# 2023.04.06 미팅 자료

Scenario, TextRank, Recommend Module, Multi-Classification

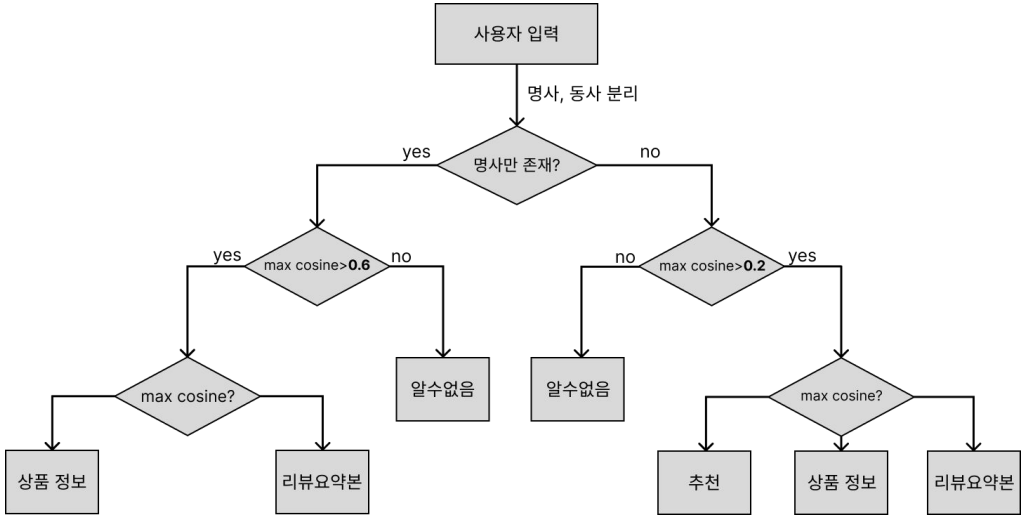
팀원 : 배한성, 김은서, 조현아, 조유진

# CONTENTS

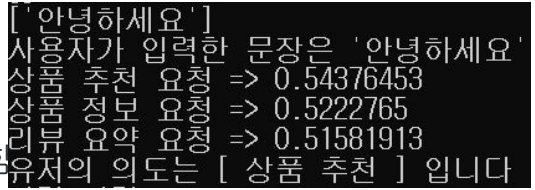
# Scenario

# Scenario

- 진행 상황
  - 카테고리 내의 물품들에 대한 질문을 작성했을 경우 대부분 의도를 잘 분류한다.



# Scenario




- **전혀 관계없는 문장임에도 유사도가 높게 나오는 경우 존재**
  - ex) 안녕하세요, 감사합니다 등
  - 최대 코사인 유사도가 0.4 이상이기 때문에 유효한 문장으로 판단함
  - 임시 해결 방안
    - 현재) “딥러닝 할만한 노트북 추천해줘” 입력 시,
      “딥러닝 할만한 노트북 추천해줘” 그대로 넣어 유사도 구함
    - **해결방안)** “딥러닝 할만한 노트북 추천해줘” 입력 시,
      “~~딥러닝~~ 할만한 ~~노트북~~ 추천해줘” 넣어 유사도 구함
      + 0.4 대신 더 높은 값(0.65) 사용
      + 추천 유사도 판단하는 문장 추가
- **오타가 있는 경우**
  - ex) **노ㅌ一북** 추천해줘
  - 네이버 맞춤법 검사기 API인 hanspell 로 문제해결을 시도해 봤으나 오타에는 별 효과가 없음

# TextRank

# TextRank란?

- 문서 내의 문장 또는 단어를 이용하여 문장의 Ranking을 계산하는 알고리즘
- 문서 집합을 요약하는 방법으로 키워드와 핵심 문장을 선택
- summarization(문서 집합을 요약하는 분야)에서 extractive approaches(통계 기반으로 작동)

**알고리즘 설명**

- word graph나 sentence graph를 구축한 뒤 Graph ranking 알고리즘인 PageRank를 이용하여 각각 키워드와 핵심 문장을 선택 -> 이들을 이용하여 주어진 문서 집합을 요약

- 핵심 문장을 선택하기 위해서
  1. 문장 간 유사도를 기반으로 sentence similarity graph를 생성
  2. 각각 그래프에 PageRank를 학습하여 각마디(단어 혹은 문장)의 랭킹을 계산
  3. 이 랭킹이 높은 순서대로 키워드와 핵심 문장이 됨.

# TextRank 모듈 테스트

- tokenizer - KoNLPy의 Komoran을 사용
- 명사, 형용사, 동사, 어간의 품사만 이용하여 단어 그래프 만들기
- textrank의 KeysentenceSummarizer의 인자로 문장의 최소 유사도를 0.6로 설정
- 데이터 - keyboard 리뷰 데이터에서 1점, 5점만 사용
- summarization - 1점 리뷰 데이터 요약 (735개 문장)

  5점 리뷰 데이터 요약 (11,510개 문장)

```
==========부정 요약==========
벌크 제품인지 누가쓰다 환불한 제품인지판매자만 알겠지만 돈받고 파시는건데최소한 배송중에 파손은 안되겠끔 해주고 보내셔야죠
스마트픽이라고 더 좋을줄 알고 했는데 보관이 엉망 박스는 밟았는지 찌그러져 있고 상품박스 헌거 처럼 얼룩지고다행이 상품은 정상인것 같아서 그냥 쓰기로 다시는 스마트픽 안할것 같네요
제품자체에 뽁뽁이같은것도안들어있는데 상자를이렇게큰데다보내면....받아서들고오는데도 이리저리왔다갔다부딪히는데 배송되면서얼마나부딪히면서왔을지새제품샀는데 기분이안좋네요
==============================

==========긍정 요약==========
아이패드를 구매하고, 키보드를 찾다가 제일 많이 추천하는 걸 사자해서 사게되었는데, 단점이라곤 한영 바꾸는거랑, 영어 대소문자 구분하는 데에 있어서 조금 힘들분 사용하는 부분에서는 크
블루투스 연결도 편하고 사용 중인 노트북 키보드보다 소리가 작고 키감이 좋아 요즘은 이 키보드만 쓰고 있습니다
본래 사용하던 아이패드 폴리오 키보드가 고장 위기에 있어서 미리 구매하였는데, 휴대용이 목적이였는데, 생각보다 무거워서 패드랑 각각 들고 다녀야 할것 같아서 번잡스러우나, 키감이 좋아
==============================
```

# Recommend Module

# Recommend Module

- Apache Solr에 5점 리뷰에 대한 Embedding Vector를 삽입
  - Vector를 삽입하기 위해 Dense vector search를 사용
  - Dense vector search를 사용하기 위해 solr의 스키마에 Dense vector field를 삽입

- 현재 Laptop에 대한 리뷰만을 저장해 두었다.

- 아래의 그림은 전체 vector를 query한 예이다.


```
"response":{"numFound":705075,"start":0,"maxScore":1.0,"numFoundExact":true,"docs":[
    {
      "product_id":[2],
      "vector_id":[5545],
      "sentence":["드라이브에 setup 클릭해야지만 나머지 드라이브를 설치해주니깐 필수적으로 꼭 백업 해야 합니다."],
      "vector":["-0.7030784",
        "0.3167229",
        "0.4722493",
        "0.43294865",
        "-1.0788196",
        "-0.53048223",
        "-0.6283272",
        "-0.9949646",
```

# Recommend Module

- solr에서 vector를 search할 때 유사도 검색을 할 수 있다.
  - solr 공식 문서에서 vector간 유사도를 검색하는 예시
  - &q={!knn f=vector topK=10}[1.0, 2.0, 3.0, 4.0]&fq={!frange cache=false l=0.99}$q
- 따라서 추천 받을 질문을 임베딩 벡터로 변환한 후 query에 함께 넣어 검색하였다.
- cosine 유사도를 검색할 때 상품 별로 검색하여 하나의 상품에 대한 유사도 점수를 구하였다.
- 하지만 순차적으로 처리를 하니 실행 시간이 오래 걸리는 문제가 있었다.
  - 하나의 상품마다 스레드를 생성하여 동시에 query를 하니 시간이 상당히 감소되었다.

# Recommend Module

- Solr에서 cosine 유사도를 구할 때 KNN(K-Nearest Neighbor)를 사용한다.
  - 따라서 K값을 적절하게 설정해야한다.

- 적절한 K 값을 구하기 위해 하나의 상품에 대한 벡터의 개수를 적절하게 나누어 실행 시간을 구하였다.
  - 실행시간이 빠르고 적당한 표본이 모이는 구간을 찾을 수 있었다.
  - 15%, 10%, 5%를 사용하는 구간은 10초에서 30초 사이의 시간이 소요되었다.
  - 또한 순서대로 114, 93, 42개의 유사한 값이 추출되었다.

- 팀원간 상의한 결과 10%와 5%의 중간 값인 7%를 사용하기로 결정

```
query : 가벼운 노트북 추천해줘
run time : 16.593363285064697
추천하는 상품은  삼성전자 갤럭시북 S NT767XCM-K38 입니다.
점수는 64 입니다.
```

# Recommend Module

- 기존 query에서 구한 유사한 값을 가지는 벡터의 값이 올바른지 확인하기 위해 원래 문장을 확인해 보았다.

- 결과를 확인해보니 “가벼운 노트북 추천해줘”를 임베딩 벡터로 변환하여 유사도를 구하면 리뷰에 노트북. 추천이라는 문구가 있으면 가벼운이라는 단어가 없음에도 유사하다고 출력되었다.
  - 위 현상은 Solr의 Dense vector search를 사용하지 않고 cosine similarity 라이브러리를 사용해도 발생하는 문제였다.
  - K의 값이 클 수록 이러한 현상이 많이 발생하였다.
  - 선택한 값인 7%는 이러한 현상이 적지만 추천 결과에 약간 영향을 주었다.

- 따라서 “가벼운 노트북 추천해줘”라고 질문을 하였을 때 “가벼운”만 입력하면 더 높은 정확도를 얻을 수 있을 것이라고 판단하였다.

# Recommend Module

- “가벼운 노트북 추천해줘”라고 입력하였을 때 “가볍다”를 분리하는데 성공하였다.
  - 이 작업은 별도의 모듈로 제작하였다

- “가벼운”만 입력하였을 때 질문의 의도에 맞는 리뷰들이 선택되었다.

```
query : 가벼운
run time : 16.10829186439514
[('LG전자 그램16 16ZD90P-GX50K', 48), ('삼성전자 갤럭시북 S NT767XCL-KLTE', 46), ('삼성전자 갤럭시북 이온2 NT950XDZ-A58AW', 35),
추천하는 상품은  LG전자 그램16 16ZD90P-GX50K 입니다.
점수는 48 입니다.
```

- “가벼운” 과 “가볍다”의 결과가 달라 적절한 작업이 필요해 보인다.

```
query : 가볍다
run time : 12.050699472427368
[('삼성전자 갤럭시북 S NT767XCL-KLTE', 44), ('LG전자 그램16 16ZD90P-GX50K', 42), ('삼성전자 갤럭시북 이온2 NT950XDZ-A58AW', 26)
추천하는 상품은  삼성전자 갤럭시북 S NT767XCL-KLTE 입니다.
점수는 44 입니다.
```

# multi-classification

# multi-classification

- labeling 진행상황

desktop, labtop, mouse, keyboard, monitor 5가지의 카테고리 내에서
디자인, 무게, 성능, 소음, 사이즈, 만족감 6가지 label 로 분류하여 labeling 진행.

- multi-classification 모델 진행상황

Simple GRU를 사용하는 multi-classification 모델을 발견하여 쇼핑몰 Review Dataset에 적용.
Simple GRU = LSTM 모델의 간소화된 버전. 텍스트 분류, 텍스트 생성, 언어번역에서 사용되며
학습이 빠르다는 장점이 있고 데이터의 양이 적을때도 괜찮은 성능을 보이는 모델이다.

# multi-classification

labeling

| review | score | label | design | weight | performance | noise | size | satisfication |
|---|---|---|---|---|---|---|---|---|
| 사무실 업무용이라 교환할 시간도 없어서 직접 LG전자 서비스센터 갖고 가서 수리 맡기고 익일 다시 찾아오고 서비스센터 직원분 말씀으로는 공장에서 확인 후 출고하기 t | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 새 제품을 받고 이렇게 고생할 줄이야 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 참고 바랍니다 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 웬만해서는 리뷰 같은 건 남기지 않으나 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 최악의 판매업체입니다 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 배송도 늦어서 불편하게 하더니 모니터 자체도 이상한 제품을 보내놓고 교환도 구매자 보고 AS 점검받고 하라고 하는 업체입니다 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 절대 이곳에서 구매 안 하시길 권장합니다 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 별 하나도 아까워요 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 오줌 액정 쩝니다 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 그냥 모니터는 눈 건강을 위해 비싸고 더 좋은 거 사세요 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 오줌 액정 신경 안 쓰시는 분들은 괜찮을 듯 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| LG 제품 비추천 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 후회막급 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 화질이 개떡입니다 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 같이 산 동료도 후회 중입니다 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 눈이 아파서 블루 라이트 차단 안경 써서 하루 종일 쓰고 있습니다 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 색감도 이상하고 글자 번짐도 심합니다 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 돈 없는 학생들의 인강용 모니터입니다 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 업무용으로 절대 사지 마세요 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 왼쪽 위부터 아래로 색이 사진과 같이 연두색으로 변색 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 자세히 보니 왼쪽 윗부분 베젤이 유격되어 있었음 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 눌러서 제대로 끼우니 그나마 조금 나아진 상태 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 제가 tv 스크린 모니터가 필요 했는데 잘못 구입했습니다 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 교환 안된다니 그냥 씁니다 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 이참에 티브이 끊을라고요 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 같은 거로 듀얼로 사용하려고 구매 했는데 눈에 띄게 모니터 색상이 달라요 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 모니터 한 개는 갑자기 계속 화면이 덜덜거리네요 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 화질이 진짜 별로네요 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |

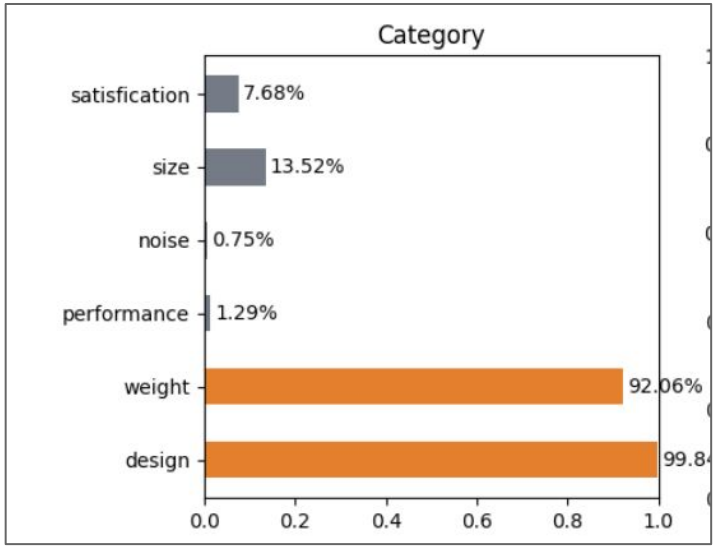
# multi-classification

accuracy

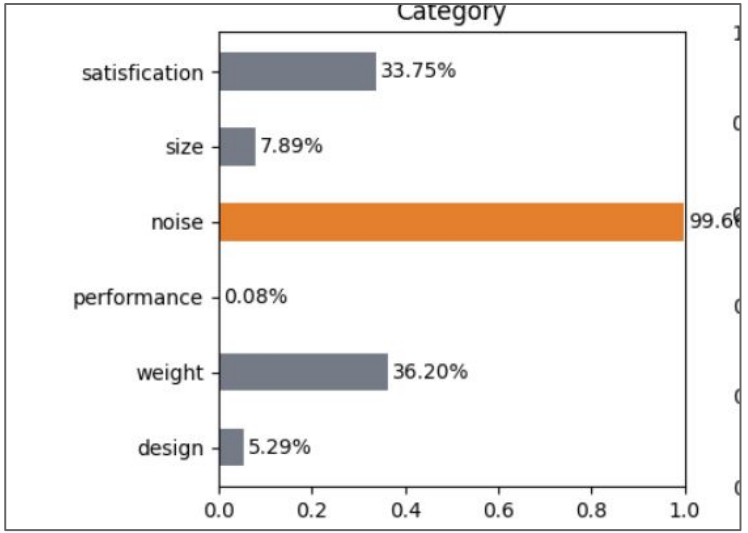
0.8239763379096985]

predict

디자인은 이쁜데 무거워요



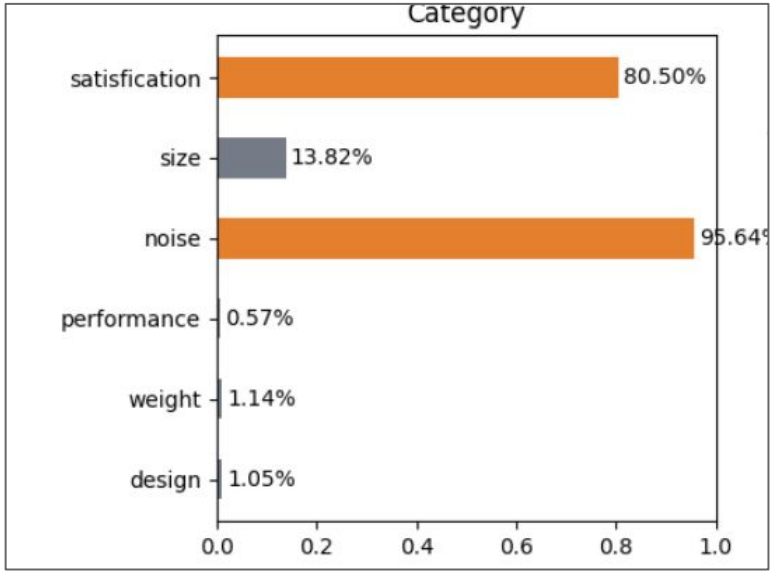

소리가 생각보다 조용해서 좋아요 근데 조금 무거워서 들고 다니기



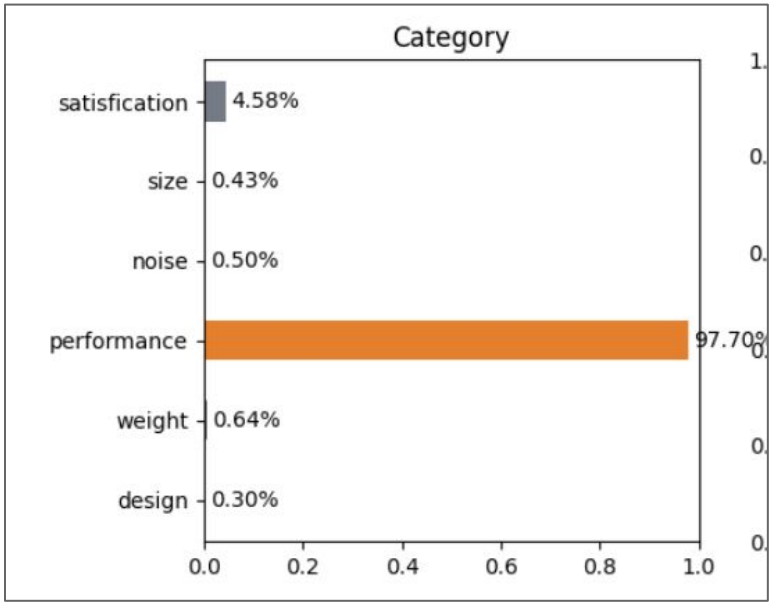
# multi-classification

predict

마우스가 생각보다 조용해서 맘에드네요



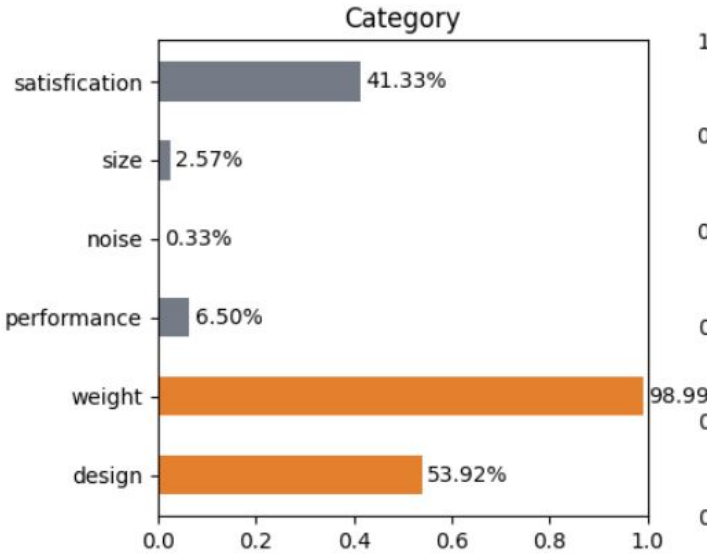

너무 버벅여서 사용하기 힘드네요



# multi-classification

★★★★★ 5 HP공식스토어 · lees**** · 21.03.26.

깔끔 빠른 배송 사은품까지 일단 화이트 너무 이쁘고 마감처리도 고급스럽구요 무게도 많이 무거운 편은 아니에요 여성분들도 한손으로 그냥 쉽게 들 수 있을 정도?? 자판은 너무 이쁘

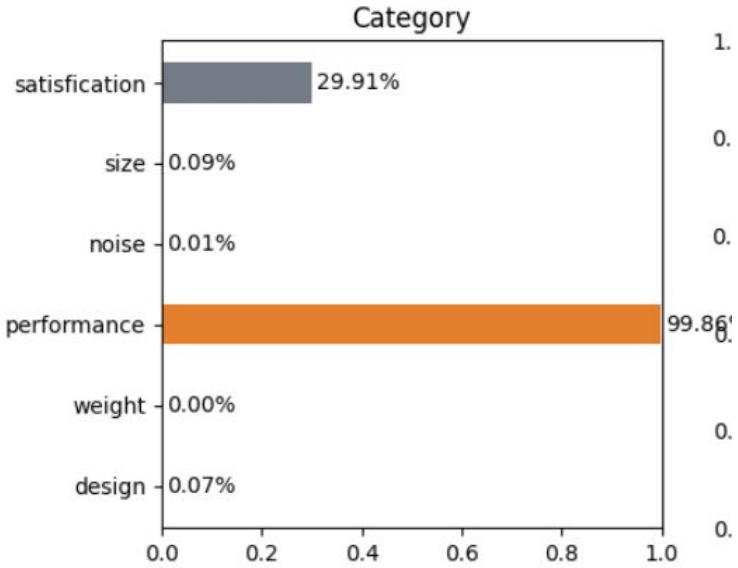


# multi-classification

★★★☆☆ 3 11번가 · da****** · 21.06.23.

보통이에요

전반적인 성능은 만족스럽습니다. 검증된 르누아르라 뭐 딱히 흠잡을 데는 없습니다. 기본 램 4G 는 좀 아쉽긴 합니다. 최소한 4G 하나는 더 추가하는 걸 추천합니다.



# multi-classification

"좋습니다. 일단 건전지 포함해도 무게가 많이 나가지 않아서 g102랑 비슷하고 사진에서 보이듯이 g102보다는 살짝 작은게 손에 쏙 들어옵니다. 손이 큰 남성분이 아니라면 크기 때문에 어려움을 겪지는 않을 것 같습니다"