Testing : Applications to text analysis

Testing high-dimension multinomials with applications to text analysis

Mose Park

Department of Statistical Data Science
University of Seoul

Selective. Lab

Oct 21th , 2024

Index

Overview

Estimation
&
Testing

Theorical
Properties

Text
Analysis

1

Introduction

텍스트와 다항분포의 관계

$X_i \sim \text{Multinomial}(N_i, \Omega_i), 1 \leq i \leq n \ , \quad X_i \in \mathbb{R}^p$ 라 가정하면

$$\mu_k = \frac{1}{n_k \bar{N}_k} \sum_{i \in S_k} N_i \Omega_i, \quad 1 \leq k \leq K \qquad (2)$$

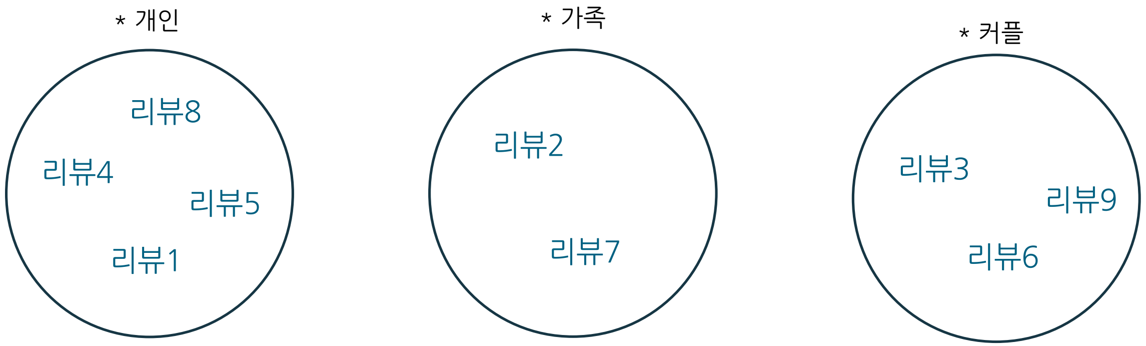
'그룹 k에서 총 리뷰 수 * 그룹 k에서 평균 리뷰 길이' = '그룹 k의 리뷰 빈도 정보' 로 스케일링

- 확률변수 X_i는 서로 다른 p개의 단어로 구성된 리뷰 i 에서 각 단어들의 출현 빈도를 나타내는 벡터
- N_i는 리뷰 i의 전체 길이, Ω_i는 리뷰 i의 확률질량함수(pmf)

가설검정 디자인

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$$

* k = 3 , 여행객 유형



$$H_0 : \mu_1 = \mu_2 = \mu_3$$

- Null : 그룹의 평균 응답 간 통계적으로 유의한 차이가 존재하지 않는다.
- 예를 들어, 여행객 유형별 각 리뷰의 응답 스타일을 비교해볼 수 있다. → 그룹별로 리뷰 스타일이 유사한지, 리뷰 다양성은 어떨지

문제 응용 사례

- ✓ 토픽 모델 글로벌 검정
- ✓ 논문 저자 추정
- ✓ 이산 분포간 근접 테스트

$$\text{Multinomial}(N_1, \mu), \text{Multinomial}(N_2, \theta)$$

μ = θ 를 검정, K=2, $n_1 = n_2 = 1$ 의 사례

- 오늘 발표는 주로 관심이 있었던 이산 분포간 Closeness Test 위주로 준비해왔습니다.
- 실제 텍스트와 생성 텍스트 간의 정량적인 차이를 보기 위해 3번째 task 위주로 읽었습니다.

➢ DELVE 라는 **estimator**로 텍스트 코퍼스와 관련된 이산 분포의 그룹을 비교하는 것이 논문의 목적입니다.

2

Estimation

노테이션

가정

$$\begin{cases} X_1, \ldots, X_n \ \text{독립} \\ X_i \sim \text{Multinomial}(N_i, \Omega_i) \ \ 1 \le i \le n \end{cases}$$

$$n_k = |S_k| \longrightarrow \text{그룹 k의 리뷰 개수}$$

$$\overline{N}_k = n_k^{-1} \textstyle\sum_{i \in S_k} N_i \longrightarrow \text{그룹 k의 리뷰 평균 길이}$$

$$\overline{N} = n^{-1} \textstyle\sum_{i=1}^{n} N_i \longrightarrow \text{전체 리뷰 평균 길이}$$

기본적으로

- X_i : 리뷰 i의 단어 수
- n : 리뷰 개수
- N_i : 리뷰 i의 총길이
- Ω_i : 리뷰 i의 단어 빈도 PMF

DELVE

전체 평균에 대해 다음과 같이 정의 합니다.

$$\mu := \frac{1}{n\overline{N}} \sum_{k=1}^{K} n_k \overline{N}_k \mu_k = \frac{1}{n\overline{N}} \sum_{i=1}^{n} N_i \Omega_i. \quad (5)$$

K 그룹별로 평균 PMF간 변동을 측정하는 양을 다음과 같이 정의합니다.

$$\rho^2 := \sum_{k=1}^{K} n_k \overline{N}_k \|\mu_k - \mu\|^2. \quad (6)$$

- 그룹별 평균 PMF 간의 차이가 있는지 없는지가 검정의 목표였습니다.
- 따라서 Null H_0는 식 (6) 이 0일 때 성립합니다. 이 식을 **검정 통계량**으로 develop 하고자 합니다.

추정량과 검정통계량의 관계

그룹평균과 전체평균 MVUE를 다음과 같이 구해볼 수 있고,

$$\hat{\mu}_k = \frac{1}{n_k \overline{N}_k} \sum_{i \in S_k} X_i \quad \text{그리고} \quad \hat{\mu} = \frac{1}{n\overline{N}} \sum_{k=1}^{K} n_k \overline{N}_k \hat{\mu}_k = \frac{1}{n\overline{N}} \sum_{i=1}^{n} X_i. \qquad (7)$$

수리통계학 시간 추정 단원에서 배웠던 직관을 활용해 나이브 추정량을 식 (8)과 같이 생각해볼 수 있습니다.

$$\tilde{T} = \sum_{j=1}^{p} \tilde{T}_j, \quad \text{여기서} \quad \tilde{T}_j = \sum_{k=1}^{K} n_k \overline{N}_k (\hat{\mu}_{kj} - \hat{\mu}_j)^2 \qquad (8)$$

선행연구 Cai et al. (2023)에서 위 estimator (8)을 비편향화(de-bias)한 것을 다음과 같이 보였습니다.

$$T = \sum_{j=1}^{p} T_j, T_j = \sum_{k=1}^{K} \left[n_k \overline{N}_k (\hat{\mu}_{kj} - \hat{\mu}_j)^2 - \left(\frac{1}{n_k \overline{N}_k} - \frac{1}{n\overline{N}} \right) \sum_{i \in S_k} \frac{X_{ij}(N_i - X_{ij})}{N_i - 1} \right]$$

- μ̂_k는 각각 그룹 k 내에서 sample들의 평균 단어 빈도를 가중하여 구한 값, μ̂ 의 경우는 전체 sample들에 대한 가중평균
- de-bias에 대한 풀이는 표본 분산 추정 풀이과정을 생각해보면 이해가 쉬울 것 같습니다. (Lemma 1)

분산 추정

오메가를 p로 이입해 생각해보기

- $E[X_{ij}X_{mj}] = N_i N_m \Omega_{ij}\Omega_{mj}$
- $E[X_{ij}^2] = N_i^2\Omega_{ij}^2 + N_i\Omega_{ij}(1-\Omega_{ij})$
- $E[X_{ij}(N_i - X_{ij})] = N_i(N_i-1)\Omega_{ij}(1-\Omega_{ij})$

mild regularity conditions 하에 Lemma 1 아래에 있는 식 (10)을 정의할 수 있고 이를 이용해 분산 추정을 합니다.

분산 추정 식은 다음과 같습니다.

$$V = 2\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{j=1}^{p}\left(\frac{1}{n_k\bar{N}_k}-\frac{1}{n\bar{N}}\right)^2\frac{X_{ij}^2-X_{ij}}{N_i(N_i-1)}+\frac{2}{n^2\bar{N}^2}\sum_{k\neq\ell}\sum_{i\in S_k}\sum_{m\in S_\ell}\sum_{j=1}^{p}X_{ij}X_{mj}+$$
$$2\sum_{k=1}^{K}\sum_{i\in S_k, m\in S_k, i\neq m}\sum_{j=1}^{p}\left(\frac{1}{n_k\bar{N}_k}-\frac{1}{n\bar{N}}\right)^2 X_{ij}X_{mj}$$

- 이항분포에서 평균, 분산을 구하는 방식에서 확장해보면 다항분포의 성질에 따라 성립하는 것을 이해할 수 있습니다.
- 분산 추정은 논문에서 제시하는 estimator를 검정통계량으로 응용하기 위한 발판이라고 생각하시면 됩니다.

검정통계량

논문에서 제안하는 검정통계량은

$$\psi = \frac{T}{\sqrt{V}}$$

귀무가설 H_0 하에서
+
mild regularity 하에서

$$\psi \to N(0,1)$$

* 기각역

$\psi > z_\kappa$, 여기서 z_κ는 표준 정규분포 $N(0,1)$의 $(1-\kappa)$분위수입니다.

- 검정통계량 psi는 **DE**-biased and **L**ength-adjusted **V**ariability **E**stimator라 불립니다. (DELVE)
- K = 2, K = n 등 특수한 경우는 논문을 참고바라고 이 통계량의 이론적인 성질을 다음 섹션에서 다루고자 합니다.

3

Theoretical Property

Regularity conditions

$$\min_{1\le i\le n} N_i \ge 2, \quad \max_{1\le i\le n} \|\Omega_i\|_\infty \le 1 - c_0, \quad \max_{1\le k\le K} \frac{n_k \bar{N}_k}{n\bar{N}} \le 1 - c_0. \tag{21}$$

- 조건 1 : 가장 길이가 작은 텍스트가 적어도 2 이상이라는 것을 보장해주는 조건
- 조건 2 : PMF에서 특정 카테고리가 지나치게 높은 확률을 제한해주는 조건
- 조건 3 : 특정 그룹이 큰 비율을 차지하는 불균형한 경우를 제한해주는 조건

$$\alpha_n := \max\left\{\left(\sum_{k=1}^{K}\frac{\|\mu_k\|_3^3}{n_k\bar{N}_k},\quad \sum_{k=1}^{K}\frac{\|\mu_k\|_2^2}{n_k^2\bar{N}_k^2}\right)\right\}\Big/\left(\sum_{k=1}^{K}\|\mu_k\|_2^2\right)^2 \xrightarrow{n\bar{N}\to\infty} o(1)$$

$$\beta_n := \max\left(\sum_{k=1}^{K}\sum_{i\in S_k}\frac{N_i^2}{n_k^2\bar{N}_k^2}\|\Omega_i\|_3^3,\quad \sum_{k=1}^{K}\|\Sigma_k\|_F^2\right)\Big/\left(K\|\mu\|_2^2\right) \xrightarrow{n\bar{N}\to\infty} o(1)$$

$$\Sigma_k = \frac{1}{n_k\bar{N}_k}\sum_{i\in S_k}N_i\Omega_i\Omega_i'$$

$$\frac{\|\mu\|_4^4}{K\|\mu\|^4} = o(1)$$

- α 식의 경우 k 그룹에서 평균 PMF의 크기, k 그룹 내 큰 원소값이 얼만큼 영향을 끼치는지(불균형)을 의미합니다.
- β 는 확률분포 자체의 불균형과, 그룹 내 샘플들의 변동성을 의미합니다.
- 기본적으로 normal 로 수렴하는 과정에서 안좋은 상황에도 안정적으로 수렴하는지 보장해주는 조건들입니다.

점근 정규성 (Asymptotic normality)

정리 1, 2

$X_i \sim \mathrm{Multinomial}(N_i, \Omega_i), 1 \le i \le n$, $\mu_k = (n_k \bar{N}_k)^{-1} \sum_{i \in S_k} N_i \Omega_i, \quad 1 \le k \le K$ 이고

귀무가설이 성립한다고 가정하자. 이론적 성질에서 다룬 조건들이 모두 만족한다면 다음 분포 수렴이 성립한다.

$$T/\sqrt{\Theta_n} \xrightarrow{d} N(0,1) \quad \text{as} \quad n\bar{N} \to \infty$$

위 정리 1 조건 하에서

$$V/\Theta_n \to 1 \quad \text{as} \quad n\bar{N} \to \infty$$

$$\psi := T/\sqrt{V} \xrightarrow{d} N(0,1) \quad \text{as} \quad n\bar{N} \to \infty$$

- 이전 섹션에서 제시했던 조건들을 이용해 논문에서 제안하는 검정통계량의 점근 정규성을 보인다. (증명은 마팅게일 CLT 활용)

검정력

정리 3, 4

$X_i \sim \text{Multinomial}(N_i, \Omega_i), 1 \leq i \leq n$, $\mu_k = (n_k \bar{N}_k)^{-1} \sum_{i \in S_k} N_i \Omega_i, \quad 1 \leq k \leq K$ 이고

이론적 성질에서 다룬 조건들이 모두 만족한다면 estimator T의 평균과 분산은 다음과 같다.

* scaled rho

$$E[T] = n\bar{N}\|\mu\|^2 \omega_n^2 \ , \qquad V(T) = O\left(\sum_{k=1}^{K} \|\mu_k\|^2\right) + E[T] \cdot O\left(\max_{1 \leq k \leq K} \|\mu_k\|_\infty\right)$$

그리고 대립가설 가정하에 $n\bar{N} \to \infty$ 이면

$$SNR_n := \frac{n\bar{N}\|\mu\|^2 \omega_n^2}{\sqrt{\sum_{k=1}^{K} \|\mu_k\|^2}} \to \infty$$

- 분자는 signal, 분모는 noise에 해당한다. 왜냐하면 H_a 하에 검정 통계량의 기댓값이 분자고 분산이 분모이기 때문입니다.
- 따라서 signal >> noise일수록 POWER가 커진다는 것을 이해할 수 있다.

K=2

$X_i \sim \mathrm{Multinomial}(N_i, \Omega_i), \quad G_j \sim \mathrm{Multinomial}(M_j, \Gamma_j)$ 이라하자.

- 가설 디자인

$$H_0 : \eta = \theta, \quad \eta = \frac{1}{n\bar{N}} \sum_{i=1}^{n} N_i \Omega_i, \quad \theta = \frac{1}{m\bar{M}} \sum_{i=1}^{m} M_j \Gamma_j$$

- 조건
  - regularity conditions
  - alpha, beta conditions ⟶ 저자 식별 문제에서 데이터 불균형 문제로 조건 수정

- 검정력 (정리 6)

$n\bar{N} \to \infty$ 이면 $$\frac{\|\eta - \theta\|^2}{\left(\frac{1}{n\bar{N}} + \frac{1}{m\bar{M}}\right) \max(\|\eta\|, \|\theta\|)} \to \infty$$

K=n

$X_i \sim \mathrm{Multinomial}(N_i, \Omega_i), \quad G_j \sim \mathrm{Multinomial}(M_j, \Gamma_j)$ 이라하자.

➢ 가설 디자인

$$H_0 : \Omega_i = \mu, \quad 1 \leq i \leq n$$

➢ 조건

- regularity conditions
- alpha, beta conditions

➢ 검정력 (정리 7)

$$\omega_n = \omega_n(\Omega_1, \Omega_2, \ldots, \Omega_n) = \frac{1}{n\bar{N}\|\mu\|^2} \sum_{i=1}^{n} N_i \|\Omega_i - \mu\|^2$$

$n\bar{N} \to \infty$ 이면 $\dfrac{n\bar{N}\|\mu\|^2 \omega_n^2}{\sqrt{\sum_{i=1}^{n} \|\Omega_i\|^2}} \to \infty$

4

Experiments

시뮬레이션

실험 1: 점근적 정규성

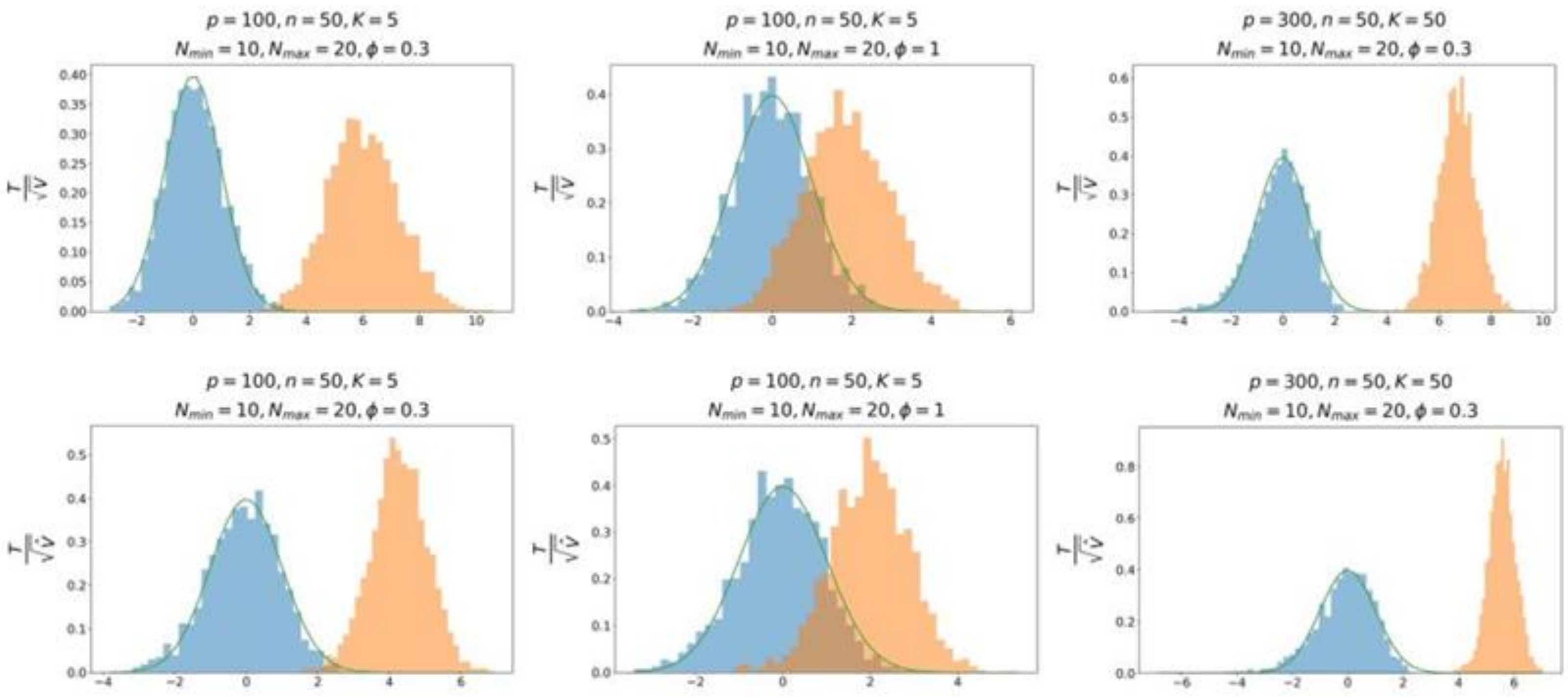
다음과 같은 매개변수 $(n, p, K, N_{\min}, N_{\max}, \phi)$를 기반으로 데이터를 생성합니다:

1. 먼저, $\{1, \ldots, n\}$을 크기가 동일한 K개의 그룹으로 나눕니다.
2. 그 다음, $\Omega^1_{\text{alt}}, \ldots, \Omega^n_{\text{alt}}$을 Dirichlet$(p, \phi 1_p)$에서 독립적으로 추출합니다.
3. 세 번째로, $N_i \sim$ iid Uniform$[N_{\min}, N_{\max}]$을 따르도록 하고, $\Omega_{\text{null}} = \mu$로 설정합니다. 여기서 $\mu := \frac{1}{nN} \sum_i N_i \Omega_i^{\text{alt}}$입니다.
4. 마지막으로, 모델 (1)을 사용하여 $X_1, \ldots, X_n$을 생성합니다.

세 가지 하위 실험을 고려합니다:

- **실험 1.1**: $(n, p, K, N_{\min}, N_{\max}, \phi) = (50, 100, 5, 10, 20, 0.3)$
- **실험 1.2**: ϕ 값을 1로 변경하고, 나머지 매개변수는 동일하게 유지합니다.
- **실험 1.3**: 실험 1.1의 매개변수를 유지하되, (p, K)를 $(300, 50)$으로 변경

시뮬레이션 결과

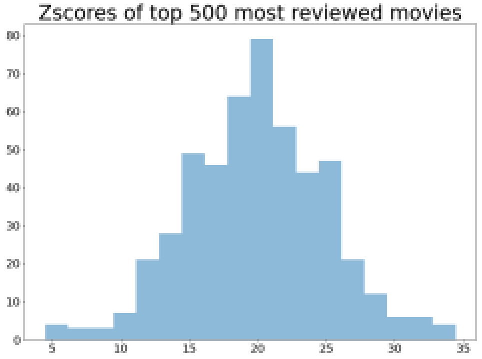


✓ p, K 가 커질수록 어떤 의미?

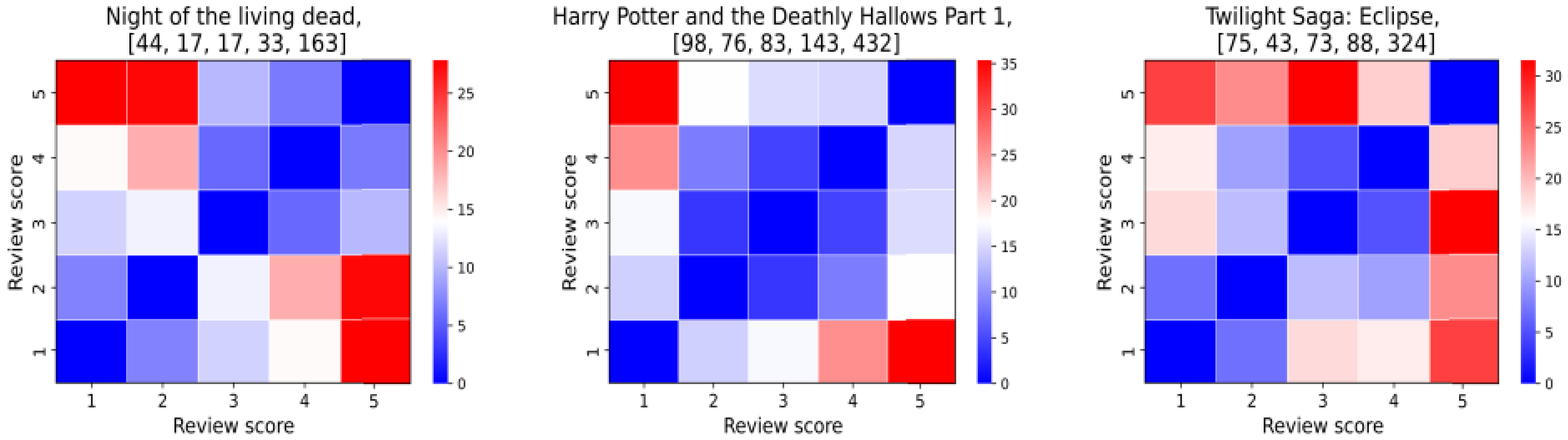
✓ ϕ = 1 vs ϕ = 0.3 비교

- DELVE와 분산 정의가 약간다른 DELVE+의 histogram
- Null 하에서 검정 통계량 DELVE의 히스토그램을 파란색으로 그림. → N(0,1)에 적절히 맞춰짐 (asymptotic normality)

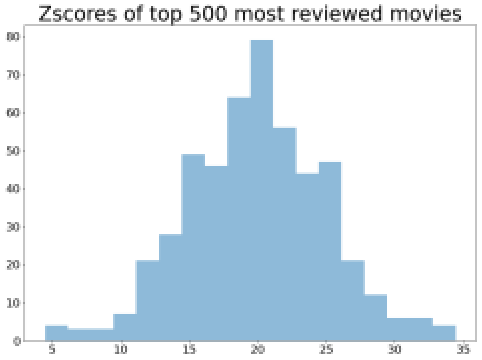
리뷰 텍스트 분석



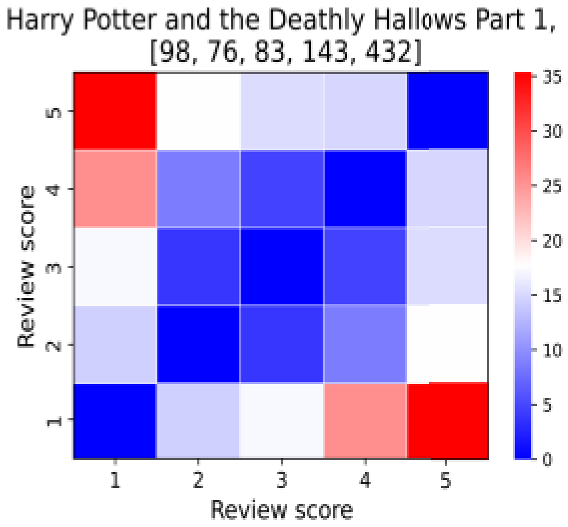
Rank	Title	Z-Score	Total reviews
1	Prometheus	34.44	813
2	Expelled: No Intelligence Allowed	34.17	830
3	V for Vendetta	32.24	815
4	Sin City	31.72	828
⋮	⋮	⋮	⋮
17	Cars	19.98	902
18	Food, Inc.	17.81	876
19	Jeff Dunham: Arguing with Myself	4.96	860
20	Jeff Dunham: Spark of Insanity	4.46	877

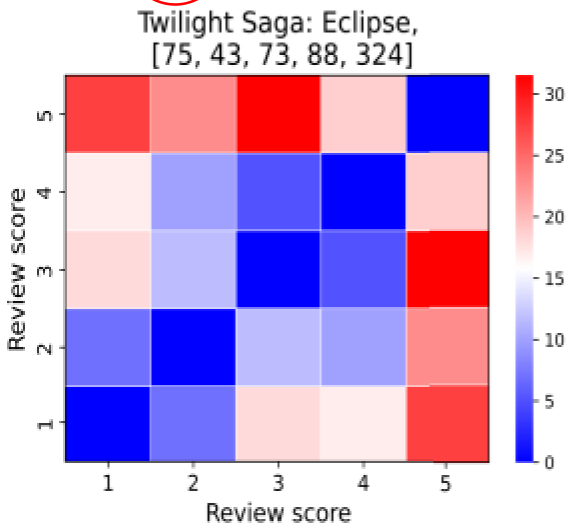


- 리뷰 텍스트 → 어간 추출 (Stemming) → **K = n** 으로 **DELVE+** 를 적용해 검정



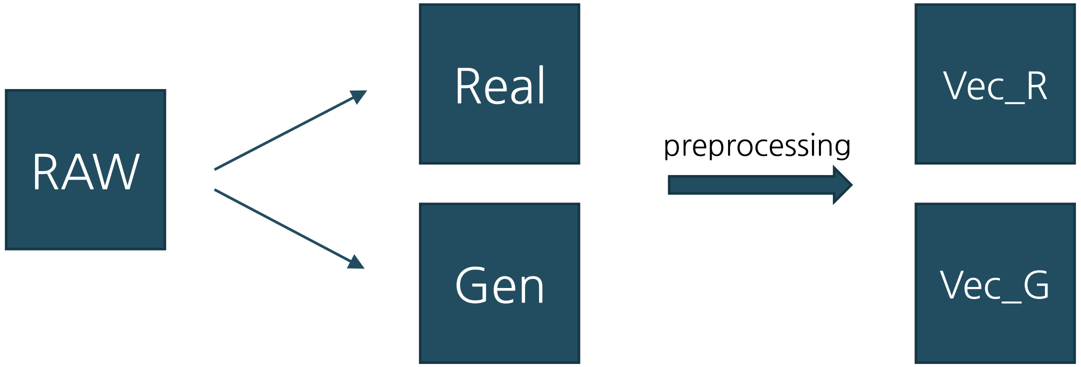
Rank	Title	Z-Score	Total reviews
1	Prometheus	34.44	813
2	Expelled: No Intelligence Allowed	34.17	830
3	V for Vendetta	32.24	815
4	Sin City	31.72	828
⋮	⋮	⋮	⋮
17	Cars	19.98	902
18	Food, Inc.	17.81	876
19	Jeff Dunham: Arguing with Myself	4.96	860
20	Jeff Dunham: Spark of Insanity	4.46	877



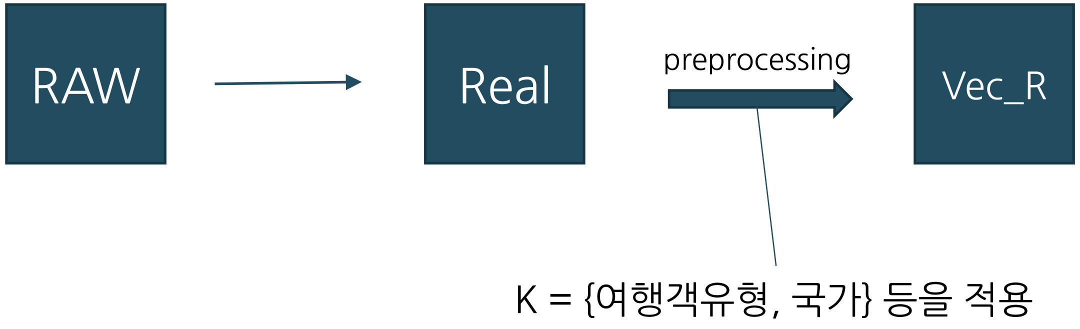


- Z-score가 작은 하위 4개 영화 리뷰에 대한 결과는? 리뷰끼리 강한 동질성을 갖고 있다.
- 아래 시각화는 k=5 (점수 5개 구간) 사이의 차이를 보는 검정을 진행, (해리포터의 예로 2~4점 리뷰들이 비슷하다고 판단)

내 연구에 응용



$$H_0 : \eta = \theta, \quad \eta = \frac{1}{n\overline{N}} \sum_{i=1}^{n} N_i \Omega_i, \quad \theta = \frac{1}{m\overline{M}} \sum_{i=1}^{m} M_j \Gamma_j$$



$$H_0 : \Omega_i = \mu, \quad 1 \leq i \leq n$$

- 텍스트에 대한 분포 가정도 어렵고 어떻게 텍스트를 활용할 것인가 고민하다가 이 논문을 찾게 되었습니다.

논문 고른 이유와 느낀점

- 제 연구에 적용해볼 수 있는 연구라고 생각했습니다.
  - 텍스트에 대한 분포를 count 기반의 PMF 로 정의하는 것에 매력을 느꼈습니다.
  - 실제 텍스트와 생성 텍스트를 closeness test를 통해 구분하면서 실제 데이터에 있는 카테고리를 쓰고 싶었습니다.

- 4대저널 논문이라고 해서 반드시 어려운 개념만 논문에 넣는것이 아닌 것을 느꼈습니다.
  - 디테일은 분명 어렵지만 분명 수리통계학, 회귀분석 같은 이론 수업에서 배운 큰 흐름을 따른다는 것을 느꼈습니다.
  - 오히려 자주 접하고 입증된 익숙한 것들(CLT, asymptotic 등)을 잘 활용한다는 느낌이 들었습니다.

어간 추출 (Stemming)

어간 추출 전 : ['formalize', 'allowance', 'electricical']

어간 추출 후 : ['formal', 'allow', 'electric']

['was', 'this', 'Billy'] → ['wa', 'thi', 'billi']

- 가끔 원하지 않는 결과도 나올 수 있다.
- 논문에서는 구체적으로 어떻게 전처리를 했는지에 대한 자세한 디테일은 없었다.