303: Text Preprocessing

단어를 전처리하는 방법

1. 영어 전처리와 한국어 전처리의 차이
2. 전처리의 목적과 그 흐름
3. 전처리에서 사용되는 라이브러리들

언어적 차이: 영어와 한국어

	영어	한국어
유형	굴절어	교착어
데이터의 띄어쓰기	대체로 엄격함 (풀어쓰기)	엄격하지 않음 (모아쓰기)
기준	띄어쓰기	형태소, 품사태깅

전처리

1. 데이터 정제 및 정규화
   - 특수문자 제거, 오타 및 맞춤법 교정, 띄어쓰기 교정
2. 토큰화 (Tokenizing)
   - 어간 및 표제어 추출
3. 형태소 분석 (POS-Tagging)
4. 불용어 (Stopwords) 제거
5. 패딩 (Padding)

정제 및 정규화

- 정제 : 데이터로부터 노이즈를 제거
- 정규화 : 표현방법을 통합

일반적인 정제 및 정규화 작업

1. 대소문자 통합
2. 불필요한 단어의 제거
    - 특수문자, 빈도가 너무 낮은 단어 등
3. 날짜표기, 숫자 표기 등
    - 두번째, 2번째 …

토큰화가 필요한 이유

말뭉치 corpus 를 의미를 가지고 있는 최소단위인 토큰 token 으로 분해하는 과정

분해되면 의미를 잃는 경우

1. 속담 또는 관용어구
   - Piece of cake (O)
   - Piece + of + cake (X)
2. 지명
   - New York (O)
   - New + York (X)
3. 단어에 구두점이 있는 경우
   - Ph.D (O)
   - Ph + D (X)

분해되어야 하는 경우

1. Apostrophe
   - I've, They've
2. 한국어 용언의 활용
   - 놀며, 놀았, 놀고는 모두 다르다.

토큰화

요약

- 의미를 가지는 최소단위이나 상황에 따라 다르다.
  - 한국어는 토큰 외에도 의미를 가지는 최소단위인 ‘형태소’ 가 있기 때문.
- 보통 띄어쓰기를 기준으로 한다.
- 나누어야 하는 상황과 나누면 안되는 상황을 명확히 알아야 한다.
- 한국어와 영어의 토큰화 방법이 다르다. : 한국어의 토큰화가 더 까다롭다.
- 활용할 수 있는 라이브러리 : (영어) NLTK, SpaCy

형태소 추출과 POS Tagging

Part-of-Speech

토큰화된 문장의 토큰에
품사를 일대일 대응

- 품사의 기준 : Penn Tree bank
- 규칙보다는 확률로 접근: 문맥이 다양하므로
- 방법론
  1. 규칙기반 : NLTK의 Pos tagger
  2. 예측기반 : 맥락을 읽는다.
     1. Deep Learning: 더 효과적
     2. Hidden Markov Model: 데이터 요구량 적음

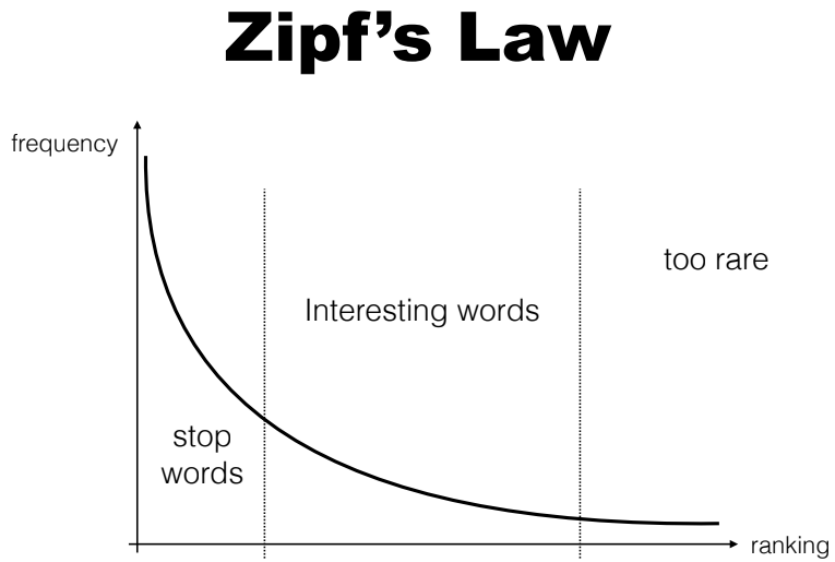
불용어란?

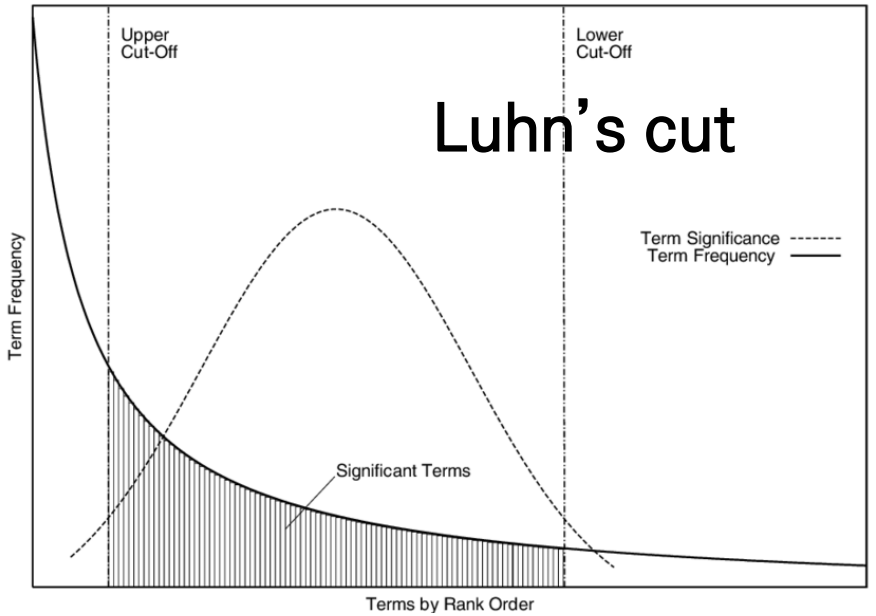
불용어: 자주 등장하지만 큰 의미는 없는 단어들

Count-based Model에서만 사용한다.

- DTM을 만들 때 문장 구성에 필요하지만 큰 의미는 없는 단어 (e.g. be동사, 관사)의 빈도가 높게 나오는 것을 확인했었다.
- NLTK에서는 100여개 이상의 단어를 불용어로 지정하고 있다.

불용어란?





- 어떤 언어든 말뭉치 속 단어를 빈도수가 높은 순으로 역정렬하면 다음 공식을 따른다. $f = \frac{C}{rank}$
- 수학적 통계를 바탕으로 밝혀진 경험적 법칙

- 불용어 찾기의 기반이 되는 이론.
- 너무 많이 쓰이는 단어(Upper cut-off)는 의미가 없고
- 너무 많이 쓰이지도, 적게 쓰이지도 않는 단어가 중요하다.

패딩

Zero Padding 전

This	Is	my	NLP	exercise
I	love	my	parrot	
I	am	tired		

Zero Padding 후

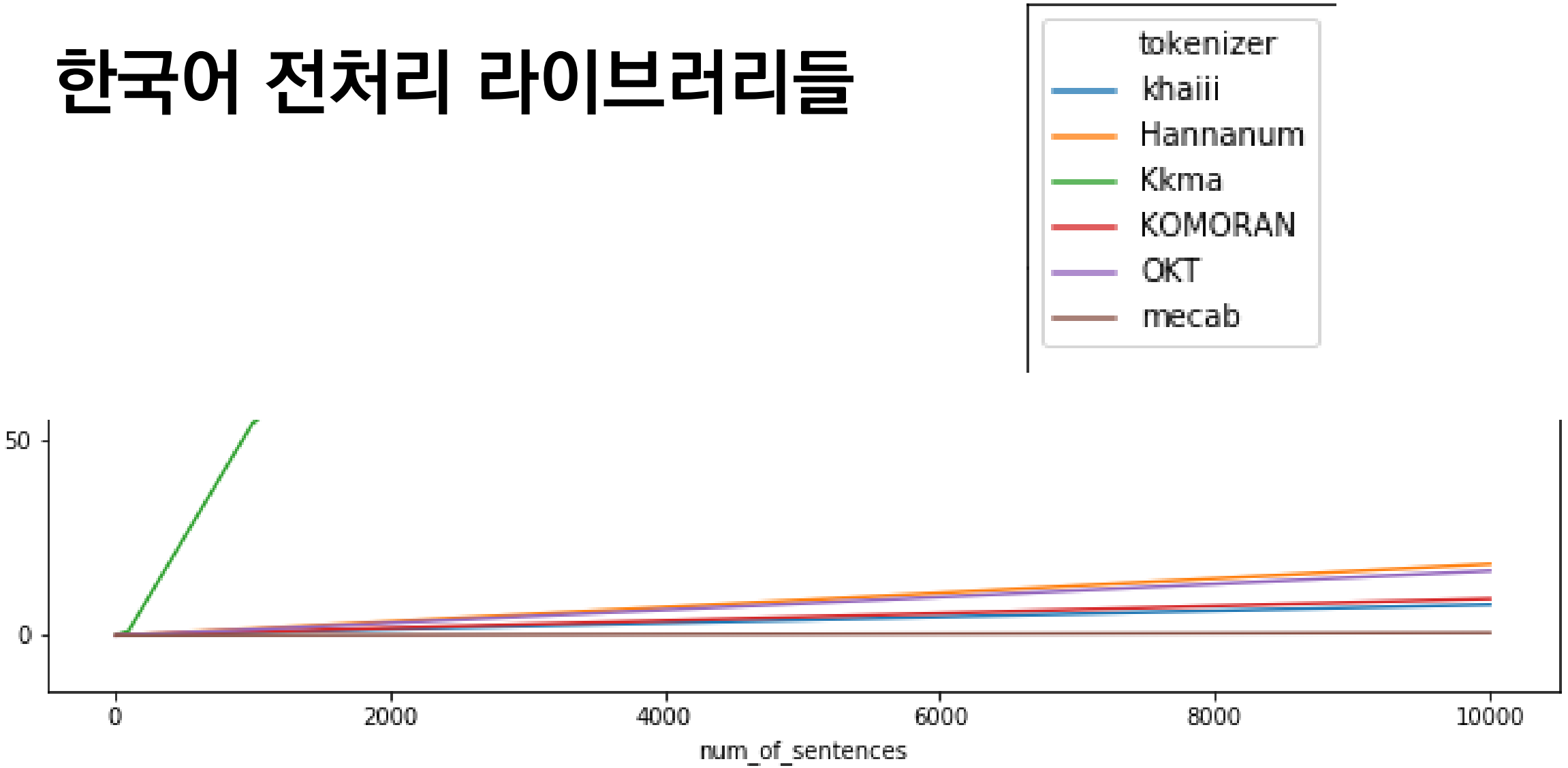
This	Is	my	NLP	exercise
I	love	my	parrot	0
I	am	tired	0	0

한국어 전처리 라이브러리들

모델명	종류	특징
Khaiii	지도학습	카카오 개발, 형태소 분석기, CNN에서 사용
KoNLP		Mecab, 꼬꼬마, 한나눔, Okt, 코모란 등의 5개 오픈소스 형태소 분석기를 한번에 사용할 수 있게 한 한국어 자연어 처리 패키지
SoyNLP	비지도학습	형태소분석, 품사 판별 등을 지원. 대규모 데이터나 문서 집합에서 효과적으로 작동
SentencePiece		구글에서 개발한 subword segmentation

Out Of Vocabulary 일부 해소 가능

한국어 전처리 라이브러리들



한국어 전처리 라이브러리들

모델명	종류	특징
Khaiii	지도학습	카카오 개발, 형태소 분석기, CNN에서 사용
KoNLP		Mecab, 꼬꼬마, 한나눔, Okt, 코모란 등의 5개 오픈소스 형태소 분석기를 한번에 사용할 수 있게 한 한국어 자연어 처리 패키지
SoyNLP	비지도학습	형태소분석, 품사 판별 등을 지원. 대규모 데이터나 문서 집합에서 효과적으로 작동
SentencePiece		구글에서 개발한 subword segmentation

Out Of Vocabulary 일부 해소 가능