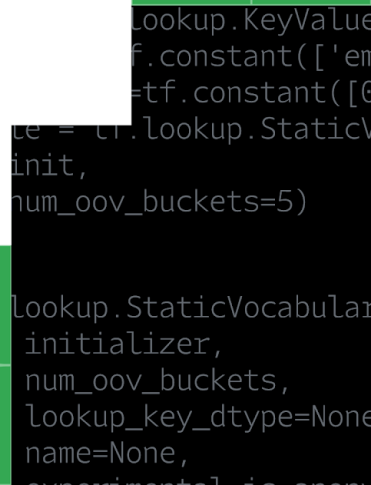
Google Developer Student Clubs

4주차 2조

팀원: 황동욱, 권도혁, 송성근

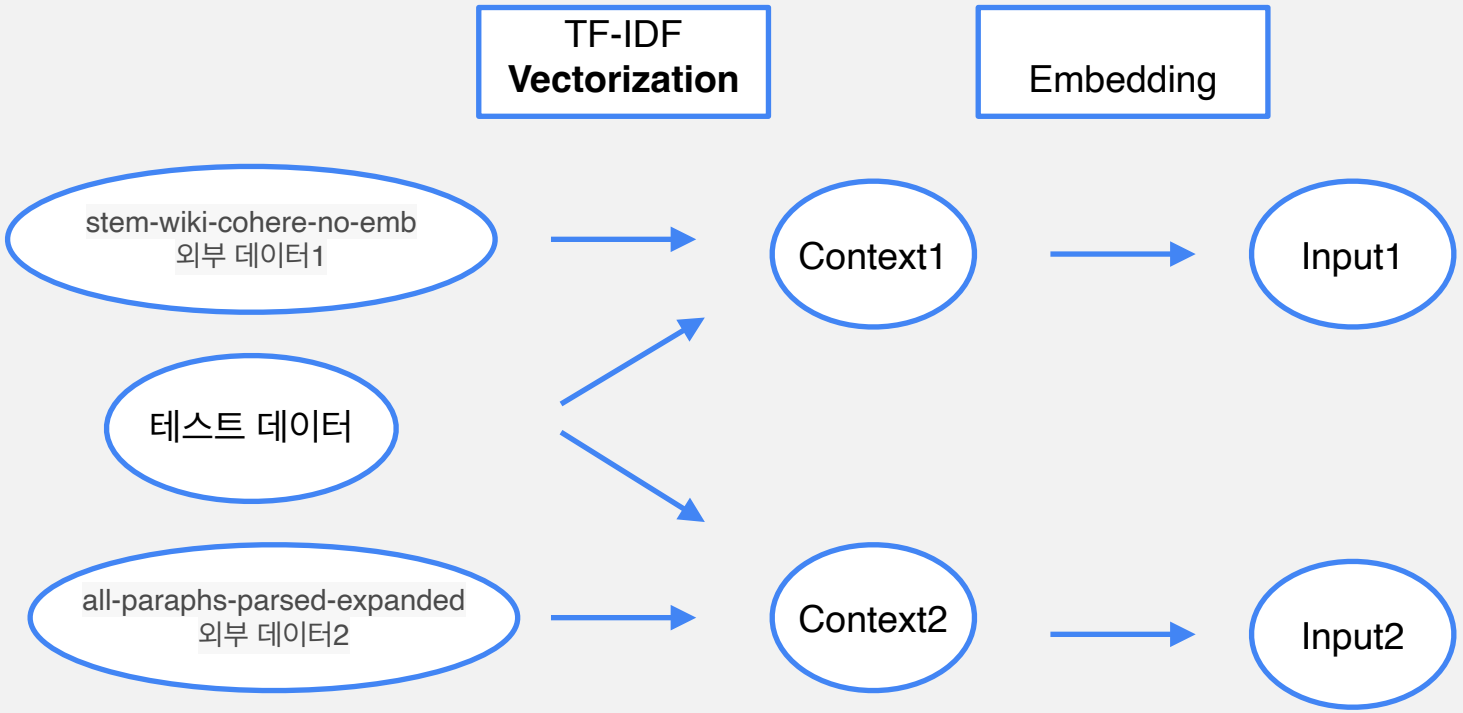


TF-IDF

벡터라이저란? 그리고 생성하는 이유

- 벡터라이저는 텍스트 데이터를 수치형 벡터로 변환하는 도구입니다.
- TF-IDF(Term Frequency-Inverse Document Frequency) 벡터라이저는 텍스트 데이터를 벡터로 변환하는 방법 중 하나입니다. TF-IDF는 각 단어의 중요도를 나타내는 값으로, 단어가 문서 내에서 얼마나 자주 등장하는지(TF)와 전체 문서 집합에서 그 단어가 얼마나 희귀한지(IDF)를 고려하여 계산됩니다.

데이터 전처리


TF-IDF
Vectorization
Embedding
stem-wiki-cohere-no-emb
외부 데이터1
테스트 데이터
all-paraphs-parsed-expanded
외부 데이터2
Context1
Context2
Input1
Input2

Longformer

- LongFormer는 BERT와 같은 표준 트랜스포머 모델의 한계를 극복하려는 시도로 개발된 모델입니다. 표준 트랜스포머 모델들은 입력 텍스트의 길이에 제한이 있지만, LongFormer는 긴 문서도 처리할 수 있도록 설계되었습니다.
- "제한된 GPU 메모리에서 훨씬 긴 접두사 컨텍스트를 가질 수 있다"는 말은 LongFormer 모델을 사용하면 GPU 메모리의 한계 내에서도 더 긴 입력 텍스트를 처리할 수 있다는 것을 의미합니다.

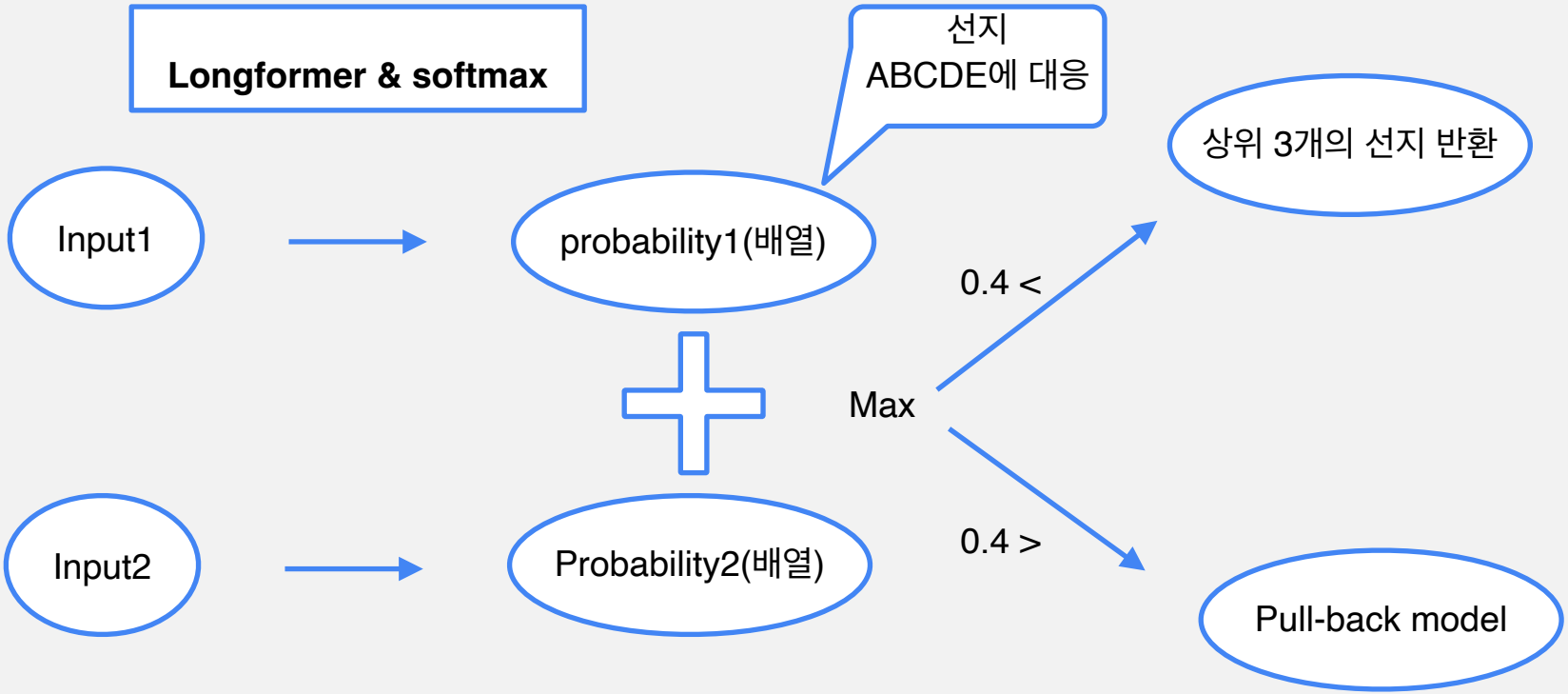
- **장점**:
  - **긴 문서 처리**: LongFormer는 긴 문서를 처리할 수 있는 특별한 attention 메커니즘을 가지고 있어, BERT와 비교할 때 훨씬 긴 텍스트를 처리할 수 있습니다.
  - **효율성**: 긴 문서를 처리할 때 효율적인 attention 계산을 위해 설계되었습니다.
- **단점**:
  - **특수성**: LongFormer는 긴 문서를 처리하는 특수한 경우에 특화되어 있어, 모든 NLP 작업에 적합하지 않을 수 있습니다.

Longformer

Attention Scope:

- **BERT**: BERT는 표준적인 self-attention 메커니즘을 사용합니다. 이는 입력 시퀀스의 모든 토큰 간의 상호 작용을 계산합니다. 따라서 BERT의 attention은 O(n^2)의 복잡도를 가집니다. 여기서 n은 시퀀스 길이입니다. 이러한 계산 복잡도 때문에 BERT는 긴 문서를 처리하는 데 어려움이 있습니다.
- **Longformer**: Longformer는 긴 문서를 처리할 수 있도록 설계된 attention 메커니즘을 사용합니다. 이는 global attention과 local attention의 조합으로 구성됩니다. 특정 중요한 토큰은 전체 문서에 걸쳐 attention을 받는 반면, 나머지 토큰은 주변의 토큰만을 대상으로 attention을 수행합니다. 이러한 방식으로 Longformer는 O(n)의 복잡도로 확장됩니다.

학습과 결과 반환


Longformer & softmax
선지
ABCDE에 대응
상위 3개의 선지 반환
Input1
probability1(배열)
0.4 <
Max
Input2
Probability2(배열)
0.4 >
Pull-back model

시도

1. TfidfVectorizer parameter를 튜닝

   max_df: 문서 빈도의 최대 임계값

   min_df: 문서 빈도의 최소 임계값입니다

   token_pattern: 토큰화 할 때 사용할 정규 표현식 패턴

2. Stop word 수정

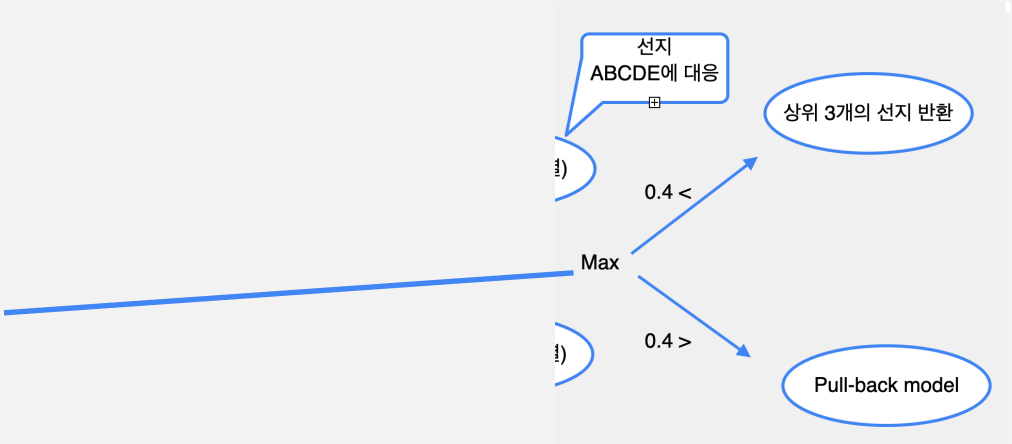
   무시할 단어의 목록

3. 임계값 수정

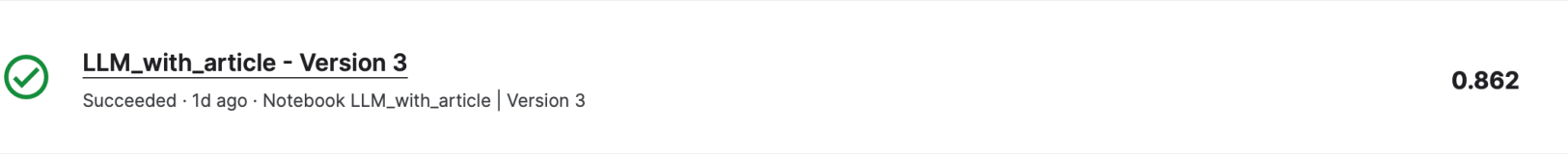
   임계값을 0.4에서 다른 값으로 수정

4. 불필요한 문자 제거

   특수 문자와 같은 문자 제거



Kaggle 제출 결과


LLM_with_article - Version 3
Succeeded · 1d ago · Notebook LLM_with_article | Version 3
0.862