Lecture 10: Sentiment Analysis

Pilsung Kang
School of Industrial Management Engineering
Korea University

AGENDA

01 Overview

02 Architecture

03 Lexicon-based Approach

04 Machine Learning-based Approach

Lexicon-based Approach

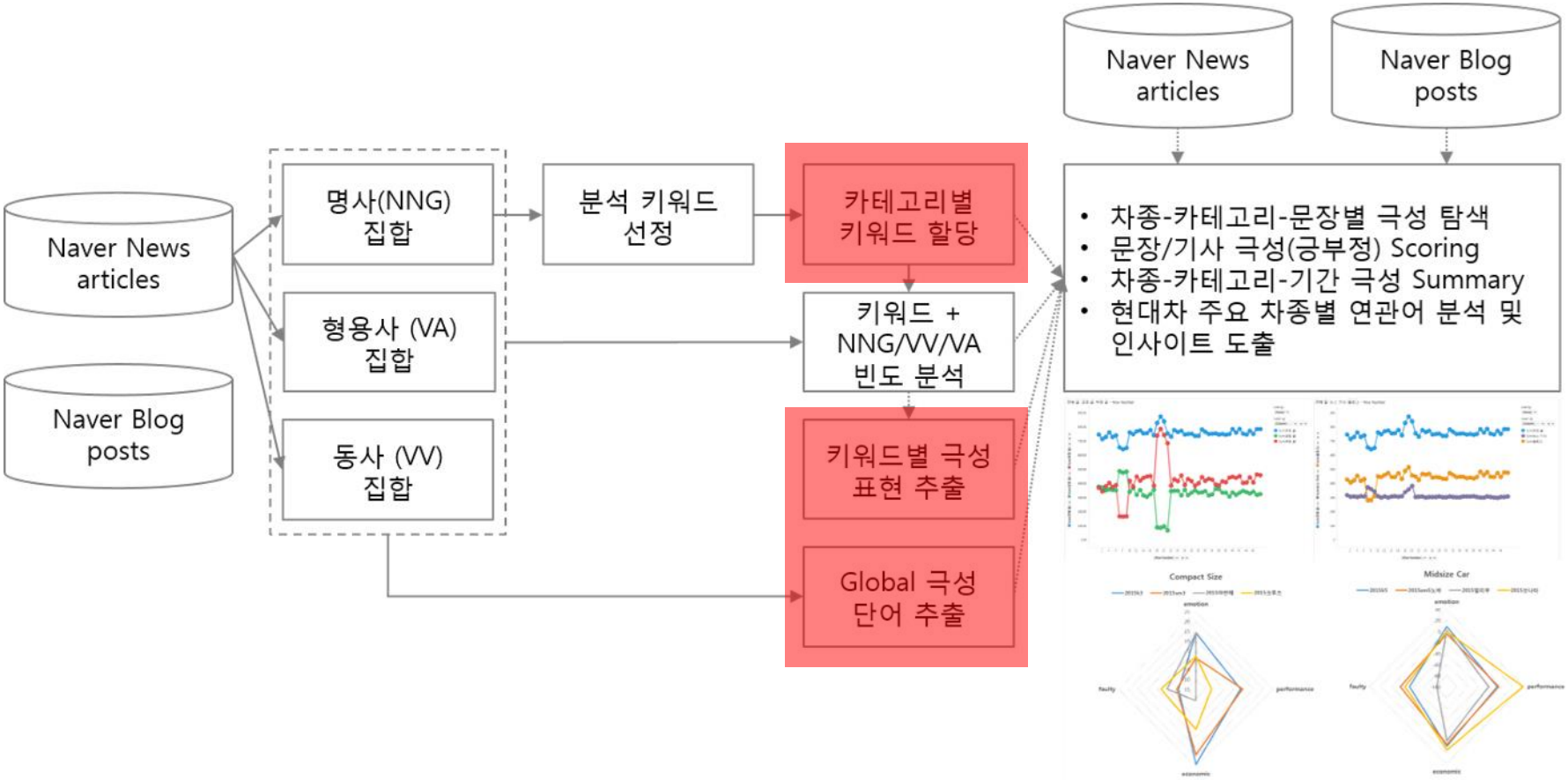
- 감성 사전 활용
  - ✓ 감성 사전은 극성, 주관성 등 감성을 지닌 감성 어휘들의 모음
  - ✓ 대부분 영어에 대한 감성 사전이 구축 (ex. SentiWordNet)
  - ✓ 한국어의 경우 준지도학습을 통해 감성 사전을 구축하려는 시도가 있었음 (서덕성 외, 2017)
  - ✓ 문서 또는 문장을 구성하는 단어의 감성 점수를 사용하여 비지도학습 기반 감성 분석을 진행할 수 있음

긍정		부정	
단어	감성 점수	단어	감성 점수
슬펏	1	별루였어	-0.5458
괜찮	0.5956	드럽	-0.4140
신나요	0.4357	부족	-0.3713
눈시울	0.4089	심해	-0.3449
짱	0.4002	어설퍼	-0.3123
존잘	0.3851	빈약	-0.2531

감성사전 예시 (출처: 서덕성 외, 2017)

Manual Approach

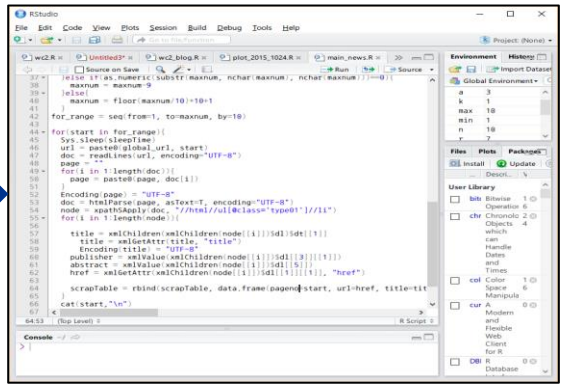
- Sentiment Analysis for Car Models



Manual Approach

- Data Collection
  - ✓ News articles and blog posts from NAVER (2010. 01. 01 ~ 2015. 07. 31)

원본 기사/블로그 게시글 검색



웹 크롤러를 이용한 게시글 수집

article_date	title	sentence_id	sentence
2010-01-04	2009 `베스트셀링카'는 쏘나타	2	(서울=연합뉴스) 권혁창 기자 = 지난해 국내에서 가장 많이 팔린 베스트셀링카는 현대차의 쏘나타로 나타났다
2010-01-04	2009 `베스트셀링카'는 쏘나타	3	4일 관련업계에 따르면 현대차 쏘나타는 지난 한해 구형 NF 8만4천981대, 신형 YF 6만1천345대 등 총 14만6천326대가 팔려 준중형 아반떼(11만5천378대)를 누르고 내수 판매 1위를 차지했다
2010-01-04	2009 `베스트셀링카'는 쏘나타	4	특히 신형 쏘나타는 지난해 9월 출시된 이후 매달 1만5천대 이상 팔려나가며 현대차의 내수판매 증대를 이끌었다
2010-01-04	2009 `베스트셀링카'는 쏘나타	5	기아차의 경차 모닝은 10만2천82대로 3위에 올랐으며, 현대차의 소형트럭 포터가 7만8천846대로 4위에 랭크됐다
2010-01-04	2009 `베스트셀링카'는 쏘나타	6	이어 현대차 그랜저(7만5천844대), 싼타페(5만8천324대), 르노삼성 SM5(6만1천10대)가 5~7위를 차지했다
2010-01-04	2009 `베스트셀링카'는 쏘나타	7	기아차 포르테(5만1천374대), 로체(4만9천54대), 르노삼성 SM3(4만5천906대), GM대우 라세티 프리미어(4만4천464대)가 그 뒤를 이었다
2010-01-04	2009 `베스트셀링카'는 쏘나타	8	faith@yna.co.kr
2010-01-04	2009 `베스트셀링카'는 쏘나타	9	[관련기사]
2010-01-04	2009 `베스트셀링카'는 쏘나타	10	<저작권자(c)연합뉴스
2010-01-04	2009 `베스트셀링카'는 쏘나타	11	무단전재-재배포금지.>

데이터베이스에 Raw Data를 저장

Manual Approach

- Data Collection: 520,000 news articles & 2,360,000 blog posts
  - ✓ At least 10,000 and at most 38,000 news articles are collected for each model
  - ✓ At least 89,000 and at most 248,000 blog posts are collected for each model

원본 문장	형태소 분석 결과
지난해 국내에서 가장 많이 팔린 베스트셀링카는 현대차의 쏘나타로 나타났다.	지난해(NNG) 국내(NNG) 팔리(VV) 베스트(NNG) 셀링카(NNG) 현대(NNG) 차(NNG) 쏘나타(NNG) 나타나(VV)
4일 관련업계에 따르면 현대차 쏘나타는 지난 한해 구형 NF 8만4천981대, 신형 YF 6만1천345대 등 총 14만6천326대가 팔려 준중형 아반떼(11만5천378대)를 누르고 내수 판매 1위를 차지했다.	관련(NNG) 업계(NNG) 따르(VV) 현대(NNG) 차(NNG) 쏘나타(NNG) 지나(VV) 한해(NNG) 구형(NNG) 신형(NNG) 팔리(VV) 중형(NNG) 아반떼(NNG) 누르(VV) 내수(NNG) 판매(NNG) 차지(NNG)
특히 신형 쏘나타는 지난해 9월 출시된 이후 매달 1만5천대 이상 팔려나가며 현대차의 내수판매 증대를 이끌었다.	신(NNG) 형(NNG) 쏘나타(NNG) 지난해(NNG) 출시(NNG) 이후(NNG) 매달(NNG) 이상(NNG) 팔(VV) 나가(VV) 현대(NNG) 차(NNG) 내수(NNG) 판매(NNG) 증대(NNG) 이끌(VV)
기아차의 경차 모닝은 10만2천82대로 3위에 올랐으며, 현대차의 소형 트럭 포터가 7만8천846대로 4위에 랭크됐다.	기아(NNG) 차(NNG) 경차(NNG) 모닝(NNG) 오르(VV) 현대(NNG) 차(NNG) 소형(NNG) 트럭(NNG) 포터(NNG) 랭크(NNG)
이어 현대차 그랜저(7만5천844대), 싼타페(5만8천324대), 르노삼성 SM5(6만1천10대)가 5~7위를 차지했다.	잇(VV) 현대(NNG) 차(NNG) 그랜저(NNG) 싸(VA) 타(NNG) 르노(NNG) 삼성(NNG) SM5(NNG) 가(VV) 차지(NNG)
기아차 포르테(5만1천374대), 로체(4만9천54대), 르노삼성 SM3(4만5천906대), GM대우 라세티 프리미어(4만4천464대)가 그 뒤를 이었다.	기아(NNG) 차(NNG) 포르테(NNG) 로체(NNG) 르노(NNG) 삼성(NNG) SM3(NNG) 대우(NNG) 라세티(NNG) 프리미어(NNG) 가(VV) 뒤(NNG) 잇(VV)

Manual Approach

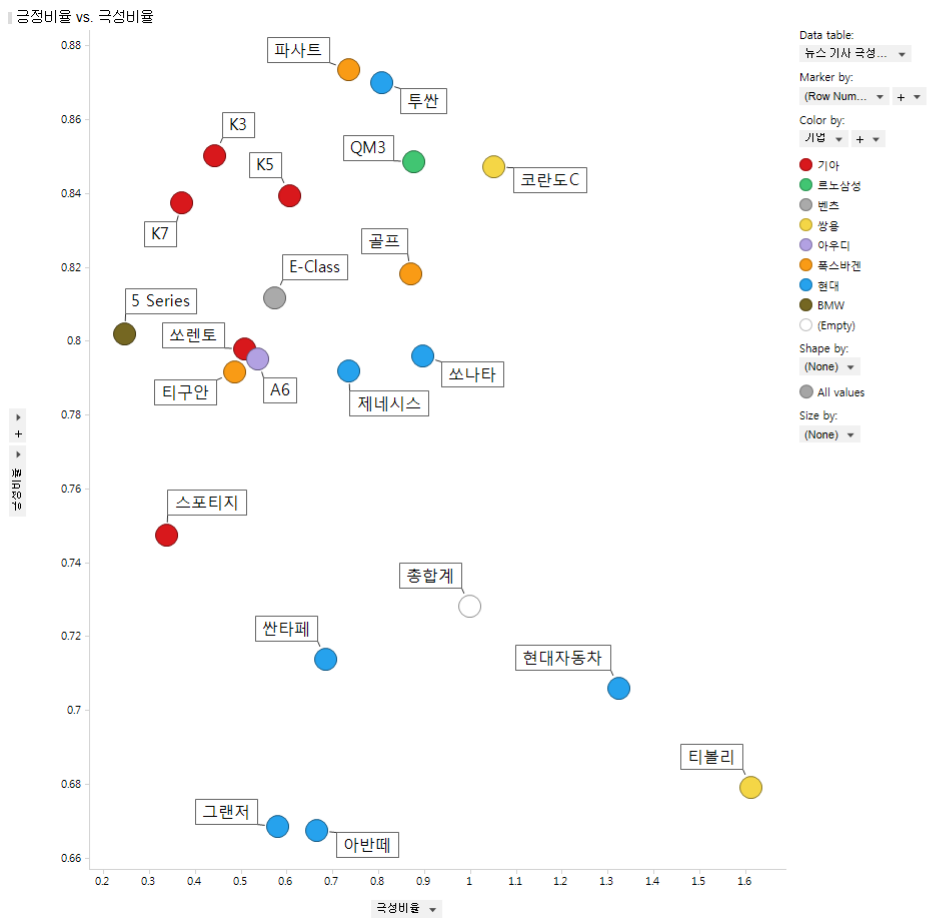
- Keyword selection and sentiment dictionary construction
  - ✓ Purely done by X people within 7 days

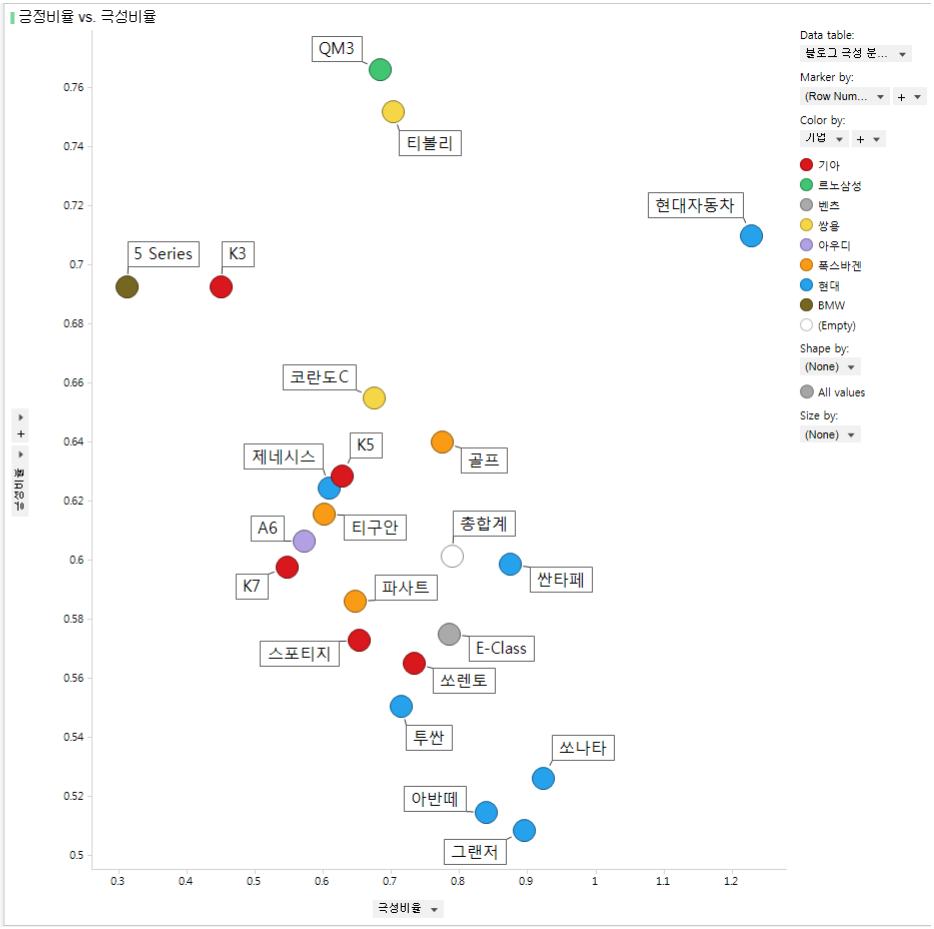
카테고리	키워드	키워드+표현	극성
Performance	기술	기술 갖(VV)	Positive
Economic	시장	시장 티볼리(NNG)	Negative
Reputation	노사	노사 연속 파업(NNG)	Negative
Performance	엔진	엔진 선택 있(VV)	Positive
Economic	매출	매출 하락(NNG)	Negative
Reputation	직원	직원 노동 조합(NNG)	Negative
Performance	하이브리드	하이브리드 고객(NNG)	Positive
Faulty	경보	경보 장치 급제동(NNG)	Positive
Performance	마력	마력 엔진 장착(NNG)	Positive
Economic	경영	경영 환경 악화(NNG)	Negative
Emotional	프리미엄	프리미엄 이미지(NNG)	Positive
Emotional	사양	사양 정숙(NNG)	Positive
Economic	시장	시장 다변화(NNG)	Positive
Emotional	출시	출시 스포츠(NNG)	Positive
Economic	시장	시장 판매 증가율(NNG)	Positive
Economic	환경	환경 차량 선보이(VV)	Positive
Emotional	콘셉트	콘셉트 세계 최초(NNG)	Positive
Reputation	노동자	노동자 정규직 전환 촉구(NNG)	Negative
...	...	...	...

"문제는 한국 사회에서 시스템이 필요하다고 지시를 내릴 사람은 많은데 전통적으로 "노가다"를 뛸 사람은 없다는 겁니다. 이런 일은 남이 해야 하는 거라 생각하죠. 아니면 남이 했다가 자기한테 해가 되면 안 되니까 오만 가지 이유를 대서 이런 일은 하면 안 된다고 하고, 이런 일이 의료계에서만 있는 줄 알았어요. 사회 전반이 바뀌지 않으면 이 문제는 나아지지 않아요"라고 했다.

Manual Approach

- Sentiment Analysis Results

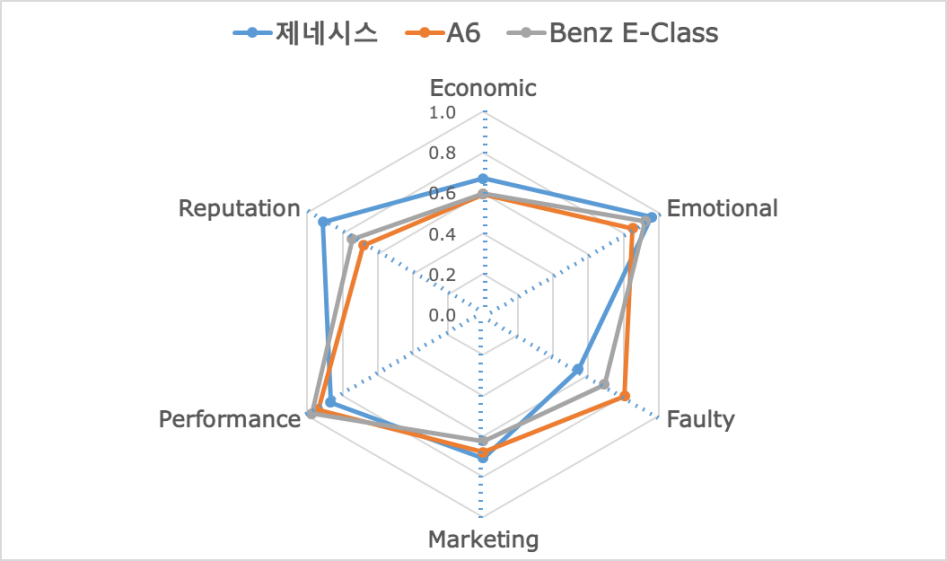




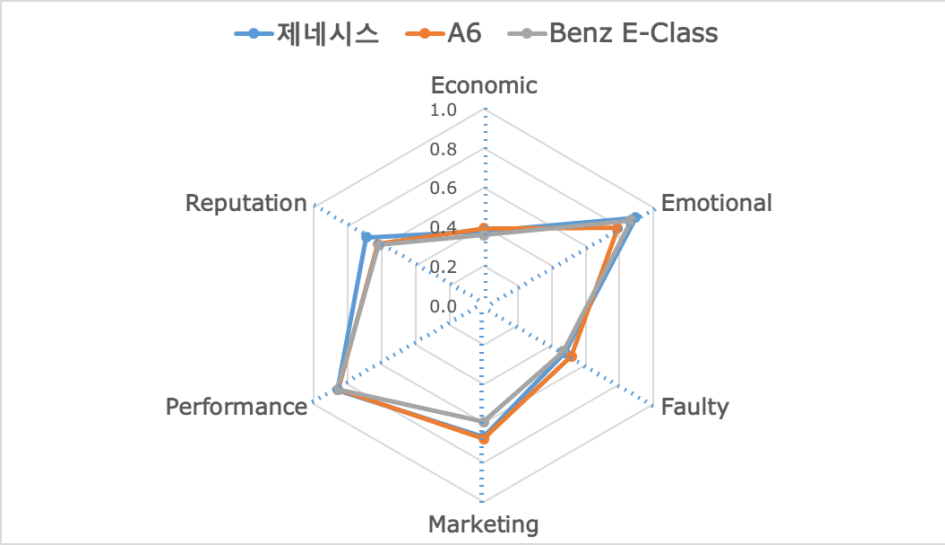
Manual Approach

- Sentiment Analysis Results

뉴스



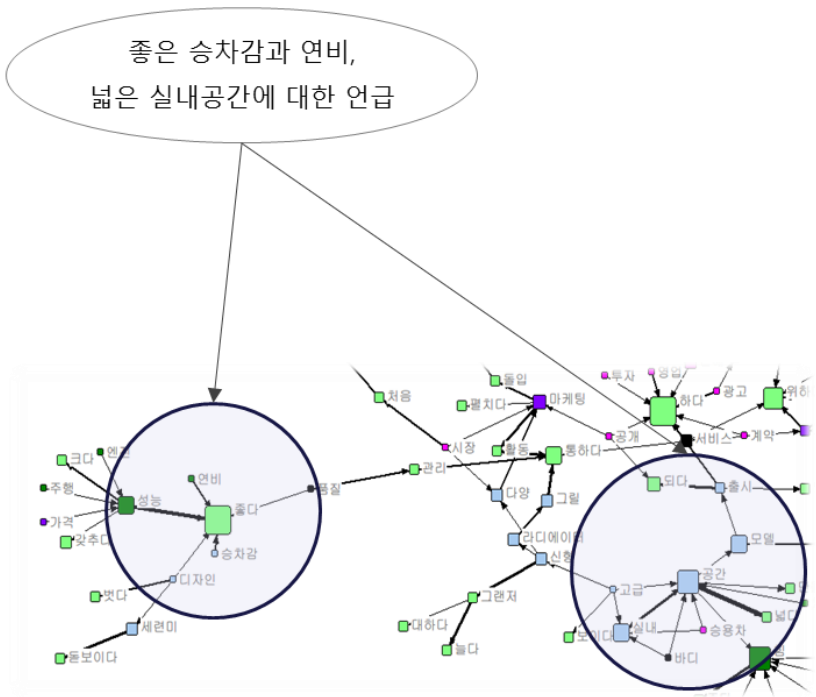
블로그



Manual Approach

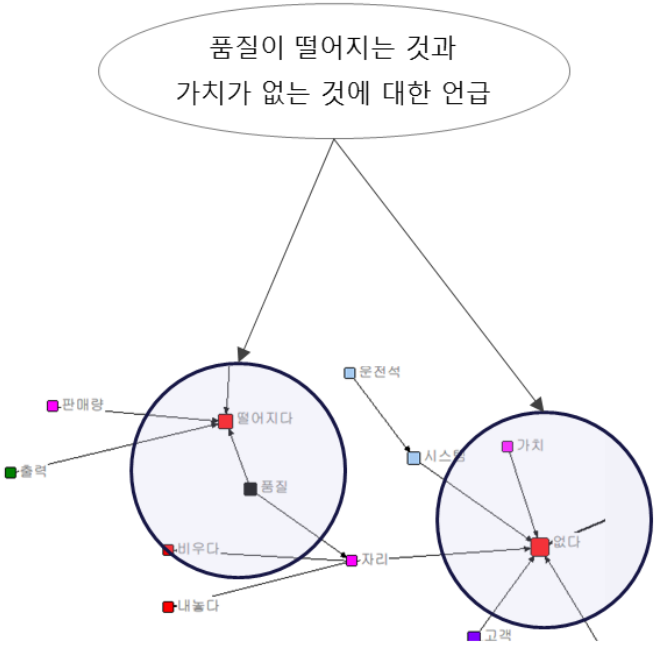
- Sentiment Analysis Results

긍정 연관어



긍정 연관어 네트워크 분석 결과 일부(뉴스)

부정 연관어



부정 연관어 네트워크 분석 결과 일부(뉴스)

Dictionary- vs Corpus-based Approaches

Feldman (2013)

- Dictionary-based approaches
  - ✓ Typically use WordNet's synsets and hierarchies to acquire opinion words
    - Start with a small seed set of opinion words
  - ✓ Usually do not give domain or context dependent meanings
- Corpus-based approaches
  - ✓ Often use a double propagation between opinion words and the items they modify
    - Knowing an aspect can find the opinion word that modifies it
      - The **rooms** are ***spacious***
    - Knowing some opinion words can find more opinion words
      - The **rooms** are **spacious** and ***beautiful***
  - ✓ Can find domain dependent orientations
  - ✓ Require a large corpus to get good coverage

Dictionary-based Approaches

- Popular Sentiment Lexicon Database for English

Multi-Perspective Question Answering

Main (MPQA Home) | Corpora (News, debates, etc.) | Lexicons (Subj. clues, etc.) | Annotation (GATE, MPQA, gfbf) | OpinionFinder (Subjectivity detector)

- MPQA Opinion Corpus
- Subjectivity Lexicon
- Subj. Sense Annotations
- Arguing Lexicon
- +/-Effect Lexicon
- Product Debate Data
- Political Debate Data
- goodFor/badFor Data
- OpinionFinder System

PLEASE NOTE: our URL has recently changed from www.cs.pitt.edu/mpqa/ to mpqa.cs.pitt.edu. Please update your bookmarks accordingly.

MPQA Opinion Corpus

The MPQA Opinion Corpus contains news articles from a wide variety of news sources manually annotated for opinions and other private states (i.e., beliefs, emotions, sentiments, speculations, etc.). To download the MPQA Opinion Corpus click here.

For **sample documents and instructions for MPQA annotation in GATE**, click here. *Updated July 2011.*

To learn more about the subjectivity and sentiment research that produced MPQA, please refer to the following publications:

Janyce Wiebe, Theresa Wilson , and Claire Cardie (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation, volume 39, issue 2-3, pp. 165-210.*

Theresa Wilson (2008). Fine-Grained Subjectivity Analysis. *PhD Dissertation, Intelligent Systems Program, University of Pittsburgh.*

Lingjia Deng and Janyce Wiebe (2015). MPQA 3.0: An Entity/Event-Level Sentiment Corpus.*NAACL-HLT, 2015.*

Subjectivity Lexicon

Made available under the terms of GNU General Public License. They are distributed without any warranty.

The Subjectivity Lexicon (list of subjectivity clues) that is part of OpinionFinder is also available for separate download. These clues were compiled from several sources (see the enclosed README). This is the version of the lexicon used in:

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proc. of HLT-EMNLP-2005.*

http://mpqa.cs.pitt.edu/

SentiWordNet

SentiWordNet is a lexical resource for opinion mining. SentiWordNet assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity. SentiWordNet is described in details in the papers:

SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining
SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining

How to obtain SentiWordNet

The current "official" version of SentiWordNet is 3.0, which is based on WordNet 3.0.

SentiWordNet is distributed under the Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0) license. Among the other possibilities, this license allows the use of SentiWordNet in commercial applications, provided that the application mentions the use of SentiWordNet and SentiWordNet is attributed to its authors.

Click here to download SentiWordnet 3.0

micro-WordNet-Opinion 3.0

micro-WordNet-Opinion 3.0 is the automatic mapping of the micro-WordNet-Opinion corpus to WordNet 3.0.

SentiWordNet has been used in...

Check Google for a list of the papers that use SentiWordNet 3.0

Check Google for a list of the papers that use SentiWordNet 1.0

http://sentiwordnet.isti.cnr.it/

Dictionary-based Approaches

- Popular Sentiment Lexicon Database for English

Opinion Lexicon (or Sentiment Lexicon)

- **Opinion Lexicon**: A list of English positive and negative opinion words or sentiment words (around 6800 words). This list was compiled over many years starting from our first paper (Hu and Liu, KDD-2004).
- **Comparative words**: A list of non-standard English compariative words and phrases for sentiment analysis. This list was compiled over many years starting from our first paper (Jindal and Liu, SIGIR-2006)
- Although necessary, having an opinion lexicon is far from sufficient for accurate sentiment analysis. See this paper: *Sentiment Analysis and Subjectivity* or the *Sentiment Analysis* book.
- Try Search for the Best Restaurant based on specific aspects, e.g., "best burger," "friendliest service." The system is a demo, which uses the lexicon (also phrases) and grammatical analysis for opinion mining.

Data Sets

- **Annotated: Customer Review Datasets (5 products)** associated with the paper (Hu and Liu, KDD-2004).
- **Annotated**: Additional Customer Review Datasets (9 products) some used in (Ding, Liu and Yu, WSDM-2008), which improves the lexicon-based method proposed in (Hu and Liu, KDD-2004)
- **Annotated**: More Customer Review Datasets (3 products) used in (Liu et al., IJCAI-2015)
- Amazon Product Review Data (more than 5.8 million reviews) used in (Jindal and Liu, WWW-2007, WSDM-2008; Lim et al, CIKM-2010; Jindal, Liu and Lim, CIKM-2010; Mukherjee et al. WWW-2011; Mukherjee, Liu and Glance, WWW-2012) for opinion spam (fake review) detection. You can also use it for sentiment analysis. It has information about reviewers, review texts, ratings, product info, etc. Due to the large file size, you may need to use *Download Accelerator Plus* (DAP) to download. If you use this data, please cite (Jindal and Liu, WSDM-2008).
- **Pros and cons dataset** used in (Ganapathibhotla and Liu, Coling-2008) for determining context (aspect) dependent sentiment words, which are then applied to sentiment analysis of comparative sentiences (comparative sentence dataset). The same form of Pros and Cons data was also used in (Liu, Hu and Cheng, WWW-2005).
- **Comparative sentence dataset** used in (Jindal and Liu, SIGIR-06) and (Jindal and Liu, AAAI-2006). **Comparative sentence dataset** used in (Ganapathibhotla and Liu, Coling-2008).
- **Blog author gender classification data set** associated with the paper (Mukherjee and Liu, EMNLP-2010)
- **Debate data set** used in (Mukherjee and Liu, ACL-2013; Mukherjee et al. ACL-2013).
- **Yelp Filtered Reviews** for Opinion spam or fake detection associated with the paper (Mukherjee et al.ICWSM-2013).

https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

SenticNet

Helping machines to learn, leverage, love.

about downloads sentic computing publications demo projects sentic api team

Talking about SenticNet is talking about concept-level sentiment analysis, that is, performing tasks such as polarity detection and emotion recognition by leveraging on semantics and linguistics in stead of solely relying on word co-occurrence frequencies.

In this context, SenticNet can be one of the following things:

1) a concept-level knowledge base;

2) a multi-disciplinary framework;

3) a private company.

As a ***knowledge base***, SenticNet provides a set of semantics, sentics, and polarity associated with 50,000 natural language concepts. In particular, semantics are concepts that are most semantically-related to the input concept (i.e., the five concepts that share more semantic features with the input concept), sentics are emotion categorization values expressed in terms of four affective dimensions (Pleasantness, Attention, Sensitivity, and Aptitude) and polarity is floating number between -1 and +1 (where -1 is extreme negativity and +1 is extreme positivity). The knoweldge base is downloadable for free as a standalone XML file and its latest version (released every two years) is also accessible as an API.

As a ***framework***, SenticNet consists of a set of tools and techniques for sentiment analysis combining commonsense reasoning, psychology, linguistics, and machine learning. In this context, SenticNet is more commonly referred to as *sentic computing*, a multi-disciplinary paradigm that goes beyond mere statistical approaches to sentiment analysis by focusing on a semantic-preserving representation of natural language concepts and on sentence structure.

As a ***company***, finally, SenticNet puts together the latest findings in concept-level sentiment analysis to offer easy-to-use state-of-the-art tools for big social data analysis that enable the automation of tasks such as brand positioning, trend discovery, and social media marketing in different modalities.

http://sentic.net/

Dictionary-based Approaches

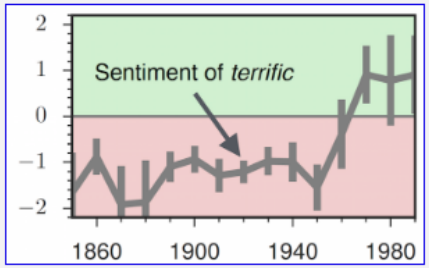
- SocialSent (Hamilton et al., 2016)

SocialSent: Domain-Specific Sentiment Lexicons for Computational Social Science

William L. Hamilton, Kevin Clark, Jure Leskovec, Dan Jurafsky

Introduction

The word *soft* may evoke positive connotations of warmth and cuddliness in many contexts, but calling a hockey player *soft* would be an insult. If you were to say something was *terrific* in the 1800s, this would probably imply that it was terrifying and awe-inspiring; today, *terrific* basically just implies that something is (pretty) good.



A word's sentiment or connotation depends on the domain or context in which it is used. However, previous computational work in natural language processing largely ignores this issue, and focuses and building and deploying generic domain-general sentiment lexicons.

SocialSent is a collection of code and datasets for performing *domain-specific* sentiment analysis. The SocialSent code package contains the SentProp algorithm for inducing domain-specific sentiment lexicons from unlabeled text, as well as a number of baseline algorithms.

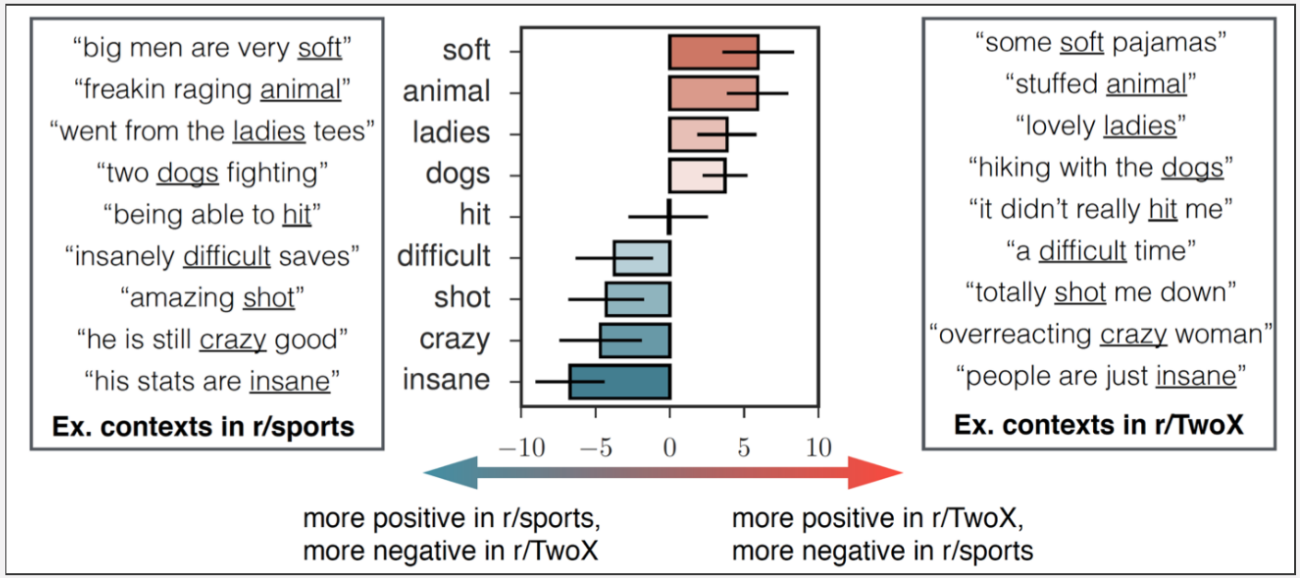
We have also released domain-specific historical sentiment lexicons for 150 years of English and community-specific sentiment lexicons for 250 "subreddit" communities from reddit.com. The historical lexicons reveal that more than 5% of sentiment-bearing words switched their polarity from 1850 to 2000, and the community-specific lexicons highlight how sentiment varies drastically between online communities.

The paper Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora details the SentProp algorithm and describes the lexicons we induced.

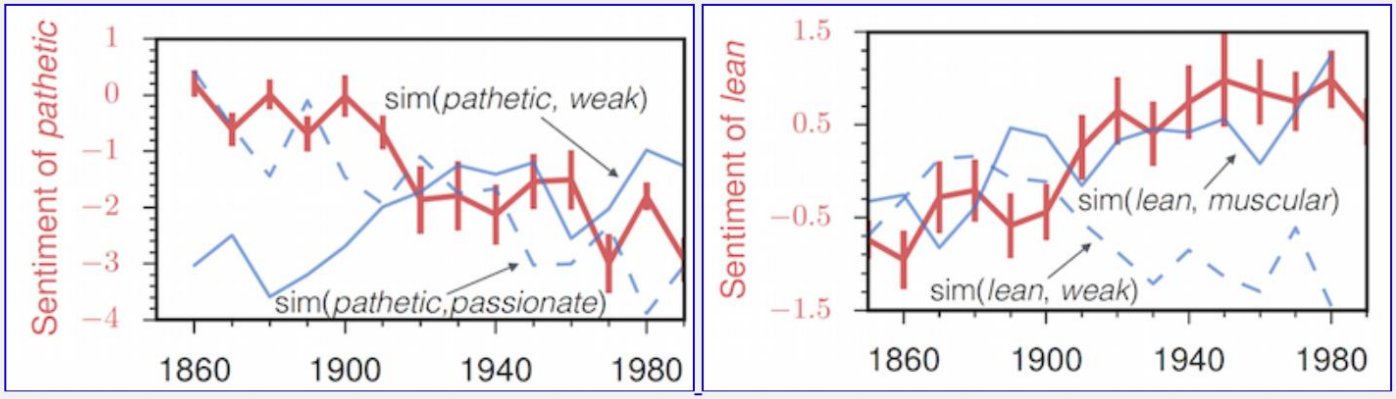
Dictionary-based Approaches

- SocialSent Highlights

1. **Word sentiment varies drastically between online communities**

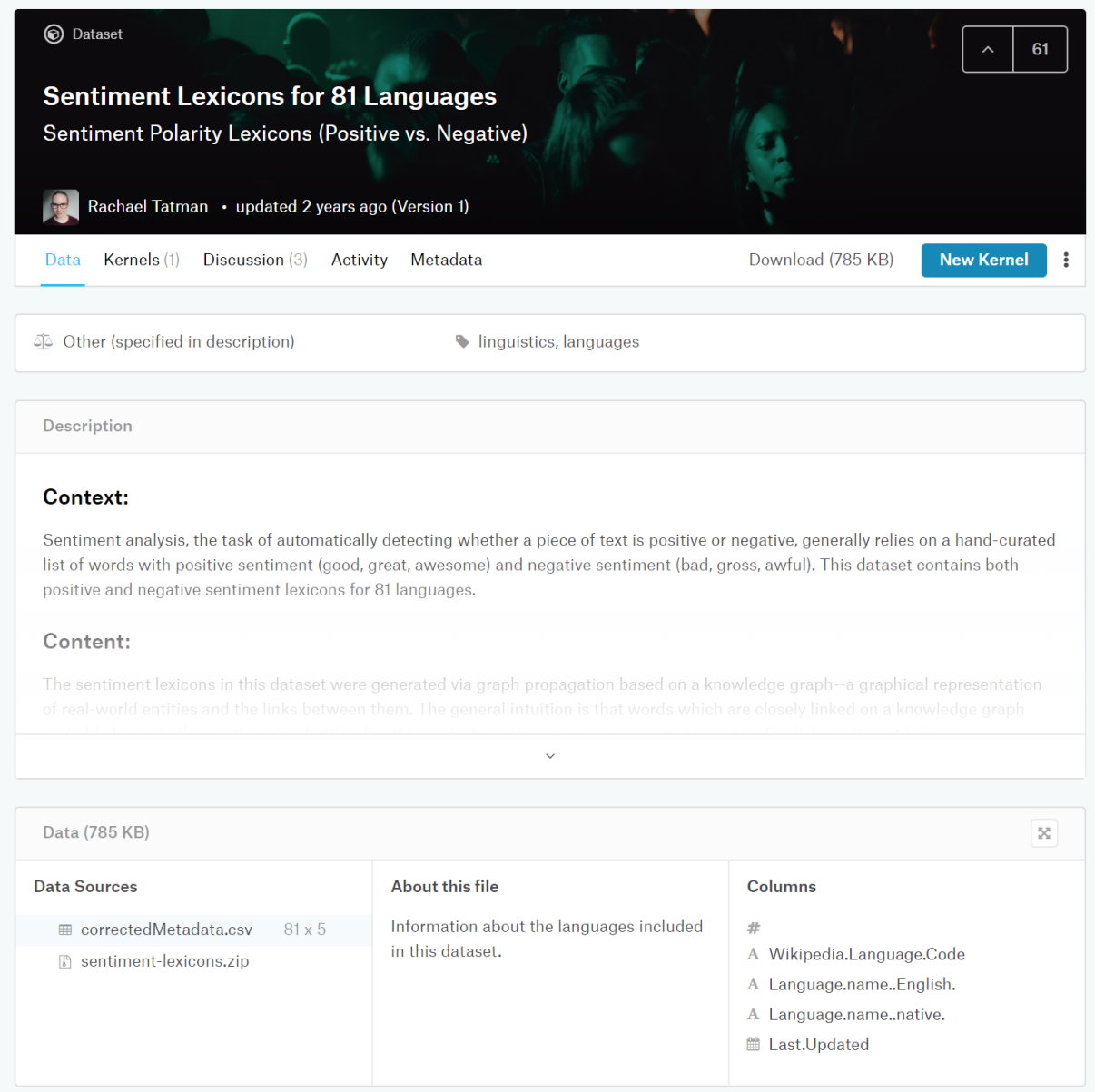


2. **Word sentiment varies dramatically over historical time-periods**



Dictionary-based Approaches

- Kaggle??



Dictionary-based Approaches

- Sentiment Lexicons by Kaggle
  - ✓ 883 positive lexicons, 1,235 negative lexicons for Korean

Positive lexicon examples

같은	정치	상	원하는	위로
더	자기	혁명	적절한	진정한
큰	고려	미	주요한	시인
주	쉽게	충분히	풍부한	일찍이
같이	최고의	아름다운	우수한	신성
매우	강력한	구조	전형적인	작업
잘	뛰어난	오른쪽	자유롭게	후한
볼	사람	상당한	특징	성공한
다양한	유명하다	지원	이해	고급
중요한	아주	단순한	장관	사랑
金	대	역사적	충분한	성공적으로
프로	개인	새롭게	참여	좋다
관련	자유	보호	비밀	유지
좋은	상당히	슈퍼	단일	수정
하다	초	지도	오른	보물
금	빠른	지속적으로	훌륭한	개혁
곧	기술	미리	종류	평화
시의	젊은	완전한	계약	힘든
유명한	right	정확히	막대한	변호사
비슷한	넓은	즉시	완전	사전
보다	인물	밝은	좋아하는	뜨거운
전체	특별한	전용	활발한	순수한
최고	德	크다	통일	굉장히
네	거대한	기능	꽤	지배
현대	정확한	방법	별의	강화
단	깊은	빨리	연	명예
사용할	양	진행	분명히	가능
강한	점	강하게	특유의	챔피언
선	원	간단한	승리	자치
작품	수상	해방	쉬운	성과

Negative lexicon examples

의	섬의	소수	지정	수상한
한	갑자기	시험	떨어지는	제외
다른	사망	어두운	문제의	힘
토론	독특한	전차	엄격한	포
특히	죽었다	결정	부족	지는
뒤	공격	죽을	분리	항구
없는	잘못된	비판	끊임없이	개인적으로
해	정의	상대	주장	폐지
없다	전쟁의	잃은	발견	테러
성	복잡한	무거운	분석	수사
크게	피해	향	타격	실패한
주의	부인	부족한	죽음	항의
만든	대상	나쁜	금지	고정
기타	적	길	암살	치열한
지방	잘못	기	김	분할
쓴	붙어	가난한	평범한	충돌
차단	아래의	화가	위험한	계승
긴	상태	색	줄거리	심하게
전혀	열	보지	소리	고문
삭제	강제	친일파	불법	위반
죽은	대신에	가을	분쟁	반란
문제	좁은	가상	연기	바이러스
사용하여	적의	노예	부하	정지
소설	뒤로	어렵다	붕괴	가사
신	발생	다리를	뚜렷한	가짜
본래	거친	경쟁	불가능한	부상
제작	약한	걸려	격렬한	살인
만화	칼	연결	범죄	망명
어려운	발표	회의	말기	짐
반대	지나치게	제한	잃었다	읽을

Dictionary-based Approaches

Feldman (2013)

- Dictionary-based Approach: Example
  - ✓ http://ws4jdemo.appspot.com/

WS4J Demo

WS4J (WordNet Similarity for Java) measures semantic similarity/relatedness between words.

Type in texts below, or use: example words | example sentences

1.	Input mode	(•) Word ( ) Sentence
2.	Word 1	love
3.	Word 2	like#
4.	Submit	Calculate Semantic Similarity

Summary

wup(love#v#2 , like#v#2) = 0.8000

jcn(love#v#2 , like#v#2) = 0.7566

lch(love#v#2 , like#v#2) = 2.6391

lin(love#v#2 , like#v#2) = 0.9086

res(love#v#2 , like#v#2) = 6.5662

path(love#v#2 , like#v#2) = 0.5000

lesk(love#v#2 , like#v#2) = 28

hso(love#v#2 , like#v#2) = 4

Each score above is the highest from 10 x 11 pairs of synsets.

WS4J Demo

WS4J (WordNet Similarity for Java) measures semantic similarity/relatedness between words.

Type in texts below, or use: example words | example sentences

1.	Input mode	(•) Word ( ) Sentence
2.	Word 1	fantastic
3.	Word 2	terrible
4.	Submit	Calculate Semantic Similarity

Summary

wup(fantastic , terrible) = -1 [Unsupported POS Pairs]

jcn(fantastic , terrible) = -1 [Unsupported POS Pairs]

lch(fantastic , terrible) = -1 [Unsupported POS Pairs]

lin(fantastic , terrible) = -1 [Unsupported POS Pairs]

res(fantastic , terrible) = -1 [Unsupported POS Pairs]

path(fantastic , terrible) = -1 [Unsupported POS Pairs]

lesk(fantastic#a#2 , terrible#a#4) = 153

hso(fantastic#a#2 , terrible#a#4) = 6

Each score above is the highest from 5 x 4 pairs of synsets.

Corpus-based Approaches

Hur et al. (2016)

- Building Up a Sentiment Dictionary based on Movie Review Data

관람 후(678) | 관람 전(33) 전체보기

★★★★½	9	내게 처음인 무성영화였다. 이렇게 5감을 쏟아가면서 집중하여 본건 처음인 것같다	widekk3	2012.03.19 신고
★★★½☆	7	한편의 좋은 옛날 영화를 보는 느낌!이지만 개인적으로 이런 영화는 좋아하지 않아서	bbollock	2012.03.19 신고
★★★★½	9	단순한 스토리지만 탄탄하게 잘 만든 듯. 다만 약간 길었다는 느낌..	un5166	2012.03.18 신고
★★★★½	9	고전적이고 뻔한 이야기? 이 시대에 무성 영화란 기획 자체가 뛰어난 발상이다.	im_loen	2012.03.18 신고
★★★★½	9	초심으로 돌아간 영화가 선사하는 마법	brego1114	2012.03.17 신고
★★★★☆	8	시나리오나연출은매우단순해도...극장에서흑백영화,무성영화를본 자체가좋은경험이었음	irlagywlsi	2012.03.17 신고
★★★★★	10	영화도예술이군요^^	cysuk2	2012.03.17 신고
★★★★☆	8	들리지 않아도 느낄 수 있다	blesspooh	2012.03.17 신고
★★★★☆	8	3D시대에 무성 흑백 4:3화면의 도전만하는 감독이 신선하다	haurifufang	2012.03.17 신고
★★★★★	10	영화의 역사를 관통하는 영리한 연출과 무성영화의 미덕을 새삼 느끼게 해준다	luki48	2012.03.17 신고

Corpus-based Approaches

Hur et al. (2016)

- Building Up a Sentiment Dictionary based on Movie Review Data
  - ✓ T-test for testing the difference of review ratings with and without a word

$$R(r_{i,j}, w_q) = \begin{cases} 0 & \text{if } \; w_q \notin r_{i,j} \\ r(r_{i,j}) & \text{if } \; w_q \in r_{i,j} \end{cases}$$

$$Score(w_q) = E(w_q) = \frac{1}{n(w_q)} \sum_{i=1}^{m} \left(\sum_{j=1}^{n_i} R(r_{i,j}, w_q) \right)$$

$$Var(w_q) = \frac{1}{n(w_q) - 1} \sum_{i=1}^{m} \left(\sum_{j=1}^{n_i} \left(R(r_{i,j}, w_q) - Score(w_q) \right)^2 \right)$$

Corpus-based Approaches

Hur et al. (2016)

- Building Up a Sentiment Dictionary based on Movie Review Data
  - ✓ T-test for testing the difference of review ratings with and without a word

$$\text{Testing}: T_w = \frac{E(W) - E(w)}{\sqrt{\frac{s_W^2}{n(W)} + \frac{s_w^2}{n(w)}}} \quad \begin{cases} \text{Positive if} & T_w > t_{(\alpha;v)} \\ \text{Negative if} & T_w < -t_{(\alpha;v)} \\ \text{Neutral} & Otherwise \end{cases}$$

$$\text{where} \quad v = \frac{\left(s_W^2/n(W) + s_w^2/n(w)\right)^2}{\frac{\left(s_W^2/n(W)\right)^2}{n(W)-1} + \frac{(s_w^2/n(w))^2}{n(w)-1}}$$

Corpus-based Approaches

Hur et al. (2016)

- Building Up a Sentiment Dictionary based on Movie Review Data
  - ✓ Example

Collocation	Avg	Stdev.	Count	NMovies	T	t-dist	Diag
멋지/vv	9.52	1.22	1,617	341	61.88	1.96	Positive
실망/NNG_시키/XSV_지/ECD_않/VXV	9.68	0.97	829	260	60.87	1.96	Positive
재밌/VA_게/ECD_보/VXV	9.13	1.70	4,679	565	59.90	1.96	Positive
재미/NNG_잇/VV	9.57	1.45	2,004	379	59.74	1.96	Positive
깔끔/XR_하/XSA	9.24	1.22	1,336	358	48.02	1.96	Positive
잘/MAG_만들/VV_ㄴ/ETD_영화/NNG	9.26	1.49	1,903	342	47.47	1.96	Positive
못/MAG_보시/VV_는/ETD_분/NNG	7.83	2.37	12	9	0.29	2.2	Neutral
남자/NNG_둘/NNG	7.78	3.18	36	34	0.26	2.03	Neutral
영웅/NNG_이야기/NNG	7.71	2.76	48	24	0.18	2.01	Neutral
복수극/NNG	7.68	2.75	104	38	0.17	1.98	Neutral
코믹/NNG_성/XSN	7.7	2.83	44	31	0.16	2.02	Neutral
도저히/MAG_모르/VV	4.27	3.28	15	15	-3.99	2.14	Negative
손발/NNG_오글거리/VV	4.41	3.8	22	18	-3.99	2.08	Negative
무리수/NNG	5.62	3.11	39	25	-4.06	2.02	Negative
안타깝/VA_ㄴ/ETD_영화/NNG	6.42	3.21	119	84	-4.13	1.98	Negative
좀/MAG_지겹/VA	6.62	2.2	89	72	-4.38	1.99	Negative
억지/NNG_이/VCP	6.3	2.96	121	88	-4.98	1.98	Negative


ANY
questions?