데이터와 분석방법

분석 과정

1. 데이터 모음
2. 텍스트 전처리
3. Feature 선택
4. 극성 또는 감성 사전 분류
5. 문장과 문서 분석하여 감성 measure
6. 데이터

* 총 231,699 문서, 2005-Dec~2017
* MPB Minutes, News articles, bond분석보고서

1. 영어는 tokenization과 normalization을 사용하지만 한국어는 eKoNLPy를 사용했다.

한국어 분석의 **문제점** 4가지

1. Spacing. 전치사(~는, 께서, 으로, 부터, 에서, 야,와,이여 등)
2. 외래어가 원래 사용되던 의미와 다르게 사용되
3. 같은 단어를 다르게 표현(인플레이션, 인플레, 물가)
4. 불규칙적으로 쓰이는 동사와 형용사

KoNLPy는 spacing처리에는 사용가능 그러나 나머지는 처리 불가

그래서 eKoNLPy사용. 이건 경제금융을 전문으로 만들어진 사전.

eKoNLPy **이점**

1. 나만의 사전 추가 가능.
2. 이미 정리된 동의어
3. 불규칙 conjugation은 stemming과 lemmatization으로 해결 가능

**둘의 큰 차이는:**

**KoNLPy =** 금융, 통화, 위원회

**eKoNLPy =** 금융통화위원회

1. Feature = Words that express opinions

단어 혼자만 있으면 극성 분류가 쉽지 않다. 또한 긍정적 단어와 부정적 단어가 함께 쓰일 때 sentiment를 분석하기 어렵다. ‘증가’라는 단어가 ‘실업률 증가’(부정적), ‘월급 증가’(긍정적)로 뜻이 바뀌기 때문에 n-gram으로 해결

1. 극성분류

채택된 방법 = corpus-based method.

* Seed word와 함께 나타나는 단어들에게서 패턴을 찾음. 전문용어가 많이 나오는 경제/금융 분야에서 활용이 좋음
* Market approach 와 laxical approach

1. Market approach는 한달동안 Call rates의 변화가 양인지 음인지로 새로운 뉴스 기사를 매파 또는 비둘기파로 구분함. 머신러닝
2. Laxical approach. 긍정/부정 seed word를 사전으로 정의하고 이 단어들과 함께 자주(15번 이상)나오는 단어들에게도 같은 polarity를 적용