2018학년도 2학기 언어와 컴퓨터

제19강 scikit-learn과 NLTK를 사용한 기계학습

박수지

서울대학교 인문대학 언어학과

2018년 11월 28일 수요일

오늘의 목표

1. 기계학습 과정에 개발 집합을 도입하여 더 나은 모형을 찾는다.
2. `sklearn`을 사용하여 `nltk`를 사용하여 단순 베이즈 분류와 로지스틱 회귀분석(최대 엔트로피 분류)을 수행할 수 있다.

지금 할 일

scikit-learn 설치: `python -m pip install -U sklearn`

개발 집합을 사용한 성능 개선

1. 훈련 집합에서 분류기를 학습시킨다.
2. 개발 집합에서 분류기의 성능을 실험한다.
3. (반복) 개발 집합의 분류 결과를 참조하여 분류기를 "튜닝"하고 훈련 집합에서 다시 학습시킨다.
4. 실험 집합에서 분류기의 성능을 평가한다.

분류기의 성능에 영향을 미칠 수 있는 요소

- 데이터를 어떻게 가공할 것인가?
- 단어 주머니 모형의 "단어"로 N-그램을 사용할 것인가?
- add-k 평탄화에서 k의 값을 얼마로 할 것인가?
- ⋯

텍스트 분류에 사용할 수 있는 라이브러리

- scikit-learn
  - 단어 빈도와 관련된 특성을 간편하고 다양하게 얻을 수 있다.
  - 실습 코드 `naivebayes_sklearn.py`
- nltk
  - 다양한 특성을 쉽게 추가하고 삭제할 수 있다.
  - 실습 코드 `naivebayes_nltk.py`
- …

준비

각 라이브러리에 맞는 데이터 가공