알고리즘 마케팅 9강

2023. 4. 20. (목)

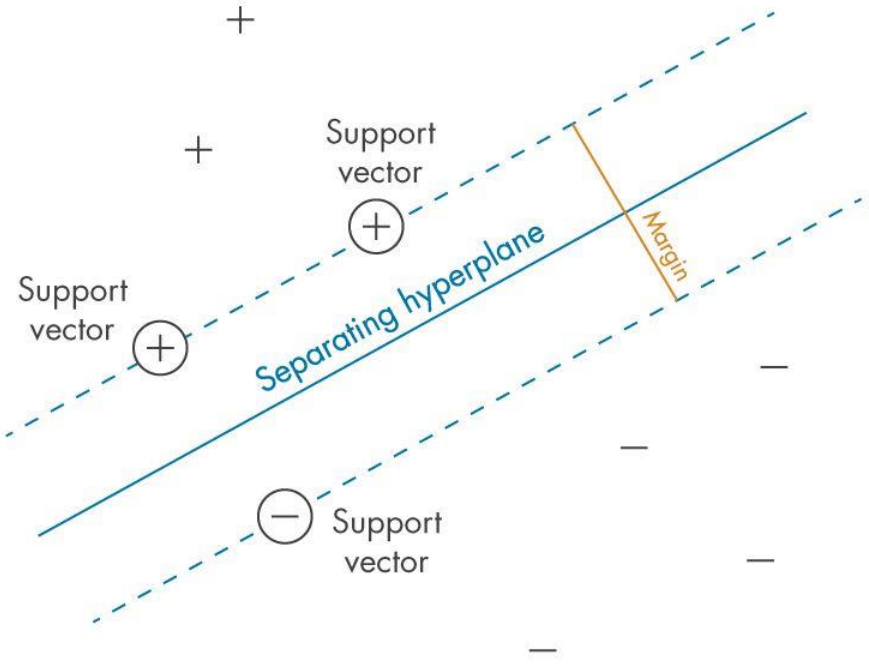
서울과학기술대학교 데이터사이언스학과
김 종 대

오늘의 강의

- 8주차 Review
- 고객 여정
- 온라인 광고
- 검색과 마케팅
- 베이지안 통계학의 기초
- 텍스트 마이닝의 기초

머신 러닝: 지도학습

- **커널 기법 추정**
  - 비선형 모형의 하나
  - 특징 공간(feature space)들을 더 높은 차원의 특징 공간으로 변환시키는 방식으로 하나 또는 그 이상의 기존 특징의 비선형 함수로 표현되는 차원을 추가
  - 분류 문제 등에 있어 특징(feature)의 조합에 관하여 상당한 유연성을 제공
  - 예: 서포트 벡터 머신 (Support vector machine; SVM)



머신 러닝: 표현학습

- **클러스터링 분석**
  - K-평균 알고리즘 (K-means algorithm)
    - 주어진 데이터를 k개의 클러스터로 묶는 알고리즘
    - K: 데이터셋으로부터 찾을 것으로 예상되는 그룹(클러스터)의 수
    - Means: 각 데이터 포인트와 그 데이터가 속한 클러스터의 중심까지의 평균 거리
      - 기본적으로 평균 거리를 최소화하는 것이 목표
    - 목표함수의 예:

$$V = \sum_{i=1}^{k} \sum | x_j - \mu_i |^2$$

머신 러닝: 표현학습

- **중요 요소 분석(Principal Component Analysis; PCA)**
  - 데이터의 압축되고 독립된 표현 방식을 찾는 방법
  - 특정 성질을 가지도록 데이터를 변형시켜 데이터의 구조를 설명하는 방법

  - 각 변수가 정규 분포를 따르고, “무상관관계”가 통계적 독립을 의미한다는 가정 하
  - 설명변수로 구성된 디자인 행렬(Design matrix; X)의 선형 변환을 통해 서로 독립인 요소로 구성된 대각 행렬(Diagonal matrix)을 도출
    - 예: 특이값 분해 (단일 가치 분해; Singular value decomposition)

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

머신 러닝: 표현학습

- **특이값 분해 (Singular Value Decomposition; SVD)**

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

  - Σ: m×n 대각 행렬(diagonal matrix)

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{pmatrix}$$

    - σ_i: 대각 행렬의 각 원소, 즉 대각 원소가 행렬 A의 특이값(singular value)
    - 특이값은 해당 차원의 정보량으로 해석할 수 있음.
    - 특이값은 큰 값 순서대로 내림차순으로 정렬됨.

선택 모형(Choice Model)

- **고객 선택 이론의 기초**
  - 소비자에게 주어진 옵션의 효용에 비례하는 가상의 수리적 지표를 구성
    - 예: $Y_{nj} = Y(x_{nj}, h_{nj})$
      - x_{nj}: 알려진 요소, 혹은 데이터로 수집할 수 있는 요소 / h_{nj}: 미지의 요소
      - 대안: $Y_{nj} = Y(x_{nj}) + \varepsilon_{nj}$
  - 랜덤 효용 모형(Random utility model)으로서 확률적으로 표현 가능
    - $P_{ni} = \Pr(Y_{ni} > Y_{nj}, \forall j \neq i) = \Pr(Y(x_{ni}) + \varepsilon_{ni} > Y(x_{nj}) + \varepsilon_{nj}, \forall j \neq i)$

타겟팅(Targeting)

- **RFM 모델**
  - RFM(= Recency + Frequency + Monetary)을 기준으로 고객을 세분화
    - 최근성(recency): 고객이 마지막으로 구매한 이후로 지난 시간
    - 빈도(frequency): 특정 시간 단위당 구매한 평균 횟수
    - 구매 금액(monetary): 특정 시간 단위당 구매한 전체 금액

    - 고객의 구매 히스토리를 3개 지표에 대해 점수를 부여함으로써 요약
    - RFM은 고객의 구매/응답 확률이나 고객생애가치와 상관관계가 있는 것으로 알려져 있음.

타겟팅(Targeting)

- **고객생애가치모델**
  - 고객생애가치(CLV; Customer Lifetime Value): 브랜드 또는 제품이 고객이 이를 사용하는 생애 동안 얼마나 고객으로부터 수입을 올릴 수 있는지를 추정한 값
  - 각종 마케팅 액션의 단기적인 이익을 넘어 장기적인 효과를 추정할 때 유용하게 활용 가능
  - 기본 공식 = 미래의 특정 기간 동안 평균 기대 이익의 총합

$$CLV(u) = \sum_{t=1}^{T} \frac{(R - C)r^{t-1}}{(1 + d)^{t-1}}$$

타겟팅(Targeting)

- **생존 분석을 이용한 타겟팅**
  - 종속 변수 혹은 관심 변수가 시간에 대한 변수일 때 유용하게 활용 가능
    - 예: 구매 확률보다 구매까지 걸리는 시간이 더 중요한 경우 (고객이 할인을 받으면 10일 내에 구매할 것을 5일 이내에 구매할 것으로 바뀔 것 vs 고객이 할인을 받으면 구매할 확률이 80%)
  - 생존 분석 모형을 통해 이벤트(예: 구매)까지의 시간을 예측 가능
  - 어떤 마케팅 액션과 고객 특성이 이벤트를 가속화 또는 감속화할 수 있는지 파악 가능

생존 분석

- **생존 분석 (Survival Analysis)**
  - 콕스 비례 위험 모형(Cox proportional hazard model): 위험 함수를 이용한 기본적인 회귀분석 방법
    - 기본적으로 설명변수 x가 위험 함수(생존 함수)에 어떤 영향을 미치는지 추정

$$h(t|w,x) = h_0(t)\exp(w'x)$$

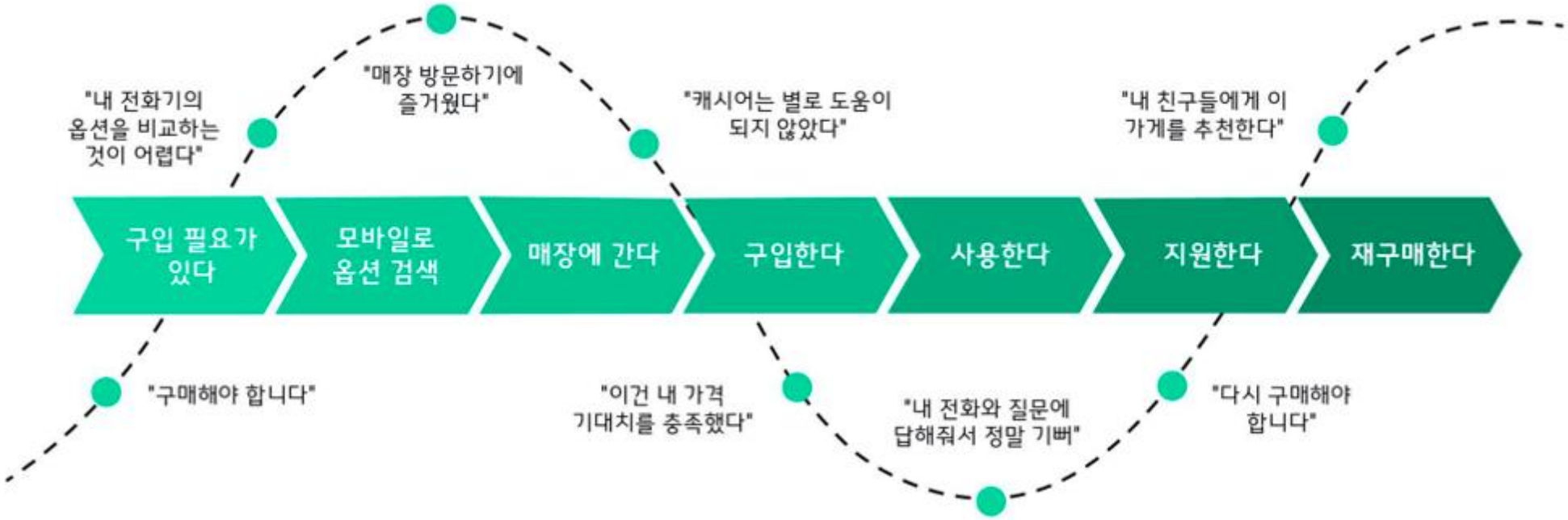
  - 기본 위험(baseline hazard; $h_0(t)$)
    - 확률 분포를 부여하고 모수적으로 추정하거나, 아예 비모수적으로 추정하는 것도 가능
    - 기본 위험으로부터 $\exp(w'x)$만큼 비례하여 증가하는 구조

고객 여정 (Customer Journey)

- **고객 여정**
  - 고객과 브랜드와의 상호 작용은 전체 구매 금액, 구매된 제품, 서비스, 마진, 웹 사이트 클릭, 광고 노출 등과 같은 형태로 표현되는 모든 종류의 거래(transaction)의 집합으로 정의 가능
  - 즉, 마케팅 최적화는 이러한 고객 여정(customer journey) 관점에서의 최적화로 해석 가능

  - 고객 여정은 매우 복잡하고 고객 행동과 의사 결정 등 매우 다양한 변수를 고려해야 모형화 가능
  - 온라인 로그 데이터 외의 구매 행동에 대해 데이터 수집이 어려운 상태

고객 여정 (Customer Journey)



온라인 광고

- **온라인 광고의 환경**
  - 수천 개의 회사가 동일한 시장 안에서 광고 캠페인을 수행하고 자동화된 프로세스를 활용하며 광고 캠페인의 품질과 효율성을 통제하고 측정

  - 광고 거래소: 광고 수요가 발생하면 퍼블리셔로부터 광고 요청을 받고 이를 광고주들에게 분배
    - 예: 실시간 입찰(real time bidding)
    - 예: 액션당 가격(Cost per action) 또는 노출당 가격(Cost per mile)

온라인 광고

- **온라인 광고의 비즈니스 목표에 관한 지표**
  - 신규 고객당 비용 (Cost per acquisition; CPA)
  - 광고 노출 후 액션 (Post-view action)
  - 클릭당 비용 (Cost per click; CPC)

  - 전환율 (Conversion rate): $R = \frac{k}{n}$
  - 업리프트 (Uplift): $L = \frac{R}{R_0} - 1$
    - 마케팅 캠페인의 효율성을 측정하기 위한 지표로서, Control 집단과 Treatment 집단 간의 전환율 차이로 정의

검색

- **검색**
  - 목적: 검색 창이나 선택된 필터로 표현되는 고객의 검색 의도에 따른 결과물을 제시
  - 개인화된 솔루션을 제공한다는 점에서 추천 서비스와도 유사한 점이 많음.

- **검색의 비즈니스 목표**
  - 일반적 목표: 사용자의 의도를 이해하고 그 의도에 적합한 결과를 전달하는 것
  - 1. 적합성: 고객의 의도와 일치하는 검색 결과를 얼마나 정확하게 제시하는가?
  - 2. 상품 통제: 마진에 따른 순위 조정 또는 고객 프로파일에 따른 필터링
  - 3. 검색 성능 지표: 컨버전 비율, 검색어 수정 비율, 페이지 비율, 검색 대기 시간 등

검색

- **검색의 기본: 매칭과 랭킹**
  - 검색 적합도 문제: 적합한 아이템과 그렇지 않은 아이템 간의 분류 문제와 유사
    - 동시에 텍스트 데이터와 랭킹 데이터에 집중하기 때문에 특이한 형태의 분류 문제
  - 기본적으로 “질의”(예: 검색어)와 “검색 결과”(예: 아이템) 간의 스코어링 또는 유사도 측정 문제

  - 토큰 매칭
    - 토큰(token): 텍스트의 토큰화(tokenization) 결과로서 “의미”를 가진 “단어” 단위
    - “질의”에 담긴 토큰과 “아이템”에 담긴 토큰(예: 드레스, 빨강, 리본 등)의 매칭
    - 단어 하나가 아닌 구절로 매칭하거나, 스테밍(stemming)을 통해 형태소 단위로 매칭 가능

검색

- **검색의 기본: 매칭과 랭킹**
  - 매칭을 통해 “검색 결과”의 집합을 찾아냈다면, 순서를 어떻게 정렬하여 제시할 것인가?
  - 랭킹: 좀 더 적합한 아이템을 상위 검색 결과에 노출시키는 방법
    - 벡터 유사도(vector similarity): “질의”와 “아이템”을 각각 벡터화하여 유사도 계산
    - TF-IDF 스코어링
    - N-gram 스코어링

검색

- **Online Search Model in Marketing**
  - Kim, J. B., Albuquerque, P., & Bronnenberg, B. J. (2010). Online demand under limited consumer search. Marketing science, 29(6), 1001-1023.
  - Kim, J. B., Albuquerque, P., & Bronnenberg, B. J. (2011). Mapping online consumer search. Journal of Marketing research, 48(1), 13-27.
  - Bronnenberg, B. J., Kim, J. B., & Mela, C. F. (2016). Zooming in on choice: How do consumers search for cameras online?. Marketing science, 35(5), 693-712.

Bayesian Statistics

- **Textbook: Hoff, P. D. (2009). A first course in Bayesian statistical methods (Vol. 580). New York: Springer.**
  - 베이지안 통계학의 이론적 입문으로서 적절한 교재
  - 사회과학 대학원 수준 또는 통계학 학부 수준

- **핵심 개념**
  - Bayes' rule
  - **Prior distribution $\times$ Sampling distribution $\propto$ Posterior distribution**
  - To calculate posterior distribution, what do we need to know?

Bayesian Statistics

- **Why Bayesian statistics?**
  - To express our information and beliefs about unknown quantities
  - Bayes' rule provides a rational method for updating beliefs in light of new information.

  - Bayesian methods provide:
    - Parameter estimates with good statistical properties (e.g. distribution information of parameter set)
    - Parsimonious descriptions of observed data
    - Predictions for missing data and forecasts of future data
    - A computational framework for model estimation, selection and validation

Bayesian Statistics

- **Frequentist vs. Bayesian**
  - 빈도주의(frequentist): 오늘날 '일반적인' 통계학에 해당되는 접근 방식
    - 특정 사건(event)이 얼마나 빈번하게 반복되어 발생하는가를 관찰
    - 확률: 반복 실험에 의해 장기적으로 발생하는 사건의 빈도
    - 모수: 고정된 상수로서 제시
  - 베이지안(Bayesian)
    - 특정 지식에 대한 믿음(belief)이 새로운 데이터에 의해 업데이트됨에 따라 변화
    - 확률: 사건 발생에 대한 믿음의 정도 (반복 실험의 개념과 무관)
    - 모수: 상수가 아닌 확률 변수로서 분포가 존재

Bayesian Statistics

- **Bayes' rule**

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

  - $\Pr(A|B)$: B라는 조건 하에 A가 일어날 확률, 즉 조건부확률
  - 주로 조건부확률의 계산, 혹은 반대 방향으로의 조건부확률의 계산에 자주 사용
  - 베이지안 통계학의 원리를 응축하고 있는 중요한 정리

Bayesian Statistics

- **Bayes' rule**

$$\Pr(\theta|y) = \frac{\Pr(y|\theta)\Pr(\theta)}{\Pr(y)}$$

  - $\Pr(y|\theta)$: 특정 parameter set이 주어졌을 때 데이터 y가 생성될 확률을 반영한 모형 (sampling model)
  - $\Pr(\theta)$: 특정 parameter set에 대한 기존에 누적된 주관적 믿음, 즉 parameter에 대한 사전 확률(prior distribution)
  - $\Pr(\theta|y)$: 기존에 누적된 주관적 믿음에 새로운 데이터 y를 업데이트한 결과, 즉 베이지안 통계학에서 최종적으로 추정하고자 하는 사후 확률(posterior distribution)

  - $\Pr(y) = \int \Pr(y|\tilde{\theta}) \Pr(\tilde{\theta})\, d\tilde{\theta}$: 일종의 상수로서 일반적으로 numerical method를 통해 계산하여야 함.

Bayesian Statistics

- **Exchangeability**
  - Let $\Pr(y_1, \ldots, y_n)$ be the joint density of $Y_1, \ldots, Y_n$. If $\Pr(y_1, \ldots, y_n) = \Pr(y_{\pi_1}, \ldots, y_{\pi_n})$ for all permutations π of {1,…,n}, then $Y_1, \ldots, Y_n$ are exchangeable.
  - $Y_1, \ldots, Y_n$ are exchangeable if the subscript labels (the order of data) convey no information about the outcomes. If $\theta \sim \Pr(\theta)$ and $Y_1, \ldots, Y_n$ are conditionally i.i.d. given θ, then marginally (unconditionally on θ), $Y_1, \ldots, Y_n$ are exchangeable.
  - 빈도주의에서 자주 사용되는 i.i.d.(independent and identically distributed)에 비해 상대적으로 약한 가정
  - Bayesian inference의 핵심 theorem인 de Finetti's theorem으로 연결
    - Bayesian fashion으로 모형화할 경우 그러한 parameter set과 분포 정보가 존재할 것임을 증명

Bayesian Statistics

- **Conjugate prior (켤레 사전분포)**
  - If the posterior distribution is in the same probability distribution family as the prior probability distribution, the prior and posterior are then called *conjugate distributions*, and the prior is called a *conjugate prior* for the likelihood function (e.g. exponential family).
  - Posterior distribution이 closed form으로 계산되므로 numerical method를 사용할 필요가 없음.

- **Conjugate prior example: Binomial model**
  - Sampling model: binomial distribution
  - Prior distribution: beta distribution
  - Posterior distribution: beta distribution

Bayesian Statistics

- **Semiconjugate prior**
  - Conjugate prior가 존재하는 경우와 달리 posterior distribution의 analytic solution이 존재하진 않으나, full conditional distribution을 이용하여 sampling할 경우 numerical method가 상대적으로 용이
  - *Full conditional distribution*: conditional distribution of a parameter given everything else (including data)
    - For y ~ $\mathrm{Normal}(\mu, \sigma^2)$, $\Pr(\mu|\sigma^2, y_1, \ldots, y_n)$
  - 이러한 full conditional distribution이 closed form으로 존재하는 경우의 prior가 *semiconjugate prior*
  - Posterior approximation can be made with **the Gibbs sampler**, an iterative algorithm that constructs a dependent sequence of parameter values whose distribution converges to the target joint posterior distribution.

Bayesian Statistics

- **Semiconjugate prior example: Normal model**
  - Sampling model: normal distribution with (μ, σ^2)

    ⇒ σ^2의 값이 known일 경우, μ에 대해서는 conjugate prior

    ⇒ σ^2의 값이 unknown일 경우, σ^2에 대해서는 semiconjugate prior
  - Prior distribution: inverse-gamma distribution
  - Full conditional distribution: inverse-gamma distribution

$$\{\frac{1}{\sigma^2}|\mu, y_1, \dots, y_n\} \sim gamma(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2(\mu)}{2})$$

Bayesian Statistics

- **Nonconjugate prior**
  - Conjugate prior처럼 posterior distribution의 closed form이 존재하지 않고, semiconjugate prior처럼 full conditional distribution의 closed form도 존재하지 않는 경우
  - When conjugate or semiconjugate prior distributions are used, the posterior distribution can be approximated with the Monte Carlo method or the Gibbs sampler. In situation where a conjugate prior distribution in unavailable or undesirable, the full conditional distributions of the parameters do not have a standard form and the Gibbs sampler cannot be easily used.
  - We can use **the Metropolis-Hastings algorithm** as a generic method of approximating the posterior distribution corresponding to any combination of prior distribution and sampling model.

  - Example: generalized linear model (GLM), longitudinal regression model (with correlated error)

Bayesian Statistics

- **Hierarchical model using a Bayesian approach**
  - We can model additional layers by probability distributions for parameters.
  - Example: normal distribution with (μ, σ^2)
    - Assume that we need to propose a model to figure out differences between groups (j).
    - For given σ^2, we can model μ as $\Pr(\mu_j|\varphi) = Normal(\theta, \tau^2)$ where $\varphi = \{\theta, \tau^2\}$.
    - We can add another layers for $\varphi = \{\theta, \tau^2\}$.

Bayesian Statistics

- **Summary: Numerical methods for Bayesian Statistics**
  - Conjugate prior ⇒ We can easily get the analytic solution (Known form of probability distribution).
  - Semiconjugate prior ⇒ **Gibbs sampler** (using the closed form of full conditional distribution)
  - Nonconjugate prior ⇒ **Metropolis-Hastings algorithm**

  - In order to estimate latent Dirichlet allocation model, we usually use **Gibbs sampler** methods.

Bayesian Statistics

- **Gibbs sampler**
  - Gibbs sampling uses distribution information of full conditional distribution which is a conditional distribution of a parameter given everything else.
  - For normal distribution $x \sim N(\theta, \sigma^2)$, full conditional distributions are:
    - $\Pr(\theta|\sigma^2, y_1, \ldots, y_n)$ for mean parameter
    - $\Pr(\sigma^2|\theta, y_1, \ldots, y_n)$ for variance parameter
  - By using these full conditional distributions, we can make the iterative sampling idea.

Bayesian Statistics

- **Gibbs sampler**
  - Given a current (or initial) state of parameters $\phi^{(s)} = \{\theta^{(s)}, (1/\sigma)^{2(s)}\}$, we generate a new state as follows:
    - 1. Sample $\theta^{(s+1)} \sim \Pr(\theta|(1/\sigma)^{2(s)}, y_1, \ldots, y_n)$
    - 2. Sample $(1/\sigma)^{2(s+1)} \sim \Pr((1/\sigma)^2|\theta^{(s+1)}, y_1, \ldots, y_n)$
    - 3. Let $\phi^{(s+1)} = \{\theta^{(s+1)}, (1/\sigma)^{2(s+1)}\}$
  - Likewise, the Gibbs sampler generates a dependent sequence of parameters $\{\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(S)}\}$.
  - S개의 parameter sample들의 평균 등으로 parameter의 추정치를 계산
    - $E[\theta|y_1, \ldots, y_n] \approx \frac{1}{S}\sum_{s=1}^{S} \theta^{(s)}$

Bayesian Marketing

- **Allenby, G. M., & Rossi, P. E. (1998). Marketing models of consumer heterogeneity. Journal of econometrics, 89(1-2), 57-78.**
  - The distribution of consumer preferences plays a central role in many marketing activities.
  - Marketing activities which target specific households (individuals) require household (individual) level parameter estimates.
  - Thus, the modeling of consumer heterogeneity is the central focus of many statistical marketing applications.

Bayesian Marketing

- **Allenby, G. M., & Rossi, P. E. (1998). Marketing models of consumer heterogeneity. Journal of econometrics, 89(1-2), 57-78.**
  - A Bayesian approach to estimate α_i (individual level parameter) and β (parameters for the distribution of α_i) by maximizing the following equation

$$\Pr(\alpha_i|data) \propto \Pr(data_i|\alpha_i)\pi(\alpha_i|\beta = \hat{\beta})$$

  - Incorporating heterogeneity in choice models

$$\beta_i \sim iid\ N(\bar{\beta}, V_\beta)$$

$$\bar{\beta} \sim N\left(\bar{\beta}, aV_\beta\right), \qquad V_\beta^{-1} \sim W(v_0, V_0)$$

Text Data and Text Mining

- **비정형데이터**
  - 텍스트 데이터 속 정보를 합리적이고 효율적으로 축약하여 숫자, 즉 정형 데이터로 변환하는 것이 하나의 목적

- **텍스트 마이닝**
  - 어떤 방식으로 텍스트 속의 유의미한 패턴을 포착하여 효과적으로 축약하는지의 문제
  - 머신 러닝, 통계학에 대한 배경 지식 필요
  - **자연어 처리 기술(Natural Language Processing; NLP)**에 대한 배경 지식 필요

Text Data and Text Mining

- [참고] 언어학의 층위
  - 음운론(phonology): 언어의 소리 체계
  - 형태론(morphology): 단어의 구조 체계
  - 구문론(syntax): 문장의 구조 체계
  - 의미론(semantics): 문장의 의미 체계
  - 화용론(pragmatics): 발화의 맥락 체계

Text Data and Text Mining

- 단어의 사전적 의미를 제대로 포착할 수 있는가?
  - 예: 사전의 여러 가지 의미 중 어떤 의미로 간주되어야 하는가?

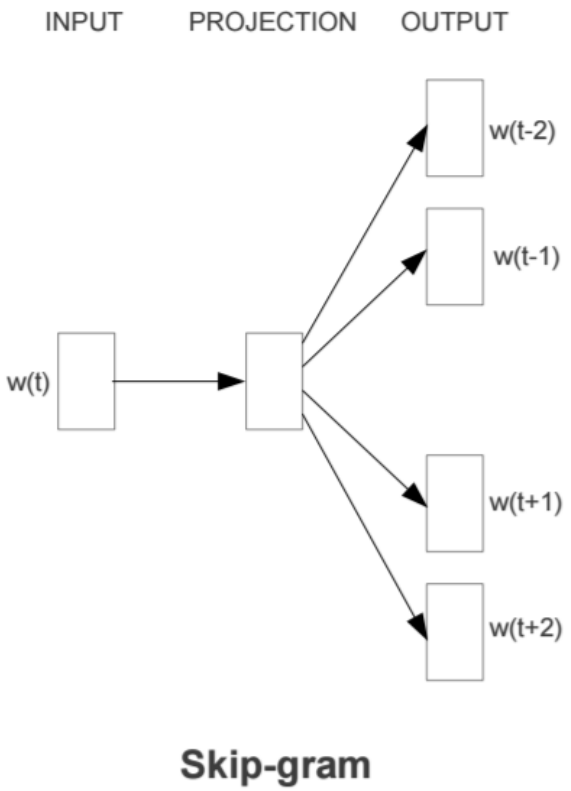
- 단어의 맥락적 의미를 제대로 포착할 수 있는가?
  - 예: 단어가 긍정적으로 사용되었는가, 부정적으로 사용되었는가?
  - 예: 특정 주제에서는 다른 의미로 사용되지는 않는가?

Key Steps for Text Mining

- **1. 데이터 준비**
  - API 호출, 웹 데이터 크롤링 등
- **2. 데이터 전처리**
  - 토큰화, 정규화, 불용어 제거 등 정제
  - 빈도 분석 등 기초 분석
- **3. 모형을 이용한 데이터 분석**
  - 국소 표현 기반의 모형 활용 (예: 토픽 모형)
  - 분산 표현 기반의 모형 활용 (예: 워드 임베딩)
- **4. 분석 결과 시각화 및 해석**

Key Models for Text Mining

- 텍스트 마이닝 모형에 대한 여러 분류 방식이 있겠으나, 여기에서는 단어 표현 방법에 따른 분류로 소개
- **단어의 표현 방법 (Word representation)**
  - **국소 표현 (Local representation)**: 해당 단어 자체로 값을 할당
    - 주로 사전에 정의한 의미 사전(dictionary)을 기반으로 값을 할당
  - **분산 표현 (Distributed representation)**: 단어를 표현할 때 함께 사용된 주변의 단어까지 참고하는 방식
    - 단어의 맥락적 의미나 뉘앙스를 비교적 잘 반영할 수 있다는 장점



Local Representation-Based Approach

- **Bag of Words**
  - 문서 내에 들어있는 단어를 하나의 가방(bag)에 전부 집어넣는 방식
  - 단어의 출현 빈도(frequency)에 주목하는 방식으로, 단어의 순서나 맥락을 반영하지 못함.
  - 한계: 단어의 맥락적 의미가 중요하거나, 한 문서 내에 여러 주제나 의견이 들어있을 경우 문서 내의 유의미한 의미를 제대로 포착하지 못할 수 있음.

- **Document-Term Matrix (DTM)**
  - Bag of words는 단어의 빈도 수 기반이므로 각 문서(행)와 각 단어(열)로 이루어진 행렬로 표현

Local Representation-Based Approach

- **Document-Term Matrix (DTM)**

	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	…
Document 1	1	3	0	0	0	0	1	…
Document 2	0	2	0	0	1	1	4	…
Document 3	3	1	0	1	0	1	0	…
Document 4	2	8	2	0	1	0	0	…
…	…	…	…	…	…	…	…	…

  - 각 문서를 단어 기반으로 수치화하여 비교할 수 있는 가장 간단한 텍스트 마이닝 중 하나
  - 불용어 등을 제거하여 좀 더 단순한 행렬로도 표현 가능

TF-IDF

- **TF-IDF (Term frequency-inverse document frequency)**
  - **단어의 빈도** + **역 문서 빈도**를 이용하여 문서 내 특정 단어의 중요도를 구하는 작업 등에 사용
  - 문서 집합 내 **핵심어 추출**이나 **문서 간 유사도**를 계산하는 데 응용 가능
  - 문서 내에 등장하는 모든 단어에 대해 빈도를 표시하는 것이 비효율적이라면, 중요한 단어에 대해 가중치를 줄 수 있음.

  - 역 문서 빈도(Inverse document frequency): 특정 단어 w가 등장한 문서의 수에 반비례하는 값
  - 원리: 불용어(the, it, as 등)는 거의 모든 문서에 자주 등장하게 됨. 즉, 일부 문서에만 등장하는 일종의 “희귀 단어”들이 의미 파악에 중요하고, 거의 모든 문서에 등장하는 단어들은 중요성이 떨어질 것이라고 가정하여 단어의 중요도 가중치를 조정하는 방법

TF-IDF

- **TF-IDF (Term frequency-inverse document frequency)**
  - 문서 d, 단어 w, 문서의 총 개수가 n일 때,
  - tf(d, w): 특정 문서 d에서 특정 단어 w가 등장하는 횟수 (예: DTM)
  - df(w): 특정 단어 w가 등장하는 문서의 수
  - idf(d, w, n): df(w)에 반비례하는 수, 즉 역 문서 빈도(Inverse document frequency)

$$idf(d,w) = \log(\frac{n}{1+df(w)})$$

  - TF-IDF: tfidf(d, w, n) = tf(d, w) × idf(d, w, n)
    - 특정 문서에서 단어 빈도가 높을수록, 그리고 단어가 등장하는 문서가 적을수록 값이 커짐.

Sentiment Analysis

- **감성 분석 (Sentiment analysis)**
  - 텍스트에 들어있는 작성자의 의견이나 평가 등 주관적인 정보를 추출
  - 해당 문서가 긍정(positive)인지 부정(negative)인지 판단하는 것이 대부분
  - **Dictionary-based Method**: 기존에 구축되어 있던 감성 사전을 이용한 분석
    - 한계 1: 단어의 맥락적 의미를 애초에 반영하지 못 한다.
    - 한계 2: 전혀 다른 주제나 분야일 경우, 단어의 의미가 아예 다르게 사용될 수 있다.
  - **Machine Learning-based Method**
    - 연구자의 데이터를 이용하여 감성 사전을 직접 구축 가능
    - Logistic classification부터 deep learning까지 다양한 모형 활용 가능

Local Representation-Based Approach

- **감성 분석 (Sentiment analysis)**
  - **감정 분석 (Emotion classification)**: 긍정 / 부정보다 더 세부적인 감정을 추출하는 방법
    - Label이 2~3개가 아닌 6~8개 혹은 그 이상으로 늘어난 경우
      - 예:{anger, happiness, neutral, surprise, sadness, fear, disgust}
    - Dictionary-based의 경우, 역시 좋은 사전의 구축이 핵심 이슈가 될 것임.

Local Representation-Based Approach

- **토픽 모형 (Topic Model)**
  - **잠재 의미 분석 (Latent Semantic Analysis; LSA)**
    - DTM 정보를 바탕으로 특이값 분해(singular value decomposition)를 이용하는 방식
  - **잠재 디리클레 할당 모형 (Latent Dirichlet Allocation; LDA)**
    - 특정 문서가 여러 토픽의 혼합으로 구성되어 있다고 가정
    - 문서의 집합을 구성하는 토픽의 잠재 구조를 추정하는 모형
    - 단어 빈도 기반의 bag of words를 가정하되, 베이지안 확률 모형을 기반으로 하여 잠재 토픽의 구조를 추정

Latent Dirichlet Allocation (LDA)

- **토픽 모형 (Topic Model)**
  - **잠재 디리클레 할당 모형 (Latent Dirichlet Allocation; LDA)**
    - 장점 1. 베이지안 확률 모형을 활용하기 때문에 확률 정보를 이용 가능
    - 장점 2. 문서의 집합을 구성하는 토픽의 수와 토픽별 의미를 추론할 수 있음.
    - 장점 3. 각 문서가 어떤 토픽으로 어떻게 구성되어 있는지 확률적으로 표현할 수 있음.

    - 단점 1. Bag of words 기반이기 때문에 단어의 맥락적 의미를 충분히 반영할 수 없음.
    - 단점 2. 토픽별 의미는 키워드를 바탕으로 연구자 등이 직접 해석해야 함.

Latent Dirichlet Allocation (LDA)

- **Basic idea**
  - 사람이 글을 쓸 때 어떤 과정을 거쳐 글을 쓰는가?
    - 1. 글로 옮기고자 하는 주제(topic)를 선정
    - 2. 주제와 관련이 있는 단어(word)들 결정
    - 3. 해당 단어들의 조합으로서 하나의 문서(document)가 탄생
  - LDA는 이 과정을 역순으로 따라 올라가는 방식으로 잠재된 주제를 추출
    - 1. 특정 문서 내 단어의 동시 출현 정보를 학습
    - 2. 자주 함께 출현하는 단어들은 특정 주제와 밀접한 연관이 있을 것으로 간주
    - 3. 잠재 토픽을 추출하고 토픽과 단어의 확률 분포를 계산

Latent Dirichlet Allocation (LDA)

- **Key assumptions**
  - 1. 하나의 문서(document)는 여러 가지의 주제가 공존하는 혼합체(mixture)다.
    - 즉, 각 문서는 토픽의 확률 분포로 표현할 수 있다.
  - 2. 주제마다 고유한 단어의 분포를 가지고 있다.
    - 즉, 토픽의 추출과 문서의 토픽별 할당은 단어의 분포에 의해 결정된다.
  - 3. 문서의 단어 분포는 각 토픽의 비율과 토픽별 단어의 분포에 따라 결정된다.
    - 즉, 토픽의 분포와 단어의 분포를 종합적으로 고려하여 추정하는 모형을 사용해야 한다.

Latent Dirichlet Allocation (LDA)

- **Definition**
  - v: word, w: document, D: corpus (i.e. a collection of documents)

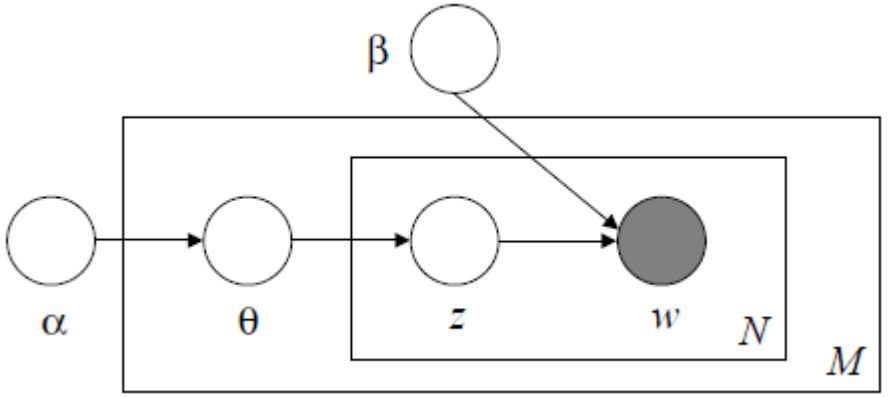
- **LDA as a Bayesian model**
  - Sampling model: multinomial distribution (for topics and words)
  - Prior distribution: Dirichlet distribution with dimension *k*

Latent Dirichlet Allocation (LDA)

- **Generative process**
  - For document w in a corpus D,
  - 1. Choose N (the number of words) ~ Poisson(ξ) (or choose N from data)
  - 2. Choose θ ~ Dirichlet(α)
  - 3. For each of the N words w_n:
    - (a) Choose a topic z_n ~ Multinomial(θ)
    - (b) Choose a word w_n from $\Pr(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Latent Dirichlet Allocation (LDA)

- **Model structure**



  - There are three levels to the LDA representation. The parameter α and β are corpus-level parameters, assumed to be sampled once in the process of generating a corpus. θs are document-level variables, sampled once per document. And z and w are word-level variables and are sampled once for each word in each document.

Latent Dirichlet Allocation (LDA)

- **Model specification**
  - k-dimensional Dirichlet distribution

$$p(\theta \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$$

    where $\Gamma(x)$ is the Gamma function.

  - k is the number of latent topics, which should be given before the estimation.
  - The Dirichlet is a convenient distribution because it is in the exponential family, has finite dimensional sufficient statistics, and is conjugate to the multinomial distribution.

Latent Dirichlet Allocation (LDA)

- **Model specification**
  - Given the parameters α and β, the joint distribution of topic mixture θ, a set of N topics z, and a set of N words w is given by:

$$p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^{N} p(z_n | \theta) p(w_n | z_n, \beta)$$

  - Integrating over θ and summing over z, we obtain the marginal distribution of a document:

$$p(\mathbf{w} | \alpha, \beta) = \int p(\theta | \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n | \theta) p(w_n | z_n, \beta) \right) d\theta$$

Latent Dirichlet Allocation (LDA)

- **Estimation**
  - 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:

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

$$p(\mathbf{w} \mid \alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta$$

  - Unfortunately, it is a function which is intractable due to the coupling between θ and β in the summation over latent topics.
  - Although the posterior distribution is intractable for exact inference, numerical approximate inference algorithms can be considered for LDA, including *Laplace approximation, variational approximation, and Markov chain Monte Carlo*.

Latent Dirichlet Allocation (LDA)

- **Perplexity**
  - They used this index to evaluate the models.
  - The perplexity, used by convention in language modeling, is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean per-word likelihood.
  - A lower perplexity score indicates better generalization performance.

$$perplexity(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\}$$

Latent Dirichlet Allocation (LDA)

- **Blei et al. (2003)'s applications and empirical results**
  - They performed experiments using a corpus of scientific abstracts from the C. Elegans community (Avery, 2002) containg 5,225 abstracts with 28,414 unique terms, and a subset of the TREC AP corpus containing 16,333 newswire articles with 23,075 unique terms.
  - They compared performances of several models such as LDA and pLSA.
  - To estimate the LDA model, they used the EM algorithm.
    - EM algorithm: (1) to calculate the Expectation of log-likelihood, (2) to calculate the Maximization of the Expectation value

Latent Dirichlet Allocation (LDA)

- **Blei et al. (2003)'s applications and empirical results**

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Latent Dirichlet Allocation (LDA)

- **Blei et al. (2003)'s applications and empirical results**

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Latent Dirichlet Allocation (LDA)

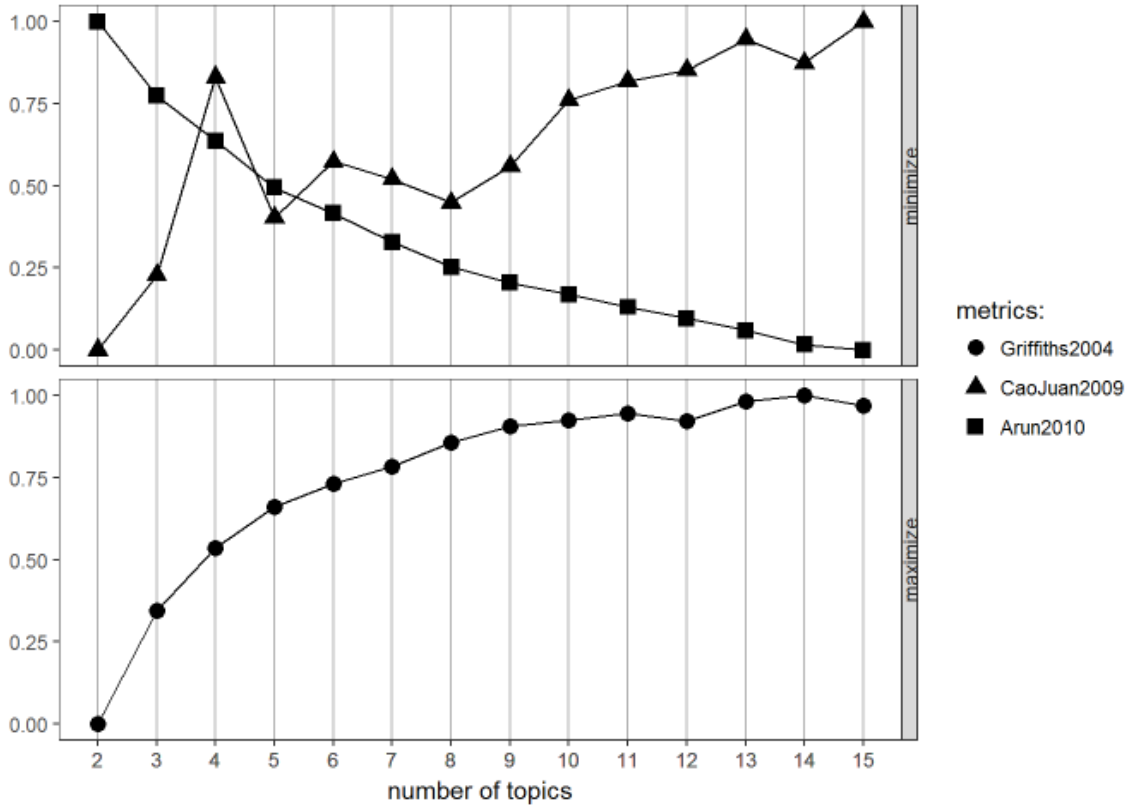
- **Model selection**
  - How do we choose the optimal number of topics (*k*)?
  - **1. Perplexity**
    - The most common way to compare the model performances in text mining
    - To measure how well a probability distribution predicts a sample (hold-out test set)
  - **2. Arun, R., Suresh, V., Veni Madhavan, C. E., & Murthy, N. (2010, June). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 391-402). Springer, Berlin, Heidelberg.**
    - The measure is computed in terms of symmetric KL-Divergence of salient distributions. The divergence values are higher for non-optimal number of topics.

Latent Dirichlet Allocation (LDA)

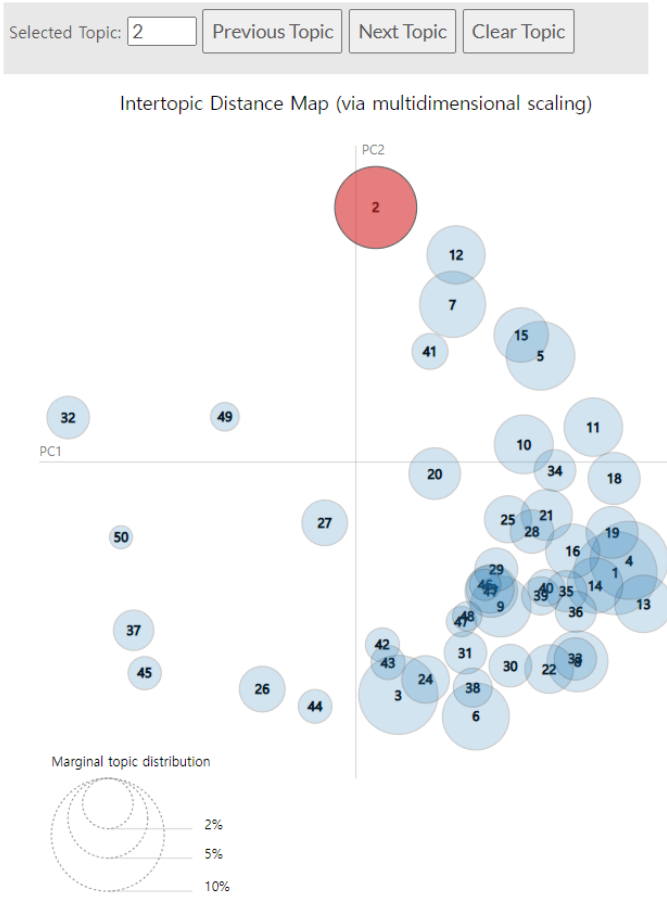
- **Model selection**
  - **3. Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7-9), 1775-1781.**
    - To adaptively choose the best LDA model based on topic density information
    - They compute the density of each topic, find the most unstable topics under the old structure, and iteratively update the parameter *k*.
  - **4. Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl 1), 5228-5235.**
    - To evaluate the results of changing the number of topics, by employing the Gibbs sampling algorithm.

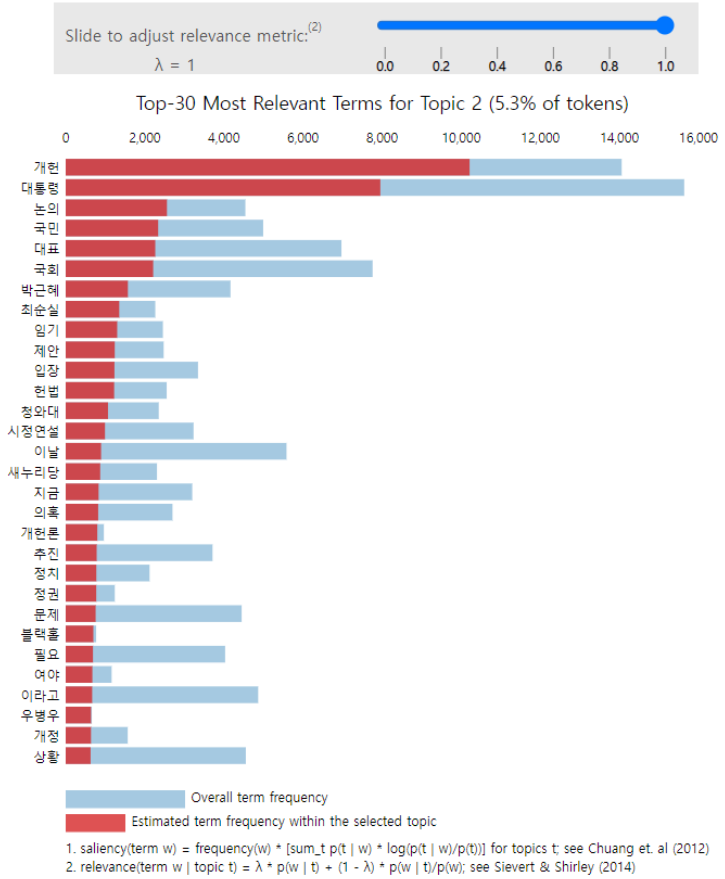
Latent Dirichlet Allocation (LDA)

- **Model selection**



Latent Dirichlet Allocation (LDA)





Local Representation-Based Approach

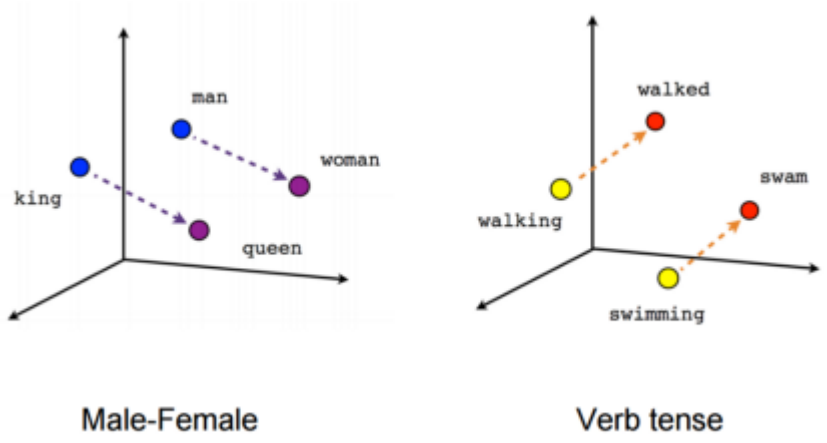
- **Joint Sentiment Topic Model (JST)**
  - 감성 분석과 토픽 모형 분석은 마케팅, 비즈니스 애널리틱스 분야에서 매우 유용하게 사용
  - 그러한 맥락에서 두 모형을 동시에 적용할 수 있는지에 대한 방법론적 확장이 관심사

  - Two-step method: 토픽 모형의 추정 결과로 나온 키워드들을 기반으로 추가적인 감성 분석을 시행
    - 특정 토픽이 긍정적인지 부정적인지 판별하는 것이 가능
  - One-step method: 토픽 모델링의 관점에서, 기존의 베이지안 모형 구조에 감성(sentiment)에 대한 층위를 추가하여 방법론적으로 확장한 것이 바로 JST 모형!

Distributed Representation-Based Approach

- **분산 표현(distribution representation) 기반의 모형**: 대다수가 머신 러닝 또는 딥러닝 기반의 모형

- **워드 임베딩 (Word embedding)**
  - 각 단어를 하나의 벡터(예: [300 X 1] 벡터)로 할당
  - 텍스트 데이터를 이용해 각 단어의 맥락적 의미를 반영하는 의미 벡터 공간을 추정하고, 각 단어의 의미를 의미 벡터 공간에서의 기하학적 위치로 표현
  - 예: LSA, Word2vec, FastText, Glove 등



Male-Female　　Verb tense

Distributed Representation-Based Approach

- **Word2vec 모형**
  - 워드 임베딩을 구현하는 대표적인 방법론으로서, 언어학의 분포 가설에 기반을 둠.
  - 분포 가설(distributional hypothesis): 비슷한 맥락에 등장하는 단어는 비슷한 의미를 가진다.

    ⇒ 텍스트 데이터 내에서 더 많은 맥락(문장에서의 위치 등)을 공유하는 단어일수록 의미가 더 비슷할 것이라고 간주
  - 분포 가설을 바탕으로 의미가 비슷하다고 판단되는 단어들을 다차원 벡터공간 상에 가깝게 위치하도록 할당
  - **Sentence2vec**: 단어가 아닌 각 문장을 벡터공간에 할당하는 모형
  - **Doc2vec**: 단어가 아닌 각 문서를 벡터공간에 할당하는 모형

Word Embedding

- **Word Representation**
  - Sparse Representation
    - One-Hot Encoding: 단어 전체 집합의 크기(N)를 차원으로 하는 벡터로 단어를 표현하되, 해당 단어의 인덱스만 1을 부여하고 나머지는 0을 부여하는 표현 방식
    - 예: {점심: [0, 1, 0, 0, 0, … , 0, 0, 0], 저녁: [0, 0, 0, 1, 0, … , 0, 0,], ...}
    - 예: Document-Term Matrix

    - 한계 1. 저장 공간 관리 면에서 매우 비효율적 (N이 매우 크다면?)
    - 한계 2. 단어의 유사도를 표현하는 것이 불가능

Word Embedding

- **Word Representation**
  - Dense Representation
    - Sparse representation과 달리, 단어 전체 집합의 크기(N)가 아닌 연구자가 사전에 부여한 값으로 차원의 크기를 설정
    - 데이터 학습을 통해 0, 1 외의 실수 값을 부여하여 좀 더 밀집된 형태로 단어를 벡터화
    - 예: {점심: [0.3, 2.0, -1.7, … , 5.7, 2.8], ...}

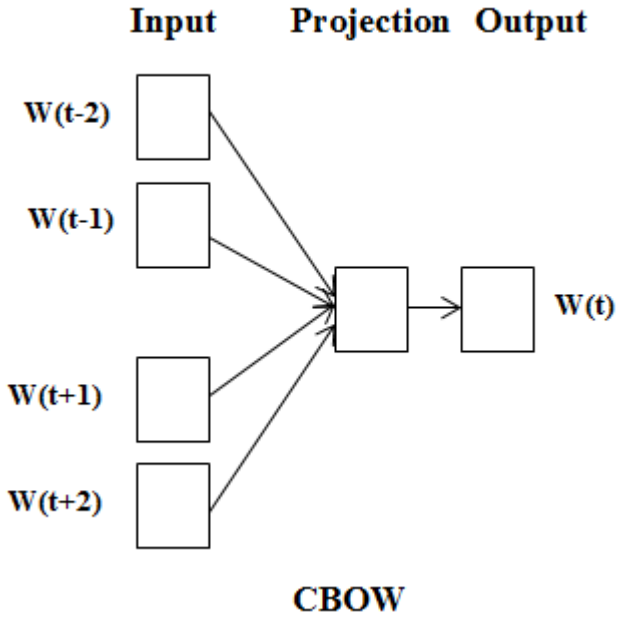
    - Word embedding: 단어를 dense representation으로 나타내는 방법론
    - LSA, Word2Vec, FastText 등이 word embedding에 해당한다고 볼 수 있음.

Word2Vec

- **Word2Vec**
  - Distributed Representation
    - 분포 가설 (distributional hypothesis): *비슷한 문맥에서 등장하는 단어들은 비슷한 의미를 가진다.*
    - 텍스트 데이터를 이용해 각 단어의 맥락적 의미를 반영하는 의미 벡터 공간을 추정하고, 각 단어의 의미를 의미 벡터 공간에서의 기하학적 위치로 표현
    - 즉, 텍스트 데이터에서 더 많은 맥락을 공유하는 단어일 수록 더 비슷한 의미를 가질 것으로 간주하고 의미 벡터 공간에 가까운 위치에 할당
  - Algorithms for Word2Vec
    - Continuous Bag of Words (CBOW)
    - Skip-Gram

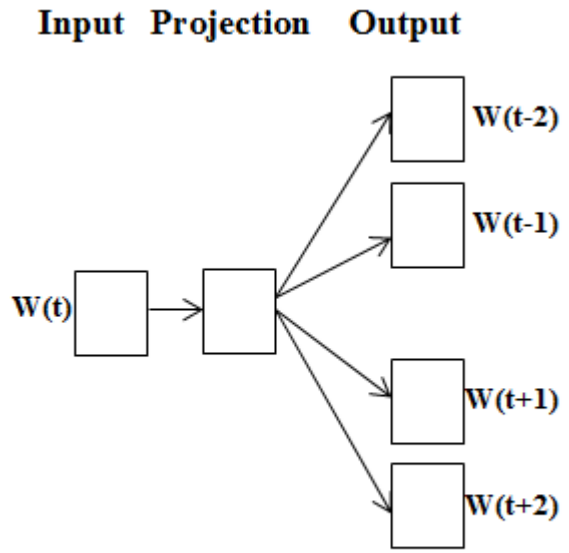
Word2Vec

- **Word2Vec**
  - Continuous Bag of Words (CBOW)
    - 주변 단어(context word)를 이용해 중심 단어(center word)를 학습, 예측
    - 주변 단어를 몇 개 고려할 것인가, 즉 학습시킬 맥락의 사이즈를 얼마로 할 것인지 윈도우(window) 값을 부여
    - Sliding window: 윈도우를 이동하면서, 즉 중심 단어와 주변 단어 조합을 바꿔가면서 데이터 학습을 진행



Word2Vec

- **Word2Vec**
  - Skip-Gram
    - 중심 단어(center word)를 이용해 주변 단어(context word)를 학습, 예측
    - 주변 단어를 몇 개 고려할 것인가, 즉 학습시킬 맥락의 사이즈를 얼마로 할 것인지 윈도우(window) 값을 부여
    - Sliding window: 윈도우를 이동하면서, 즉 중심 단어와 주변 단어 조합을 바꿔가면서 데이터 학습을 진행
    - 일반적으로 CBOW보다 우수한 성과를 보고한다고 알려져 있음.



Skip-gram

Distributed Representation-Based Approach

- **Neural Network Language Model (NNLM)**
  - 워드 임베딩의 원리를 바탕으로 단어의 유사도를 학습하여 단어 시퀀스를 예측하는 모형
  - 딥러닝 기반 언어 모형의 시초에 해당하는 모형으로 평가
- **Recurrent Neural Network-based Model**
  - Recurrent Neural Network (RNN): 입력 데이터와 출력 데이터를 sequence로서 처리하는 딥러닝 모형
  - 텍스트 생성 (Text generation)
    - 예: AI 자동 응답 채팅
  - 텍스트 분류 (Text classification)
    - 예: 스팸 메일 자동 분류 시스템

Distributed Representation-Based Approach

- **BERTopic Model (Grootendorst, 2022)**
  - Sentence-level BERT 기반의 토픽 모형 (문서 클러스터링 모형)
    - Bidirectional Encoder Representations from Transformers (BERT)

  - 1. SBERT: 주어진 텍스트 데이터에 대해 문서 단위의 임베딩 진행
  - 2. HDBSCAN (Hierarchical density-based spatial clustering of applications with noise): 유사한 문서 간의 클러스터링 진행 (하나의 클러스터 = 하나의 잠재 토픽)
  - 3. UMAP (Uniform manifold approximation and projection): 데이터 차원 축소에 활용
  - 4. c-TF-IDF (A class-based version of TF-IDF): 토픽별 키워드 추출 및 제시

Text Preprocessing

- **1. 토큰화 (tokenization)**
  - 언어학의 음운론, 형태론에 해당하는 단계로서 텍스트를 단어나 형태소 등의 단위로 쪼개는 작업
- **2. 정제 및 정규화 (cleaning and normalization)**
  - 표기법 통일(예: USA, United States), 대소문자 통일
- **3. 불용어 처리 (stopword)**
  - 관사, 전치사 등 텍스트 의미 분석에 큰 의미가 없는 단어를 제거
- **4. 어간 (stemming) 및 표제어 (lemmatization) 추출**
  - 단어의 의미를 담고 있는 핵심 부분을 추출
- **참고: 정규표현식을 이용한 전처리 (regular expression)**

Regular Expression

- **정규표현식을 이용한 전처리 (regular expression)**
  - 문자의 공통된 패턴을 이용해 텍스트 데이터를 정제할 정규적인 규칙을 부여

규칙	설명
\\	backslash (\)
\d	모든 숫자(digits)
\D	숫자를 제외한 모든 문자, [^0-9]와 동일한 의미
\s	모든 공백(space), [\t\n\r\f\v]와 동일한 의미
\S	공백을 제외한 모든 문자, [^ \t\n\r\f\v]와 동일한 의미
\w	문자 또는 숫자, [a-zA-Z0-9]와 동일한 의미
\W	문자 또는 숫자가 아닌 문자, [^a-zA-Z0-9]와 동일한 의미

Text Preprocessing

- Python code를 이용하여 영문 텍스트에 대한 전처리 실습 진행
- 필요한 라이브러리
  - NLTK: Python의 가장 대표적인 자연어 처리(NLP) 패키지
  - KoNLPy: 한국어 자연어 처리를 위한 패키지
  - Re: 정규 표현식을 이용한 텍스트 전처리를 위한 패키지

Text Preprocessing for Korean Text

- **한국어 텍스트 데이터의 특성**
  - 영어 텍스트의 경우, 띄어쓰기를 기준으로 토큰화하면 단어가 비교적 깔끔하게 분리되어 나오기 때문에 전처리의 결과가 대부분 좋음.
  - 그러나 한국어 텍스트의 경우, 띄어쓰기로 구분되는 “어절”이 반드시 “단어”와 일치하지가 않음. 이는 **“교착어”로서의 특성** 때문으로, 한국어는 영어와 달리 조사가 존재하고, 이러한 조사가 띄어쓰기 없이 붙어 있게 되어 이를 전부 분리해주는 전처리 과정이 필요
  - 즉, 한국어 전처리와 토큰화의 핵심은 조사를 잘 분리하여 토큰화하는 것으로, 이를 위해서는 **형태소(morpheme), 특히 자립 형태소**를 잘 추출할 수 있어야 함.
  - 또 하나의 어려운 점은 한국어의 경우 영어에 비해 **띄어쓰기가 잘 지켜지지 않는 경향이 존재**함. 이는 한국어의 경우 띄어쓰기가 잘 지켜지지 않더라도 의미가 잘 전달되기 때문이라는 의견이 있음.

Text Preprocessing for Korean Text

- **한국어 텍스트 데이터 전처리를 위한 대안**
  - Python 라이브러리에 내장된 “형태소 분석기”를 잘 활용하여 전처리 수행

    ⇒ 영어 텍스트의 경우, 어간 추출 등에서 형태소가 활용되지만 한국어 텍스트는 좀 더 일찍 형태소 개념을 활용하여 전처리 진행

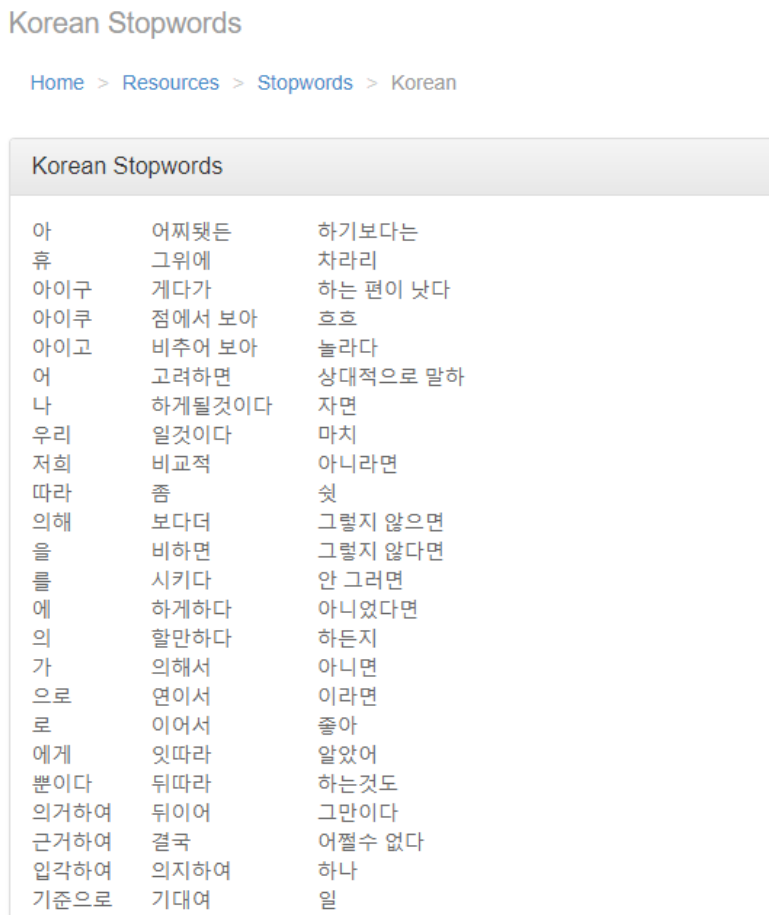
  - Okt, 꼬꼬마, 한나눔, 코모란 등 다양한 형태소 분석기가 존재하고, 성능이나 속도에 따라 다소 차이가 있음.

Text Preprocessing for Korean Text

- **한국어 텍스트 데이터 전처리를 위한 대안**
  - 한국어 텍스트 데이터 보정 알고리즘
    - PyKoSpacing: 딥러닝 기반의 띄어쓰기 보정 (https://github.com/haven-jeon/PyKoSpacing)
    - Py-hanspell: 네이버 맞춤법 검사기를 이용한 오타 보정을 위한 라이브러리 (https://github.com/ssut/py-hanspell)

Text Preprocessing for Korean Text

- 한국어 불용어 세트
  - https://www.ranks.nl/stopwords/korean
  - https://bab2min.tistory.com/544



Korean Stopwords

Home > Resources > Stopwords > Korean

Korean Stopwords

아	어찌됏든	하기보다는
휴	그위에	차라리
아이구	게다가	하는 편이 낫다
아이쿠	점에서 보아	흐흐
아이고	비추어 보아	놀라다
어	고려하면	상대적으로 말하
나	하게될것이다	자면
우리	일것이다	마치
저희	비교적	아니라면
따라	좀	쉿
의해	보다더	그렇지 않으면
을	비하면	그렇지 않다면
를	시키다	안 그러면
에	하게하다	아니었다면
의	할만하다	하든지
가	의해서	아니면
으로	연이서	이라면
로	이어서	좋아
에게	잇따라	알았어
뿐이다	뒤따라	하는것도
의거하여	뒤이어	그만이다
근거하여	결국	어쩔수 없다
입각하여	의지하여	하나
기준으로	기대여	일

KoNLPy 설치

- KoNLPy 라이브러리의 설치
  - Python의 버전, Java의 버전, JAVA_HOME 설정, JPype1 설치 여부 등에 따라 KoNLPy 설치 도중 오류가 발생할 수 있음.
  - 다음의 링크를 참고하여 설치를 진행할 수 있음.
    - https://konlpy-ko.readthedocs.io/ko/v0.4.3/install/
    - https://hcid-courses.github.io/TA/FAQ/konlpy_install_troubleshoot.html
    - https://velog.io/@recoder/KoNLPy-%EC%84%A4%EC%B9%98-%EC%97%90%EB%9F%AC-%ED%95%B4%EA%B2%B0
    - https://byeon-sg.tistory.com/entry/%EC%9E%90%EC%97%B0%EC%96%B4-%EC%B2%98%EB%A6%AC-konlpy-%EC%84%A4%EC%B9%98-%EC%98%A4%EB%A5%98-okt%EC%97%90%EB%9F%AC-already-loaded-in-another-classloader-SystemErro-1