토큰(Token)?

토큰(Token)은 자연어 처리(NLP)에서 텍스트를 작은 단위로 나누어 처리하기 위해 사용되는 기본 단위(단어, 부분 단어, 문자 등)입니다. '토큰화'(Tokenization): 텍스트를 토큰으로 나누는 단계입니다.

토큰화의 방법

- 문자 기반 토큰화: 텍스트를 문자 단위로 나누는 방법입니다.
  - 예시: 'Hello' → ['H', 'e', 'l', 'l', 'o']
- 단어 기반 토큰화: 텍스트를 단어 단위로 나누는 방법입니다.
  - 예시: "I love natural language processing." → ["I", "love", "natural", "language", "processing"]
- 서브워드 기반 토큰화: 단어를 더 작은 단위(서브워드)로 나누는 방법입니다.
  - 예시: "hugging face is awesome" → ["hug", "g", "in", "g", "face", "is", "a", "w", "e", "s", "o", "m", "e"]
  - BPE 토큰화 과정:
    - 초기 어휘 집합: {"h", "u", "g", "i", "n", "f", "a", "c", "e", "s", "w", "o", "m"}
    - 빈도수 높은 쌍 병합: "hu", "in", "fa", "ac", "gg"
    - 추가 병합: "hug", "face"
  - BPE(Byte Pair Encoding)알고리즘: 자주 등장하는 문자 쌍을 합쳐가며 서브워드를 생성하는 알고리즘

토큰의 중요성

- 성능: 토큰화의 결과에 따라 모델의 성능에 큰 영향을 줍니다.
- 문맥 이해: 모델이 문맥을 이해하고 적절하게 응답할 수 있도록 도움을 줍니다.
- 효율성: 적절한 크기의 토큰을 사용함으로써 연산 자원을 효율적으로 사용합니다.

토큰화는 자연어 처리의 기본적이면서도 중요한 단계로, 텍스트 데이터를 효과적으로 분석하고 처리하는 데 필수적인 과정입니다.

토큰사용량과 비용

- 비용 계산 방식: 대부분의 AI 모델은 입력 토큰과 출력 토큰 수에 따라 비용을 청구합니다.
- 토큰과 단어의 관계: 일반적으로 1,000 토큰은 영어로 약 750단어에 해당합니다.
- 비교 사이트: https://tiktokenizer.vercel.app/
- 모델 복잡도에 따른 차이: 더 고급 모델일수록 토큰당 비용이 높아집니다
- 입력과 출력의 구분: 대부분의 서비스는 입력(프롬프트)과 출력(완성)에 대해 별도로 비용을 청구합니다

- 언어별 차이: 영어 기반 사용 사례를 중심으로 가격이 책정되며, 다른 언어의 경우 토큰 수가 달라질 수 있습니다
- 배치 처리 할인: 일부 서비스는 배치 API를 통해 처리할 경우 할인을 제공합니다