Characteristic of Tokenization Style

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

토큰 평균 길이에 따른 성격과 특징

*OoV: Out of Vocabulary, <UNK>로 치환

짧을 수록

- Vocabulary 크기 감소
  - 희소성 문제 감소
- OoV가 줄어듬
- Sequence의 길이가 길어짐
  - 모델의 부담 증가
- 극단적 형태: character 단위

길 수록

- Vocabulary 크기 증가
  - 희소성 문제 증대
- OoV가 늘어남
- Sequence의 길이가 짧아짐
  - 모델의 부담 감소

토큰 길이에 따른 Trade-off 존재

정보량에 따른 이상적인 형태

- 빈도가 높을 경우 하나의 token으로 나타내고,

- 빈도가 낮을 경우 더 잘게 쪼개어, 각각 빈도가 높은 token으로 구성한다.

압축 알고리즘?