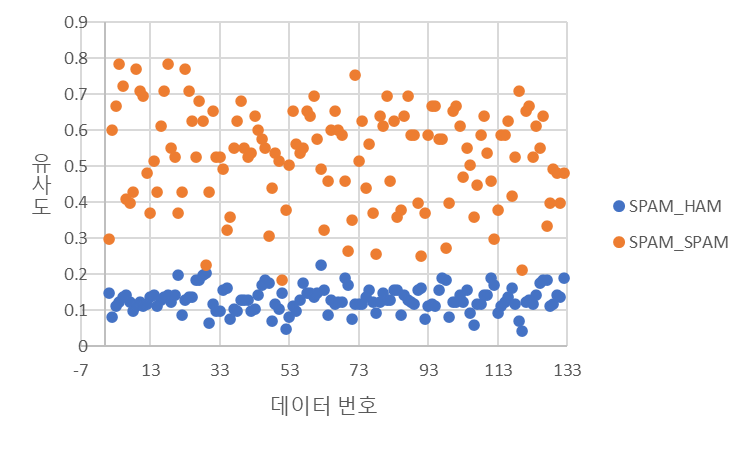
**자카드 유사도**

- 지난주 20개의 표본만으로 자카드 유사도의 평균

-> 표본의 수가 적어서 대표하지 못 할 수도 있을 것이라고 생각

--> 133개의 전체 SPAM DATA 전체를 가지고, 자카드 유사도의 평균 뿐 아니라 분포도를 통해 어떤 특징이 있는지 확인

결과(Jaccard\_out133.txt)



평균 : SPAM\_HAM : 0.130714, SPAM\_SPAM : 0.528374

분포 :

- SPAM\_HAM의 경우 대다수의 값이 0.1과 0.2사이에서 분포

-> SPAM data와 개인 SMS data간에는 대부분 유사도가 0.2 이하이기 때문에 관련도가 없고,

Spam filtering이 잘 되어지는 것으로 보여짐.

- SPAM\_SPAM의 경우 유사도가 0.2 부터 0.8에 이르기까지 분포도가 일관성이 없음.

예상 이유

1) SPAM 데이터의 종류가 다양

2) SPAM데이터 내에 SPAM 데이터가 아닌 것이 존재?

**BERT 모델 이용한 유사도 측정**

(Ko-Sentence-BERT-SKTBERT)

https://github.com/BM-K/KoSentenceBERT-SKT

SKT KoBERT를 사용하여 SentenceBERT 학습

STS 단일 데이터를 이용한 Fine-tuning

https://colab.research.google.com/drive/1uGHdcpIslcegcfmxKt0q4TTP9kmIuIwX?authuser=1

위의 링크에서 제시하는 예제 실행

->필요 패키지 설치를 비롯한 환경설정을 하느라 아직 실행 실패

colab을 사용하는데 SPAM data를 활용해도 되나요???

BERT로 유사도 측정하는 방법

기존의 BERT로는 large-scale의 유사도 비교, 클러스터링, 정보 검색 등에 많은 시간 비용이 들어간다.

Sentence-BERT(SBERT)는 BERT의 출력에 풀링 연산을 추가한 모델

풀링 방법은 [CLS] 토큰의 결과를 사용하는 방법, 모든 출력 벡터를 평균내어 사용, 출력 벡터의 max-over-time을 계산해 사용하는 방법

기본적으로 SBERT는 평균 풀링을 사용하며, 평균 풀링으로 문장 표현을 얻으면 이 표현은 본질적으로 모든 단어의 의미를 갖는다. 반면 최대 풀링으로 문장 표현을 얻을 경우 문장 표현은 본질적으로 중요한 단어의 의미를 갖는다.

(Max-over-time pooling: 문장의 길이가 다 다르면 문장마다의 feature map 개수가 달라지는데, 모든 문장마다 하나의 값을 갖도록 feature map 벡터 중 가장 큰 값 하나만 사용하는 것)

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명