**OpenAI Embedding**

**1. 학습 방식**

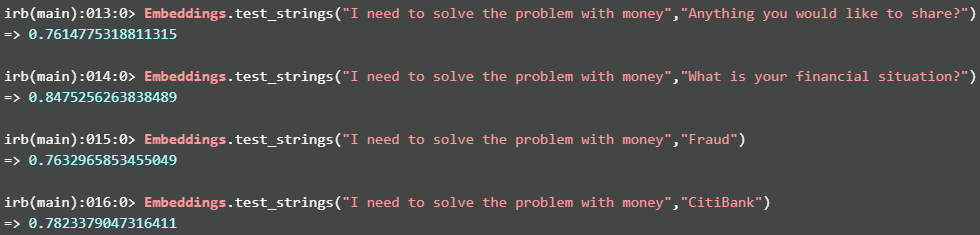
- OpenAI의 최근 임베딩 모델인 ‘text-embedding-ada-002’은 성능을 제외하고, 학습 방법이나 데이터에 대해서 공개된 것이 없어 정확한 분석이 어렵다. 이전 임베딩 모델의 경우 쌍으로 나눠진 학습 데이터로 Contrastive Pre-Training을 진행했다. 하지만 성능과 가격에 차이가 있기 때문에 text-embedding-ada-002가 동일한 방식으로 학습되었다고 보기는 어렵다.

- OpenAI 커뮤니티에서 임베딩 모델과 관련된 실험을 많이 진행해본 [raymonddavey](https://community.openai.com/t/some-questions-about-text-embedding-ada-002-s-embedding/35299/53)의 말에 의하면 OpenAI의 임베딩은 영어-영어 유사도가 한글-영어의 유사도보다 높게 나오도록 학습되었다고 한다.

**2. cosine similarity의 범위(-1~1)**

- 여러 랜덤 문장으로 임베딩을 진행해본 결과 cosine similarity의 범위가 0.7~1의 범위로 계산이 된다. 임베딩 사이 각도의 범위가 매우 작다고 할 수 있다. 이는 한국어 문장 뿐만 아니라 영어 문장도 비슷한 결과를 보인다.

- 이러한 이유는 위에서 같은 언어일 수록 유사도가 높게 나오도록 학습했기 때문이라고 볼 수 있다.



**3. text-embedding-ada-002의 영어 성능에는 문제가 없다**

[임베딩 모델 리더보드](https://huggingface.co/spaces/mteb/leaderboard)

- 태스크: Information Retrieval(right document retrieval for a given query.)

- 언어: 영어

- Metric: Normalized Discounted Cumulative Gain@k (k=10)

- 같은 언어가 높은 유사도를 가졌어도 정보 검색 태스크의 성능에는 문제가 없는 것으로 확인됐다.

- 예전 임베딩 모델 중 가장 성능이 좋은 모델인 davinci-001 보다 Retrieval 성능이 더 좋고 가격도 싸다.

- Huggingface의 Retrieval Task에서 sota는 아니지만 7등으로 좋은 성능을 보여주고 있다. Retreival에서 가장 좋은 성능을 보여주는 모델인 multilingual-e5-large와 2.1정도 차이밖에 나지 않는다.

- **한글 임베딩이 가능함과 동시에 최대 길이가 8k가 넘는 오픈 임베딩 모델은 존재하지 않는다.**

**4. 학습 데이터로 인해 text-embedding-ada-002에 bias가 존재**

- 아프리카계 미국인보다 유럽계 미국인 이름에 긍정적 감정을 가지는 경향이 있다.

- 흑인 여성에게 부정적인 고정관념을 가지는 경향이 있다.

- 이 외에 보이지 않는 편향이 존재할 수 있다.

**5. 한국어 질문으로 세션이 잘 안걸리는 이유**

1. BPE(Byte Pair Encoding)

- 한국어와 영어를 BPE로 토크나이징을 할 경우가 다르다.

- 예를 들어 **웹투비**의 경우 웹->토큰 2개, 투->토큰 3개, 비->토큰 1개로 나눠지게 된다. 따라서 **웹툰**과의 유사도가 높아진다.

- WebtoB로 토크나이징 할 경우 Web, to, B로 토크나이징 된다.

2. 영어로 이루어진 주요 Keyword들

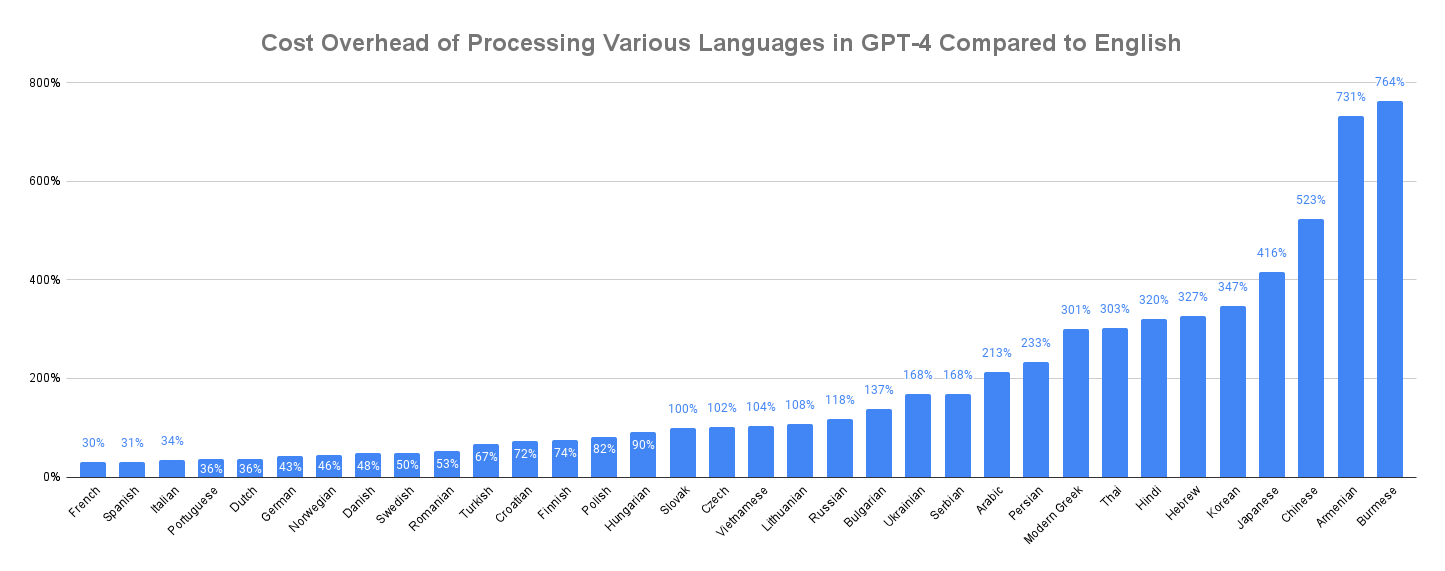
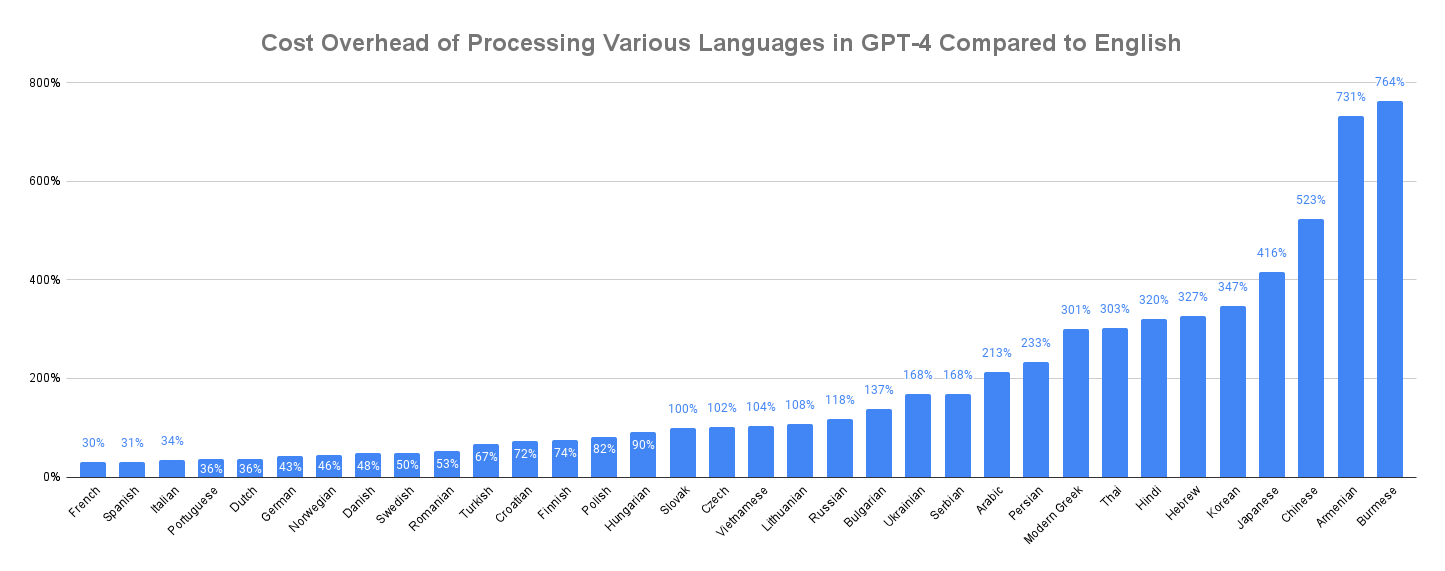
- WebtoB, Jeus 등 주요 키워드 들이 영어로 이루어져 있고, 임베딩을 할 때 위에서 말했던 것처럼 [Web, to, B]와 [Je,us] 등의 토큰으로 나눠지게 된다.

- 그래서 키워드를 한글 웹투비나 제우스로 검색을 하게 되면 영어로된 키워드를 잘 찾지 못하고, 주변에 있는 단어들에 초점을 더 준다.

- 이러한 고유 명사는 단어의 변형 없이 웹투비 또는 WebtoB를 그대로 쓰기 때문에 후처리로 웹투비를 찾아 WebtoB로 변형해주면 큰 문제가 되지 않는다.

3. 외래어(스택트레이스, 스레드 etc.)

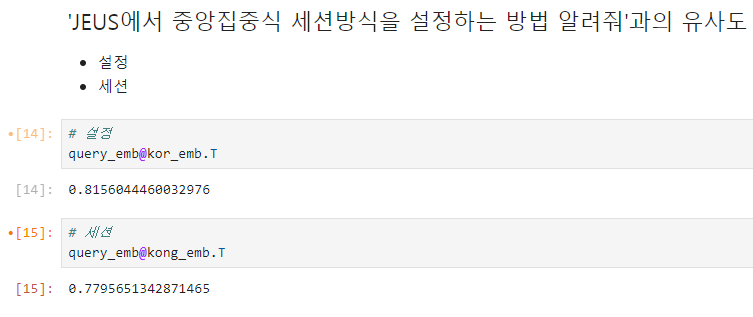
- OpenAI의 gpt3.5, gpt4의 토크나이저의 사전에 한국어가 많지 않다. 아래 그림에서 알 수 있듯이, gpt4가 한국어를 처리하는 데 드는 비용은 영어의 3.5배이다. 같은 문장이어도 한국어로 하면 토큰이 3.5배정도 더 많다고 대략적으로 알 수 있다.



- ‘스레드’ 같은 외래어의 경우 4개의 토큰으로 나눠진다. 이렇게 하나의 뜻을 가진 단어를 여러개의 토큰들로 나눠진 토큰들을 임베딩 했을 때 가지는 의미는 같은 단어를 영어‘thread’로 임베딩 했을 때 가지는 의미와 거리가 멀어진다.

- **[가설]** 외래어같이 여러개의 토큰으로 나눠지는 경우 어텐션 구조에서 주변 단어들의 어텐션 스코어가 더 커져서 결과적으로 주변 단어의 중요도가 높아지는 것 같다.

한국어 세션에 이러한 외래어가 많아지면 설정이나 방식같이 주변 단어의 주요도가 높아지고, 마찬가지로 질문의 임베딩에서도 ‘설정’의 주요도가 높아져 ‘세션’보다 ‘설정’이 들어간 문서의 유사도가 높게 나오는 것 같다. 그래서 문서에 나와있는 형식 그대로 질문을 하거나 외래어가 주요 키워드인 경우 주변에 추가 단어 없이 외래어 키워드만으로 검색해야 외래어 키워드에 집중해서 결과를 반환하는 것 같다.



**6. 결론**

- text 임베딩 모델들이 영어 데이터셋으로만 테스트 되어 다른 언어에서의 성능은 확실하지 않지만, 전체적인 성능 지표를 봤을 때 영어에서 좋은 성능을 내는 모델인 것은 확실하다.

- 길이와 한국어가 지원이 되는 모델로는 OpenAI 임베딩 모델을 대체할 공개 모델이 없다.

- BPE가 다른 언어를 쪼갰을 때 쪼개진 의미가 서로 다르기 때문에 문서와 검색문장이 서로 같은 언어로 임베딩 됐을 때 retrieval 성능이 유지된다.

- 외래어가 포함된 문서나 질문의 경우 주변 단어의 의미에 의존하게 되어 이들이 가지는 임베딩이 의도한 바와 다르게 작동할 수 있다.