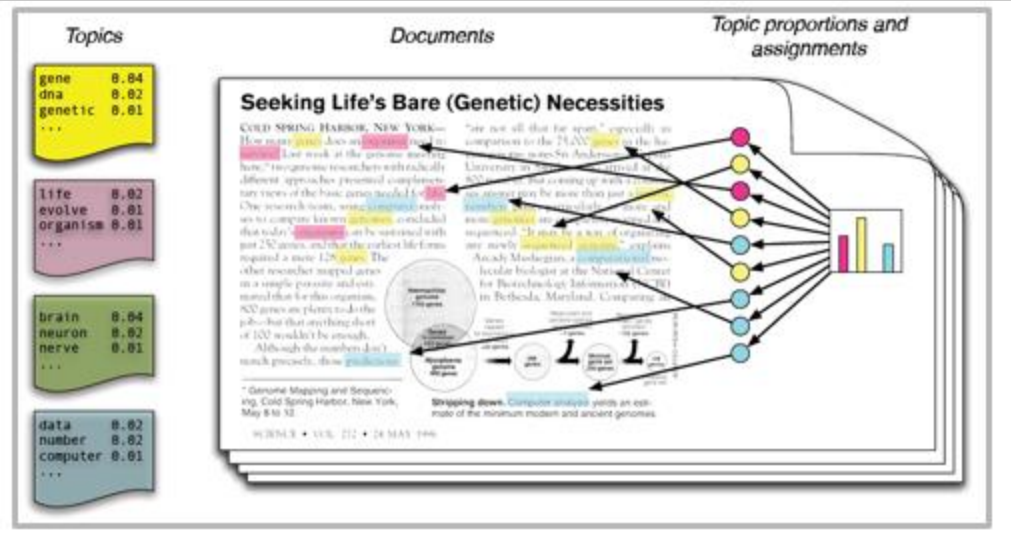
토픽 모델링

: Latent Dirichlet Allocation

LDA

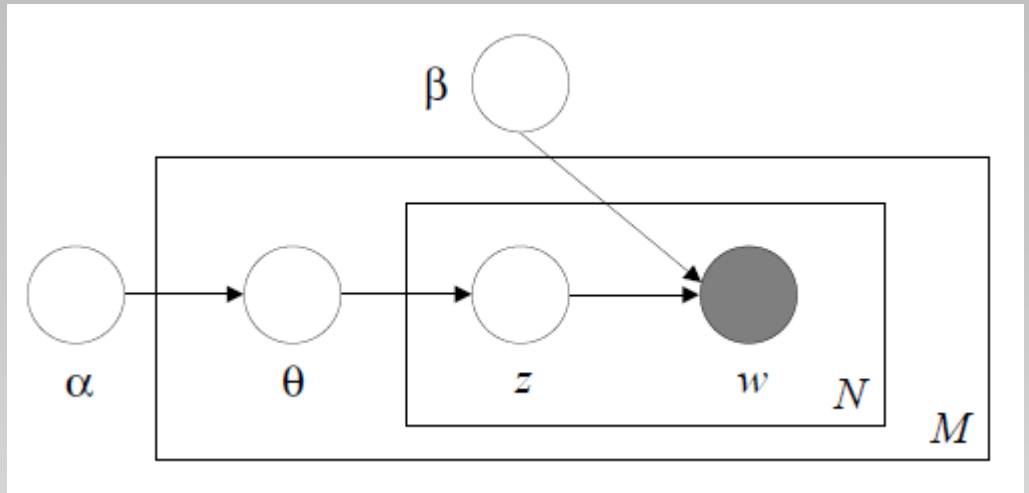
- 비지도학습 (unsupervised learning)
  - 단순 클러스터링
  - 기준 제시할 필요 없음
  - 추출할 주제의 수 지정



LDA

- 문서를 토픽의 집합으로 간주
  - 문서들은 특정 확률의 토픽들을 가지고 있다.
  - 토픽들은 단어 집합으로 구성되어 있다.
- Backtrack을 통해 문서들을 생성했을 법한 토픽들을 샘플링

LDA



- Alpha: 문서 안의 토픽 분포 parameter
- Beta: 토픽 안의 단어 분포 parameter
- Theta: 문서 안의 토픽 분포

- M: 문서 개수
- N: 단어 개수
- z: 토픽
- w: 단어

LDA sampling 문서탈트… 토픽탈트… 단어탈트…

- 토픽의 개수 K 지정
- 각 문서의 단어들에 임의로 K개의 토픽 중 하나 배정
- 각 토픽들에 대해 p(토픽|문서)와 p(단어|토픽) 계산
  - P(토픽|문서) = 전체 문서에서 특정 토픽으로 배정된 단어의 비율
  - P(단어|토픽) = 토픽으로 배정된 단어들 중 특정 단어의 비율
- 두 p의 곱을 가지고 토픽 재배정
  - P(토픽|문서) * p(단어|토픽) = 토픽이 단어를 생성했을 확률

	W1	W2	W3	Wn
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
Dn	1	1	3	0

LDA sampling

	Document X		Document Y
	Fish		Fish
	Fish		Fish
	Eat		Milk
	Eat		Kitten
	Vegetables		Kitten

	Document X		Document Y
F	Fish	?	Fish
F	Fish	F	Fish
F	Eat	F	Milk
F	Eat	P	Kitten
F	Vegetables	P	Kitten

	Document X		Document Y
F	**Fish**	?	Fish
F	**Fish**	**F**	**Fish**
F	Eat	F	Milk
F	Eat	P	Kitten
F	Vegetables	P	Kitten

	Document X		Document Y
F	Fish	?	Fish
F	Fish	**F**	**Fish**
F	Eat	**F**	**Milk**
F	Eat	**P**	**Kitten**
F	Vegetables	**P**	**Kitten**

- 얼마나 문서 안에서 많이 등장하는 토픽인지?
- 얼마나 토픽 안에서 많이 등장하는 단어인지?

LDA in Python

- pip install nltk
  - 아나콘다에 설치되어 있음
  - 자연어 처리를 지원하는 패키지
- pip install stop-words
  - 불용어(stop-words) 처리를 위한 패키지
- pip install genism
  - 파이썬에서 LDA 처리를 포함하고 있는 패키지

전처리

- Brocolli is good to eat. My brother likes to eat good broccoli, but not my mother.

- Tokenization: 문서 분할
  - brocolli / is / good / to / eat / my / brother / likes / to / eat / good / broccoli / but / not / my / mother
- Stopping: 불용어 제거
  - brocolli / good / eat / brother / likes / eat / good / brocolli / mother
- Stemming: 어간화 (동일한 의미의 단어 병합)
  - brocolli / good / eat / brother / like / eat / good / brocolli / mother

Document-term matrix

- dictionary: 각 단어에 index 부여
  - brocolli(0) / brother(1) / eat(2) / good(3) / like(4) / mother(5) / …
- doc2bow: (index, 출현 횟수) tuple 구성
  - (0, 2) / (1, 1) / (2, 2) / (3, 2) / (4, 1) / (5, 1)
  - tuple로 이루어진 vector

LDA 모델링

- LdaModel(corpus, num_topics=2, id2word = dictionary, passes=20)
  - num_topics: 추출할 토픽 개수
  - passes: 모델링 반복 횟수
- print_topics
  - 0.086*health + 0.086*brocolli + 0.086*good + 0.061*eat

꿋