오피니언 마이닝

jongbumi@gmail.com

2017/05/18

오피니언 마이닝이란?

- 감성 분석, 평판 분석 등으로 불리우며
- 글쓴이의 의견을 추출하는 텍스트 연구 분야.
- 의견은 보통 긍정, 부정, 중립으로 구분.

활용

- 제품이나 서비스에 대해 대중의 의견을 구하는 마케팅 분야에서 특히 유용.

- 상품 구매 및 정보 획득에 있어서 평판의 중요성이 더욱 중대되고 SNS 및 블로그 등을 통해 평가 내용을 공유하는 리뷰 관련 시장이 활성화 되면서 그 가치가 더욱 높아지고 있음.

예제 분석

- 목표
  영화 리뷰를 긍정과 부정으로 분류하는 방법 학습

- 소스 저장소
  https://github.com/jongbumi/mlstudy_opinion_mining

- 전체 흐름
  1. 전처리
  2. 나이브 베이즈 분류기를 통한 훈련 및 테스트
  3. 바이그램을 통해 개선
  4. 최빈도 단어를 통해 개선
  5. Doc2Vec + 로지스틱 회기 분석 모델/SVM 분류기를 통한 훈련 및 테스트

예제 분석 - 전처리

- 토큰화
  단어별로 분리

- 불용어(stopwords) 제거
  nltk.corpus.stopwords.words()

- 어간 추출
  nltk.stem.porter.PorterStemmer().stem()

예제 분석 – 나이브 베이즈 분류기

- 훈련
  classifier = NaiveBayesClassifier.train(trainfeatures)

- 분류
  classifier.classify(word_features)

예제 분석 - 바이그램

- 각 문서별 특징 추출시
  최고의 바이그램을 계산해 결과를 개선

- 핵심 코드

```
measure = BigramAssocMeasures.chi_sq
bigram_finder = BigramCollocationFinder.from_words(words)
bigrams = bigram_finder.nbest(measure, nbigrams)
```

예제 분석 - 최빈도 단어

- 전체 흐름
  1. 긍정 리뷰, 부정 리뷰 각각에서의
     전체 단어에 대한 빈도수 계산
  2. 각 단어별 점수 계산
  3. 점수가 높은 10000개의 단어만 추출
  4. 각 문서별 특징 추출시 위 10000개의 단어만 추출

- 핵심 코드

```
word_scores = {}
for word, freq in iter(word_fd.items()):
    pos_score = BigramAssocMeasures.chi_sq(label_word_fd['pos'][word],
                                           (freq, pos_words), tot_words)
    neg_score = BigramAssocMeasures.chi_sq(label_word_fd['neg'][word],
                                           (freq, neg_words), tot_words)
    word_scores[word] = pos_score + neg_score
```

예제 분석 - Doc2Vec

- 전체 흐름
  1. Doc2Vec을 통해 문서의 벡터화
  2. 벡터화된 데이터를 훈련 집합과 테스트 집합으로 분리
  3. 로지스틱 회귀 분석 모델에 훈련 데이터 할당 및 테스트
  4. SVM 분류기에 훈련 데이터 할당 및 테스트

한국어 오피니언 마이닝 예제

- 발표 자료
  https://www.lucypark.kr/slides/2015-pyconkr/#36

- 소개
  현재 네이버에서 파파고를 개발하고 있다는 박은정님이
  2015년 PyCon에서 발표한 예제

- 특징
  한국어 처리를 위해 KoNLPy라는 한국어 형태소 분석기가
  사용됨

- 소스 저장소
  https://github.com/jongbumi/mlstudy_opinion_mining_korean

용어 정리

용어	의미	관련 용어
바이그램 (bigram)	바이(bi-)'는 '둘'이라는 뜻이다. 한 단어가 나타날 확률이 앞 단어에 영향을 받는다고 가정하는 것.	유니그램 트라이그램
연어 (collocation)	특정한 뜻을 나타낼 때 함께 쓰이는 단어의 결합을 의미.	
영가설 (null hypothesis)	귀무가설'과 같은 말로 통계학에서 처음부터 버릴 것을 예상하는 가설. 차이가 없거나 의미있는 차이가 없는 경우의 가설.	