데이터 분석 캡스톤디자인

2016100946 김효준

1. 한국어 형태소 분석기 라이브러리 정리 및 비교

■ KoNLPy 라이브러리에서 5가지 형태의 세분화된 방법 제시

① Hannanum : KAIST의 SWRC (Semantic Web Research Center)에서 Java로 작성된 형태소 분석기 및 POS 태거

② Kkma (꼬꼬마 형태소 분석기) : 서울대 지능형 데이터 시스템 (IDS) 연구실에서 개발 한 자바로 작성된 형태소 분석기 및 자연어 처리 시스템

③ Komoran : 2013 년부터 Shineware에서 개발한 Java로 작성된 비교적 새로운 오픈 소스 한국어 형태소 분석기

④ Mecab : 일본 형태소 분석가에 의해 개발되었지만 추후 한국어에 적합하게 수정됨

⑤ Okt : 과거 Twitter()와 동일. 한국어 트위터 기반으로 작성된 라이브러리

■ 5가지 방법 속도 비교

형태소 분석기	한국어 욕설 데이터셋	시간
Hannanum	7896건	24.760
Kkma	7896건	248.536초
Komoran	7896건	5.726초
Mecab	7896건	1.049초
Okt(twitter)	7896건	17.44초

■ 5가지 방법 성능 비교

- 실제 욕설 데이터셋 중 5개를 샘플로 뽑아 실제로 명사만 추출하여 성능 평가

➔ 그 결과, Komoran 과 mecab 형태소 분석기가 가장 적합하다고 판단

① Komoran

- 많은 단어 명사를 포함시킴 (ex. 호빗까지도)
- 간혹 noise 존재 가능(ex. '몰', '메', '걸리', '안',)

② Mecab

- 가장 정제되고 안정된 단어
- 그러나 호빗 같은 단어는 없음 (단어장에 없는 단어는 추출 안되는 현상)