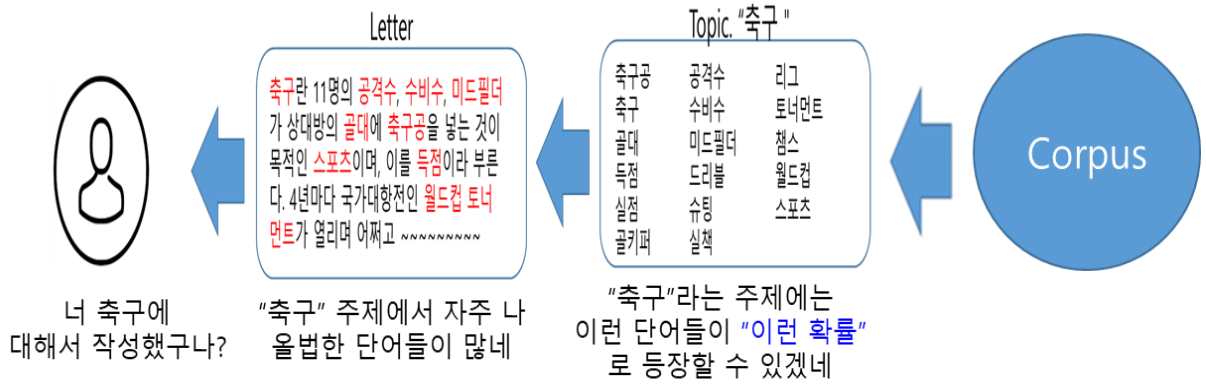
결과 1. 문서 별 주제 분포
결과 2. 주제 별 단어 분포

LDA 가 가정하는 Document 생성 과정



VS

LDA Inference 과정



2. 준비물 – 베이지안

베이지안 : 너희가 Constant라고 생각한 Parameter,,, 그거 사실 Random Variable 이야,,,

동전

동전을 N번 던져서 앞면이 나올 횟수 X,,,

1) 빈도주의 : P(앞면) = 0.5 니까,,, X ~ Bin(N, 0.5)
2) 베이지안 : P(앞면) = 0.5 라는걸 어떻게 믿어,,,?

베이즈정리

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- $P(A|B)$: 사후확률(posterior). 사건 B가 발생한 후 갱신된 사건 A의 확률
- $P(A)$: 사전확률(prior). 사건 B가 발생하기 전에 가지고 있던 사건 A의 확률
- $P(B|A)$: 가능도(likelihood). 사건 A가 발생한 경우 사건 B의 확률
- $P(B)$: 정규화 상수(normalizing constant) 또는 증거(evidence). 확률의 크기 조정

사전분포?

위에서 P(A)에 해당하며, 확률에 대한 추론 시, Beta 분포 자주 활용 (정의역 fit한 켤레분포)

$$X \sim Beta(\alpha, \beta)$$

$$f_X(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \qquad B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$(0 < x < 1,\ \alpha, \beta > 0)$$

디리클레분포?

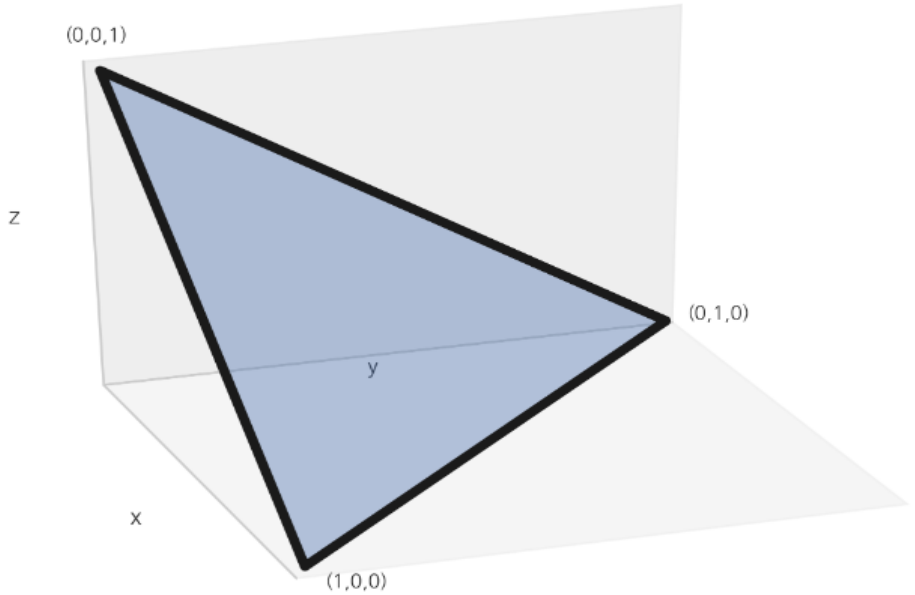
Beta dist' 가 Multinomial 하게 존재한다면??

K = 차원, 토픽의 수

$$f(x_1, \cdots, x_k; \alpha_1, \cdots, \alpha_k) = \frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}$$

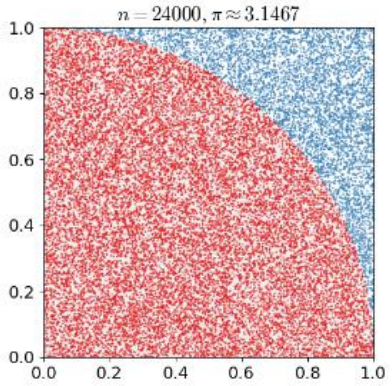
실수값 $x_1, \cdots, x_k$가 모두 양의 실수이며 $\sum_{i=1}^{k} x_i = 1$을 만족

$$\mathrm{B}(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)}$$ (Γ는 감마 함수)



2. 준비물 – 깁스샘플링

MC?



MC?

오늘/내일	맑음	흐림
맑음	0.7	0.3
흐림	0.25	0.75

오늘 날씨에 기준하여 내일 특정 날씨가 발생할 확률

MCMC?

MCMC란 마코프 연쇄에 기반한 확률 분포로부터 표본을 추출하는 몬테카를로 방법입니다. 다음과 같습니다.

> MCMC 알고리즘은 우리가 샘플을 얻으려고 하는 목표분포를 Stationary Distribution으로 가지는 마코프 체인을 만든다. 이 체인의 시뮬레이션을 가동하고 초기값에 영향을 받는 burn-in period를 지나고 나면 목표분포를 따르는 샘플이 만들어진다.

Gibbs sampling

- Joint probability $p(z_1, z_2, \ldots, z_n)$에서 샘플링할 때, 특정 하나의 확률변수(z_i)와 그것을 제외한 나머지 확률변수(z_{-i})를 conditional probability: $p(z_i \mid z_{-i})$로 샘플링하는 방법

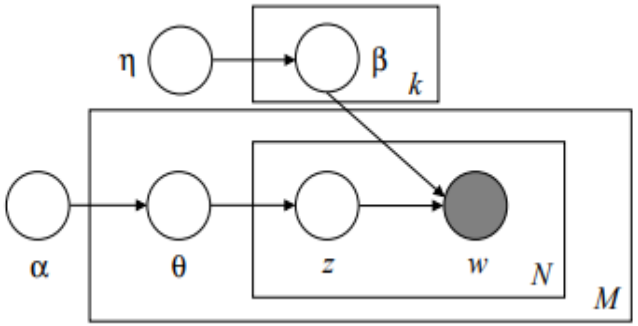
Ex) X = 첫 번째 굴린 주사위의 눈
Y = 주사위 2번 굴린 눈의 합

if Y <= 7, X ∈ {1,2,3,4,5,6}
Else if Y <= 8, X ∈ {2,3,4,5,6}
Else if Y <= 9, X ∈ {3,4,5,6}

우리의 Corpus에 등장하는 단어의 개수가 사후확률을 추정할 때 사용해야 하는 차원(Dimension)의 개수이기 때문에 Loading 너무 큼.

MCMC, 특히 깁스샘플링으로 이를 보완하려 함

3. 논문리뷰



The key inferential problem that we need to solve in order to use LDA is that of computing the posterior distribution of the hidden variables given a document:

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}.$$

Unfortunately, this distribution is intractable to compute in general. Indeed, to normalize the distribution we marginalize over the hidden variables and write Eq. (3) in terms of the model parameters:

관측 가능한 w / Hyperparameter alpha, beta
를 제외한 모든 변수는 unknown

EM

A : d번째 문서의 다른 단어들이 토픽 k에 많이 할당되어 있을 수록
B : 토픽 k에 할당된 전체 단어 중 Nd,k 의 점유율이 높을수록

$$p(z_{d,i} = j|z_{-i}, w) = \frac{n_{d,k} + \alpha_j}{\sum_{i=1}^{K}(n_{d,i} + \alpha_i)} \times \frac{v_{k,w_{d,n}} + \beta_{w_{d,n}}}{\sum_{j=1}^{V}(v_{k,j} + \beta_j)} = AB$$

위 수식의 표기를 정리한 표는 다음과 같습니다.

표기	내용
$n_{d,k}$	k번째 토픽에 할당된 d번째 문서의 단어 빈도
$v_{k,w_{d,n}}$	전체 말뭉치에서 k번째 토픽에 할당된 단어 $w_{d,n}$의 빈도
$w_{d,n}$	d번째 문서에 n번째로 등장한 단어
α	문서의 토픽 분포 생성을 위한 디리클레 분포 파라메터
β	토픽의 단어 분포 생성을 위한 디리클레 분포 파라메터
K	사용자가 지정하는 토픽 수
V	말뭉치에 등장하는 전체 단어 수
A	d번째 문서가 k번째 토픽과 맺고 있는 연관성 정도
B	d번째 문서의 n번째 단어($w_{d,n}$)가 k번째 토픽과 맺고 있는 연관성 정도

3. 논문리뷰

A : d번째 문서의 다른 단어들이 토픽 k에 많이 할당되어 있을 수록
B : 토픽 k에 할당된 전체 단어 중 Nd,k 의 점유율이 높을수록

$$p(z_{d,i}=j|z_{-i},w)=\frac{n_{d,k}+\alpha_j}{\sum_{i=1}^{K}(n_{d,i}+\alpha_i)}\times\frac{v_{k,w_{d,n}}+\beta_{w_{d,n}}}{\sum_{j=1}^{V}(v_{k,j}+\beta_j)}=AB$$

Doc1의 i번째 단어의 토픽 $z_{1,i}$	3	2	1	3	1
Doc1의 n번째 단어 $w_{1,n}$	Etruscan	trade	price	temple	market

말뭉치에 등장하는 모든 단어들을 대상으로 각각의 단어들이 어떤 토픽에 속해 있는지를 일일이 세어서 만든 표도 다음과 같다고 가정하겠습니다(이 표 역시 초기엔 랜덤 할당을 합니다)

구분	Topic1	Topic2	Topic3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
...	...	...	...

이제 깁스 샘플링을 할 차례입니다. $p(z_{1,2})$를 구해 보겠습니다. $z_{1,-2}$, 즉 첫번째 문서의 두번째 단어의 토픽(z_2) 정보를 지운 상태에서, 나머지 단어들의 토픽 할당 정보를 활용해 계산하게 됩니다.

$z_{1,i}$	3	?	1	3	1
$w_{d,n}$	Etruscan	trade	price	temple	market

이번엔 수식의 오른쪽 부분인 B를 보겠습니다. 전체 말뭉치를 대상으로 단어들의 토픽 할당 정보를 조사한 표는 다음과 같다고 칩시다. 여기서 주의할 점은 깁스 샘플링을 수행하면서 trade의 토픽 정보를 지웠으므로 단어별 토픽 분포 표에서 $v_{2,trade}$가 1이 줄어든다는 사실입니다.

구분	Topic1	Topic2	Topic3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	**10**	**8-1**	**1**
...	...	...	...

어쨌든 위 표를 보면 우리의 관심인 trade라는 단어의 토픽은 Topic1일 가능성이 제일 높겠네요. 전체 말뭉치에서 trade의 토픽은 1번, 2번, 3번 순으로 많거든요. 바꿔 말해 $v_{1,trade}=10, v_{2,trade}=7, v_{3,trade}=1$입니다. B에 적용된 β 역시 샘플링 과정에서 바뀌는 과정이 아니므로 B의 크기는 v_k에 가장 많은 영향을 받을 겁니다. 이를 그림으로 나타내면 다음과 같습니다.

어쨌든 결과적으로 $z_{1,2}$이 Topic1에 할당됐다고 가정해 보겠습니다. 그러면 Doc1의 토픽 분포(θ_1)와 첫번째 토픽의 단어 분포(ϕ_1)가 각각 다음과 같이 바뀝니다.

$z_{1,i}$	3	1	1	3	1
$w_{d,n}$	Etruscan	trade	price	temple	market

구분	Topic1	Topic2	Topic3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	**10+1**	7	1
...	...	...	...

4. 응용 – 웹툰 댓글 분석

1) 데이터 수집

*2019-08-15기준

Train data

현재 연재중인 네이버 웹툰 댓글
4,782,553건

현재 연재중인 네이버 웹툰 최신화 ~ -5화
242 작품

Test data

2019 네이버 웹툰 최강자전 예선작 댓글
16,692건

2019 네이버 웹툰 최강자전 예선작 1화
105 작품

1) 텍스트 분석과정 개요

1 댓글 필터링
- 자/모음, 영어, 숫자, 특수문자 제거
- 6글자 이상 필터링
- 웹툰 별 주인공 이름 레이블링
  댓글 4,782,553건 -> 3,819,687건

2 전처리
- PyKoSpacing으로 띄어쓰기 수정 *
- 원형 복원(Lemmatization)
- Okt와 MeCab 활용 형태소 분석*
- 불용어 처리

3 토픽 모델링
- LDA
- 정확도를 위해 Okt와 Mecab에서 중복되는 명사만 사용

4 감성분석
- 감성사전 구축
- 댓글(좋아요, 싫어요) 가중치로 사용

'신의 탑' 밤 -> 주인공

Raw data

lchi****,lchinin3,2019-08-14,신의 탑,1,1,진짜 그저 감탄만 번째정주행,2,0
icec****,응아니야,2019-08-14,신의 탑,1,1,프롤로그 의문점하츠명예아낙복수쿤권력라
gimg****,김규리,2019-08-14,신의 탑,1,1,아 정주행 번째 마약웹툰,1,0
ycr1****,BLDoo,2019-08-14,신의 탑,1,1,-,0,0
ckd0****,창준,2019-08-14,신의 팁,1,1,핸드폰 바꾸니깐 봤다는 표시가 없어져서 다시
mmal****,아네미,2019-08-14,신의 탑,1,1,정주행 원 전이신탑? 시작하며 정독은 기본
sanl****,이신,2019-08-14,신의 탑,1,1,쏘오오오오보러오미,2,0
wiw6****,wiw6****,2019-08-13,신의 팁,1,1,첫번째 베댓 마춘삡 넘이 거슬리네요,8,0
jsh0****,아아브,2019-08-13,신의 팁,1,1,내 여신도 서있나,5,0
hw10****,최흰웅,2019-08-13,신의 탑,1,1,번째다시봐도 갓작이다,4,0
sm91****,sm91****,2019-08-13,신의 탑,1,1,베댓 맞춤법 거슬리네,3,0
ldw2****,dlehdnjs,2019-08-13,신의 탑,1,1,정주행 한번하는데 일은 걸리는듯,0,2
kimj****,김준서,2019-08-13,신의 팁,1,1,아 이제 이거 정주행해볼까,1,0
2113****,KON,2019-08-13,신의 탑,1,1,와 댓글 만개 ,2,0
qkrd****,박이레,2019-08-13,신의 탑,1,1,번째 정주행,0,0
rkdj****,난찍그리고웨너블,2019-08-13,신의 탑,1,2,뱃댓 수인공도 빌지미,0,0
ciwo****,김시윤,2019-08-13,신의 팁,1,2,난 난,0,0
yhk2****,여히,2019-08-12,신의 탑,1,2,너명안,2,0

4. 응용 – 웹툰 댓글 분석

3) LDA : Topic & Keywords

Topic 1 (10.6%)

"또 대작 타는 냄새 드립이 난무하겠지만 이건 대작 같아요 다음 내용이 궁금해요"

기대감

Topic 2 (54.7%)

"작가님 제발 그림작가 구하시면 안될까요 그림작가만 있으면 웹툰 세계 정복 가능 "

그림체, 소재

3) LDA : Topic & Keywords

Topic 3 (6.9%)

"저 누렁머리 남자 여주 좋아하면 양다리 되던가 저 파랑머리가 여주 좋아하면 그럼 걍 사귀는 거지 후후후"

몰입도

Topic 4 (18.3%)

"스토리 좋고 그림체 좋고 분량도 적당한데 왜 별점이 정도지 신작이라고 별점테러 하지 좀 말자 진짜 골보기 싫다 "

스토리

Topic 5 (9.5%)

"이거 웹툰 공모전 출전했던 작품이네 게속 이 작품만 투표했었는데 네이버 정식연재로 또 뵙게 되네요 축하드려요 얼른 위로 올라가세요 "

To. 작가

스코어링 : 텍스트 점수, 이미지 점수

Title	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Color_Score	Score
신의 탑	97	95	84	81	79	72	508
복학왕	51	65	58	57	55	66	352
연애혁명	81	82	61	55	82	81	442
용이산다	97	81	84	81	79	55	477
열렙전사	88	80	84	81	79	33	445
뷰티풀 군바리	81	88	76	77	91	78	491
마음의 소리	97	95	84	81	79	69	505
유미의 세포들	97	95	84	81	79	50	486
기기괴괴	97	95	84	81	79	26	462

2019 네이버 웹툰 최강자전 본선 진출작 예측

name	실제 순위	지표 순위
오늘 죽는 너에게	1	1
오로지 오로라	2	2
하루만 네가 되고싶어	5	3
아침을 지나 밤으로	7	4
하늘은 왜 파랄까	4	5
왕년엔 용사님	9	6
⋮	⋮	⋮
붓꽃 예술 고등학교	10	15
아르테미스 신드롬	19	16
⋮	⋮	⋮
빤타스틱 스위퍼	40	32

순위상관계수 0.759

16강 진출작품 $\frac{15}{16}$ 일치

32강 진출작품 $\frac{25}{32}$ 일치