**\*\* 스팸데이터와 내데이터를 가지고 유사도 측정을 해보는 실험을 진행**

(hashing 공부)

🡪 유투브 및 구글

key -> hashcode -> Index -> list

한 인덱스에만 여러개 들어가는 collison이 일어나지 않게(분) 하는 알고리즘을 만들어야함

문서 유사도 분석

- 많이 나온 상위 100개 단어(명사)를 기준으로 자카드 유사도 측정 진행

- spam-spam(20개)의 유사도와 spam-mydata의 유사도 비교측정

spam-spam : 1-2번파일, 2-3번파일, 3-4번파일... 식으로 유사도를 측정해서 평균.

spam-mydata : spam1-mydata파일, spam2-mydata파일, spam3-mydata파일... 식으로 측정해서 평균

자카드 유사도, 코사인 유사도, 유클리디안 거리 중에서 자카드 유사도로 진행.

- 자카드 유사도

두 집합의 합집합 중 교집합의 비율

- 코사인 유사도

두 벡터 사이의 각도를 계산하여 두 벡터가 얼마나 유사한지 측정

- 유클리디안 거리

유클리드 거리를 활용하여 문서 간 유사도를 계산

결과: Jaccard\_out.txt

spam-mydata mean : 0.12480076766766927, spam-spam : 0.575712970437022

spam-mydata는 대부분 0.1 ~ 0.14으로 측정

spam-spam은 0.8에 가까운 유사도 값도 나온 적이 있고, 0.7, 0.6이상도 많음.

유사도가 상대적으로 매우 낮은 것 몇 개를 제외 시킨다면 spam-mydata mean과 spam-spam의

차이가 위의 그래프보다 더 많이 남.

어떤 특정단어가 몇번 나왔는지, 중복되는지는 고려하지 않음.

얼마나 다른 종류의 단어가 중복되었는지만 고려함.

결론적으로 현재 SKT에서 spam 데이터를 잘 필터링 해주고 있다는 생각이 듭니다.

**\*\* 트랜스포머 BERT 유투브나 구글을 통해서 스터디 및 예제실습 진행**

- 구글에서 개발한 NLP(자연어처리) 사전 훈련 기술

- 특정 분야에 국한된 기술이 아니라 모든 자연어 처리 분야에서 좋은 성능을 내는 범용 Language Model

- BERT를 사용하지 않은 일반 모델과정

: 분류를 원하는 데이터 -> LSTM, CNN 등의 머신러닝 모델 -> 분류

- BERT를 사용한 모델링 과정

: 관련 대량 코퍼스 -> BERT -> 분류를 원하는 데이터 -> LSTM, CNN 등의 머신러닝 모델 -> 분류

토큰화된 단어 하나만 보면 의미를 알기 어려움 -> 토큰의 의미를 구하기 위해 attention 벡터를 사용