리뷰의 길이를 통해 알아보는

# 영화 관객의 성향

2020 데이터 크리에이터 캠프 12호

“

숙명여자대학교 국주현
한양대학교 이나현
숙명여자대학교 이다현
한양대학교 이세민
숙영여자대학교 장은조

**데븐데븐**

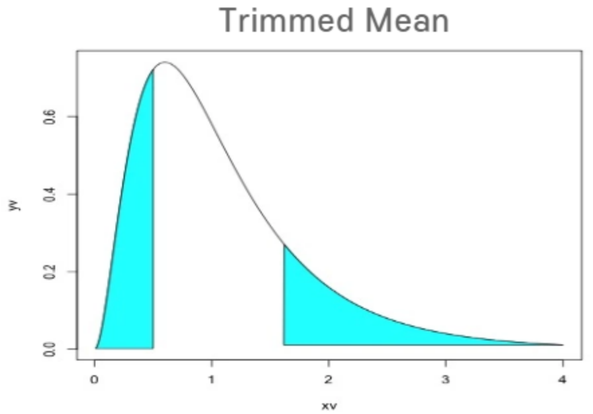
장은조

과학기술정보통신부

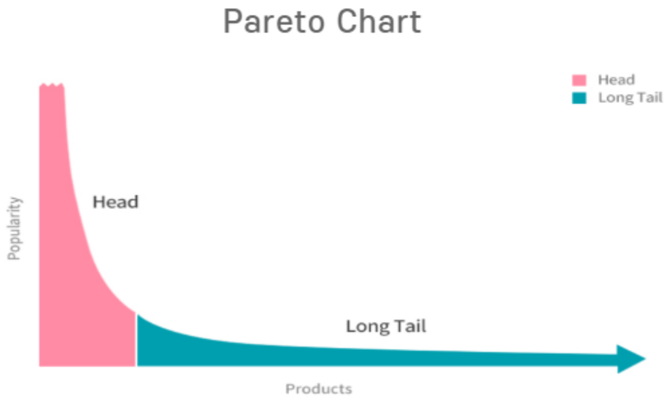
NIA 한국정보화 진흥원

# 짧은 리뷰와 긴 리뷰

그 기준을 찾기




편차가 큰 자료의 경우, 산술평균이 적합하지 않으므로, 자료의 총 개수에서 일정비율만큼 가장 큰 부분과 작은 부분을 제거 후 평균을 산출한다




전체효과를 만드는 것은 상위 20%의 사람들이다.

## 절사평균 + 파레토 차트 의 원리

정확한 산술평균을 위해 절사평균을 사용하고 상업적으로 유의미한 의미를 가지고 있는 상위 20%의 길이의 리뷰수를 참고

## 이상치 상위 8.7% , 하위 10% 제거

박스플롯을 통해 구한 상위 이상치가 8.7%

# 짧은 리뷰와 긴 리뷰

그 기준을 찾기

```
short_df = unique_df[:][int(round(total*0.087)):int(round(total*0.287))]
short_df
```

| | id | document | label | preprocessed_document | document_len | newindex |
|---|---|---|---|---|---|---|
| **79060** | 6022374 | 배우팝스타 최고 | 1 | 배우팝스타최고 | 7 | 16925 |
| **119643** | 148361 | 캠페인 동참 ㅋㅋ | 0 | 캠페인동참ㅋㅋ | 7 | 16926 |
| **169510** | 7661175 | 지나친 무리수다 | 0 | 지나친무리수다 | 7 | 16927 |
| **190797** | 7139310 | 이것도 영화라고;;;; | 0 | 이것도영화라고 | 7 | 16928 |
| **25142** | 8066167 | 최고 인듯 하네요 !!! | 1 | 최고인듯하네요 | 7 | 16929 |
| ... | ... | ... | ... | ... | ... | ... |
| **24728** | 5127702 | 후반부로 갈수록 아쉬워는 영화 | 1 | 후반부로갈수록아쉬워는영화 | 13 | 55829 |
| **106714** | 9470067 | 댓글알바 때문에 1점 벨런스맞춤 | 0 | 댓글알바때문에점벨런스맞춤 | 13 | 55830 |
| **50132** | 9616988 | 브랜든대령님 겨론해주세여 ㅠ | 1 | 브랜든대령님겨론해주세여ㅠ | 13 | 55831 |
| **146602** | 10221286 | 개노잼영화는평점떨어트리기 | 0 | 개노잼영화는평점떨어트리기 | 13 | 55832 |
| **148171** | 10017021 | 정말 실망 재미없었다 재미없음 | 0 | 정말실망재미없었다재미없음 | 13 | 55833 |

38909 rows × 6 columns

짧은 리뷰 : 8.7%~28.7% : 7~13자

# 짧은 리뷰와 긴 리뷰

그 기준을 찾기

```
lens_df = unique_df[:][int(round(total*(1-0.2-0.087))):int(round(total*(1-0.087)))]
lens_df
```

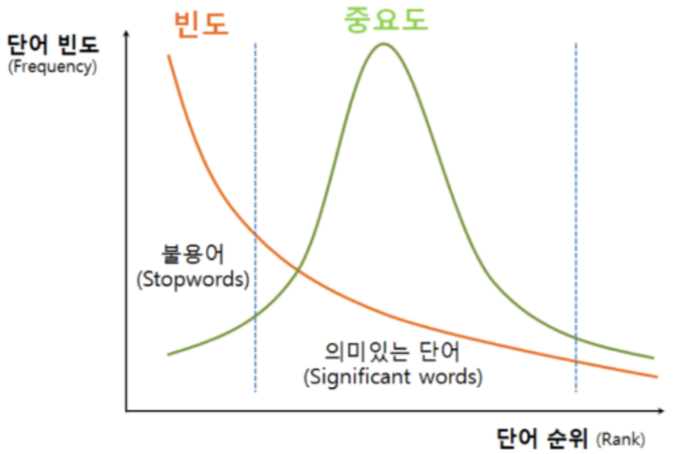
| | id | document | label | preprocessed_document | document_len | newindex |
|---|---|---|---|---|---|---|
| **167529** | 4124245 | 이런남자졸싫다 능력없이무작정쫓아다니는..말도안되는영화스토리 | 0 | 이런남자졸싫다능력없이무작정쫓아다니는말도안되는영화스토리 | 29 | 136709 |
| **76084** | 5109297 | 아이디어는 진부할지몰라도 실제 상황을 영화적으로 raw하게 잘 보여준듯 | 1 | 아이디어는진부할지몰라도실제상황을영화적으로raw하게잘보여준듯 | 29 | 136710 |
| **126820** | 8616786 | 글쎄요) 마지막엔 따뜻하고 좋았지만, 과정은 좋지않았다. 참 아쉬운 영화. | 0 | 글쎄요마지막엔따뜻하고좋았지만과정은좋지않았다참아쉬운영화 | 29 | 136711 |
| **14534** | 4800088 | 가까운 것에 대한 소중함을 느끼지 못했던 나를 부끄럽게 만들어 주었다 | 1 | 가까운것에대한소중함을느끼지못했던나를부끄럽게만들어주었다 | 29 | 136712 |
| **120286** | 2992799 | 영화에서 재미를 빼면 어쩌란 거냐.. 아무리 블랙풍자코미디라곤 하지만.. | 0 | 영화에서재미를빼면어쩌란거냐아무리블랙풍자코미디라곤하지만 | 29 | 136713 |
| ... | ... | ... | ... | ... | ... | ... |
| **89855** | 9730028 | 마지막에 송중기랑 박보영(잡미니)이 시로 끝이인고가메라 있으로 돌리는빈에시 박보영이... | 1 | 마지막에송중기랑박보영잡미니이시로끝이인고가메라있으로돌리는빈에시박보영이왔었출직으로나오게... | 61 | 177613 |
| **18310** | 9241082 | 내기 본 재난영화 중 제일 명작이라 생각되는 영화. 또 보통 2편 나오면 1편에 비... | 1 | 내기본재난영화중제일명작이라생각되는영화또보통편나오면편에비해별문데이건편에는조금못미치지만... | 61 | 177614 |
| **67031** | 9599582 | 이미 알고있는 역사에 흥미로운 주제를 곁들여 좋은 이야기를 만든 영화. 개개인의 캐... | 1 | 이미알고있는역사에흥미로운주제를곁들여좋은이야기를만든영화개개인의캐릭터가대력있고특히이성새... | 61 | 177615 |
| **162967** | 8575577 | 조자룡이 유비부인과 아들을 구하러 갈 당시 전여 이름없던 병졸로 나오는게 말도 안됨... | 0 | 조자룡이유비부인과아들을구하러갈당시전여이름없던병졸로나오는게말도안됨다조자룡은그당시이미유... | 61 | 177616 |
| **9930** | 9563837 | 대사 하나하나가 명언이쇼. 인생 사는 것이 결국은 죽기 3일전이라는 대사. 살면서 ... | 1 | 대사하나하나가명언이쇼인생사는것이결국은죽기일전이라는대사살면서시급이최악은아닐건데라는대사... | 61 | 177617 |

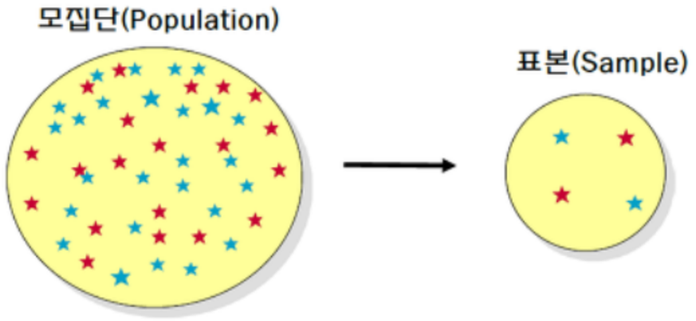
38909 rows × 6 columns

긴 리뷰 : 71.3%~91.3% : 29 ~ 61자

# 스탑 워드 제거 & 샘플링 기법 이용

효율을 위해




## 1 스탑워드 제거

- 모든 군집에서 똑같이 나오는 단어 ( 큰 의미가 없는 단어) 제거
- 분석해야 할 칼럼의 단어 수 줄이기

⇒ 형태소 분석, vocarb 상위 100, 200, 300 개만 사용하기

## 2 샘플링 기법 이용

- 모집단은 데이터의 수가 많아 모두 분석하기 효율이 낮음
- 모집단의 특성을 나타낼 수 있는 샘플을 무작위로 추출

# K = 2 선정

효율을 위해

```
[28] kmeans_clustering = KMeans(n_clusters = 2, random_state = 1121).fit(short_cv_res_array[:500, :]) # KMeans 알고리즘 (군집 수는 3개로 설정)
```

```
[29] kmeans_clustering.inertia_
```

```
1455.0252580716008
```

# 결 론

짧은 리뷰 집단과 긴 리뷰 집단, K=2

짧은 리뷰 집단

```
[('최고', 26),
 ('정말', 20),
 ('너무', 19),
 ('진짜', 17),
 ('ㅋㅋㅋ', 11),
 ('감동', 10),
 ('연기', 9),
 ('이다', 9),
 ('재밌다', 9),
 ('좋다', 9),
 ('좋아요', 9),
 ('ㅠㅠ', 8),
 ('드라마', 8),
 ('봐도', 8),
 ('평점', 8),
 ('역시', 7),
 ('재밌게', 7),
 ('ㅋㅋ', 6),
 ('ㅎㅎ', 6),
 ('다시', 6)]
```

```
[('그냥', 9),
 ('최고다', 4),
 ('영화', 3),
 ('진심', 3),
 ('잼슴', 2),
 ('필요없다', 2),
 ('가슴', 1),
 ('고마워요', 1),
 ('니까', 1),
 ('다시', 1),
 ('보는', 1),
 ('봐고', 1),
 ('봐도', 1),
 ('심심할', 1),
 ('아리구나', 1),
 ('아빠', 1),
 ('안될', 1),
 ('액션', 1),
 ('으로', 1),
 ('이건', 1)]
```

# 결 론

짧은 리뷰 집단과 긴 리뷰 집단, K=2

짧은 리뷰 집단

```
[('최고', 26),
 ('정말', 20),
 ('너무', 19),
 ('진짜', 17),
 ('ㅋㅋㅋ', 11),
 ('감동', 10),
 ('연기', 9),
 ('이다', 9),
 ('재밌다', 9),
 ('좋다', 9),
 ('좋아요', 9),
 ('ㅠㅠ', 8),
 ('드라마', 8),
 ('봐도', 8),
 ('평점', 8),
 ('역시', 7),
 ('재밌게', 7),
 ('ㅋㅋ', 6),
 ('ㅎㅎ', 6),
 ('다시', 6)]
```

```
[('그냥', 9),
 ('최고다', 4),
 ('영화', 3),
 ('진심', 3),
 ('잼슴', 2),
 ('필요없다', 2),
 ('가슴', 1),
 ('고마워요', 1),
 ('니까', 1),
 ('다시', 1),
 ('보는', 1),
 ('봐고', 1),
 ('봐도', 1),
 ('심심할', 1),
 ('아리구나', 1),
 ('아빠', 1),
 ('안됄', 1),
 ('액션', 1),
 ('으로', 1),
 ('이건', 1)]
```

# 결 론

짧은 리뷰 집단과 긴 리뷰 집단, K=2

긴 리뷰 집단

```
[('너무', 27),
 ('정말', 21),
 ('진짜', 21),
 ('ㅋㅋ', 20),
 ('감동', 16),
 ('드라마', 12),
 ('이영화', 12),
 ('마지막', 11),
 ('봐도', 11),
 ('으로', 11),
 ('봤는데', 10),
 ('수작', 10),
 ('에서', 10),
 ('평점', 10),
 ('보●', 9),
 ('사랑', 9),
 ('이다', 9),
 ('ㅠㅠ', 8),
 ('생각', 8),
 ('연기', 8)]
```

```
[('최고', 34),
 ('정말', 11),
 ('영화', 10),
 ('드라마', 7),
 ('감동', 3),
 ('인생', 3),
 ('긴장감', 2),
 ('너무', 2),
 ('마지막', 2),
 ('액션', 2),
 ('역시', 2),
 ('완전', 2),
 ('ㅋㅋㅋ', 1),
 ('ㅎㅎㅎㅎㅎ', 1),
 ('ㅠㅠ', 1),
 ('ㅠㅠㅠㅠ', 1),
 ('ㅡㅁㅡ', 1),
 ('가는줄', 1),
 ('가득한', 1),
 ('가봤던', 1)]
```

# 감사합니다!

데븐데븐였습니다!

“