

Python을 이용한 텍스트 마이닝 기초

이윤경
서울대학교 심리학과
인간공학심리연구실

강사 소개



Yoon Kyung Lee

<https://yoonlee78.github.io/>

■ 소속

현) 서울대학교 심리학과 인지심리학 박사과정

■ 주요 강의 이력

서울대학교 공학교육센터 SPLIT 프로그램 Python Tutor

서울대 심리학과 대학(원)생 대상 「파이썬과 텍스트 분석 기초」

서울대 파이썬 부트캠프 「Psybus Self-Camp」 (2018)

Naver Startup Factory 「심리학과 파이썬: PsychoPy」 특강(2018)

참고 자료(튜토리얼) : www.github.com/yoonlee78

강의 개요

■ Introduction to Text Mining

■ Text Preprocessing

■ Korean Morphological analysis

■ Vector Space Model and Visualization

■ Topic Modeling

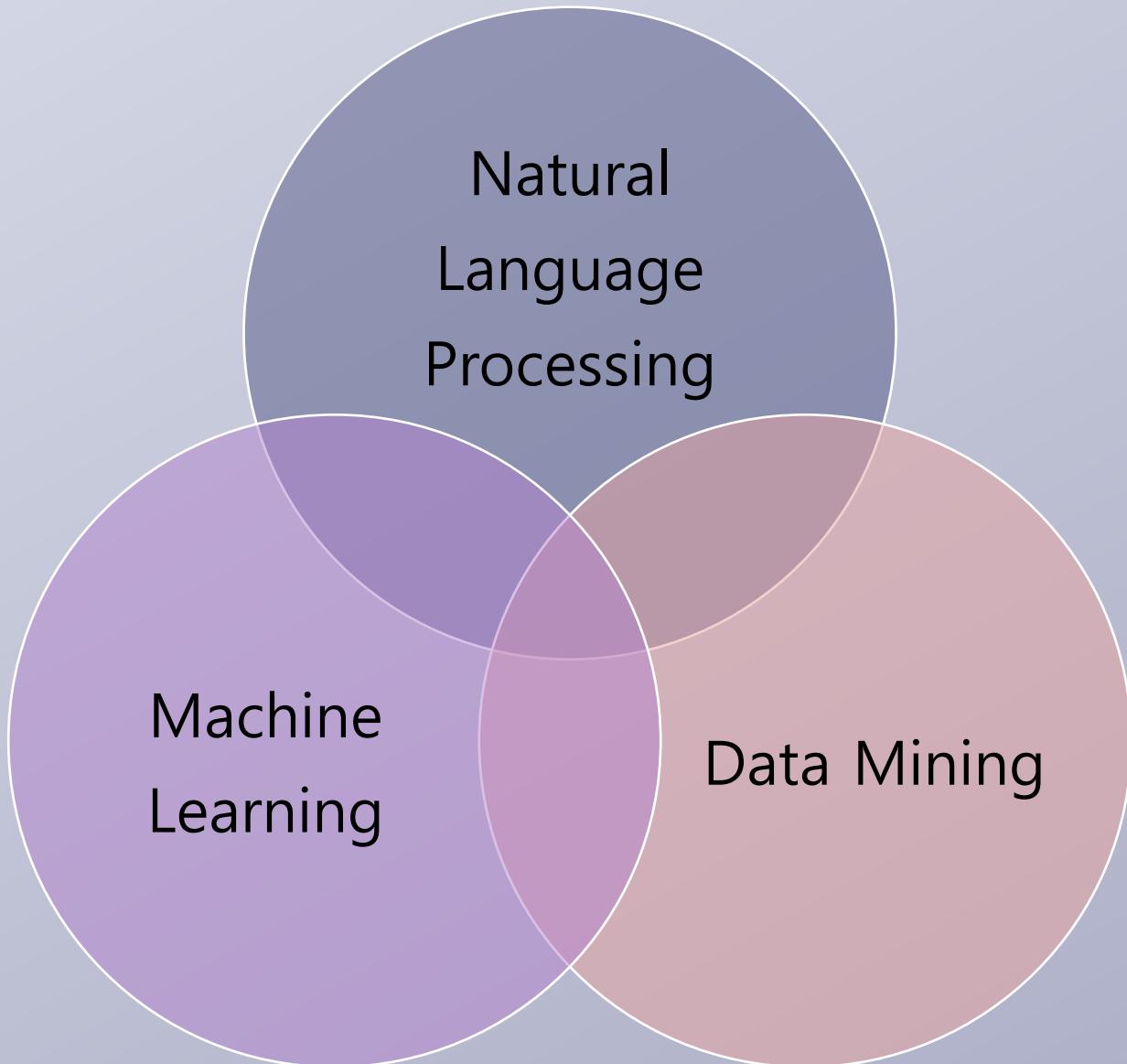
1 | Introduction to Text Mining

■ 텍스트 마이닝이란?

- 다량의 비정형 텍스트 데이터셋에서 “흥미로운” 규칙을 찾아내는 작업
- 대규모 데이터베이스에서 패턴을 찾는 ‘데이터 마이닝’의 변형
- 데이터 마이닝과 달리 ‘텍스트’로 된 데이터에서 새로운 정보를 발견함

Text mining is a variation on a field called data mining, that tries to find interesting patterns from large databases

Hearst, M. (2003), UC Berkeley



1 | Introduction to Text Mining

■ 텍스트 마이닝 기법의 적용 분야

- Search
- Sentiment Analysis
- Natural Language Understanding
- Recommendation

1

Introduction to Text Mining

▪ Search

The image displays four examples of search interfaces:

- NAVER search bar:** Shows a search bar with the query "파이썬" (Python). Below it, a dropdown menu lists search categories like 통합검색 (Unified Search), 이미지 (Image), 책 (Book), 블로그 (Blog), 카페 (Cafe), 지식백과 (Encyclopedia), 지식iN (Knowledge iN), 웹사이트 (Website), and 더보기 (More). A related search section shows terms like python, django, 파이썬 설치, 파이썬 기초, 디스코드 노래봇, 파이썬 프로그래밍, 파이썬 독학, django, 파이썬 머신러닝, 파이썬 학원, 프로그래밍, and 독학.
- Gmarket dropdown menu:** Shows a dropdown menu for the category "1인" (1 person) under the "Best" section. The menu lists items such as 1인용소파, 1인용컴퓨터책상, 1인용침대, 1인용소파, 1인용텐트, 1인소파, 1인용안락의자, 1인소파, 1인용밥솥, and 1인용침대.
- Google search suggestions:** Shows a list of search suggestions for the query "korea". The suggestions include korea times, korea university, korea herald, korea map, korea post, korea weather, korea inverse etf, korea flag, korea air, and korea gdp.
- News topic list:** Shows a list of news topics under the heading "뉴스토픽". The topics are numbered 1 through 10, with titles like 김홍일 전 의원 별세, 한국당 광화문집회, 대전 홍역 환자 2명 추가 발생, 공동방송 추진, 전북 무주 야산서 산불, 민주당 한국당, 박선영 남편 김일범, 황교안 자유한국당 대표, 김학의 수사, and 해경이 구조. Each topic has a "NEW" tag next to it.

1

Introduction to Text Mining

▪ Sentiment Analysis



★ 1/10

Not Worth Your Money Nor The Hype 2018

Warning: Spoilers

351 out of 851 found this helpful. Was this review helpful? [Sign in](#) to vote.

[Permalink](#)



★ 9/10

Excellent Film

I was amazed to see so many negative reviews; so many people are impossible to please. This movie was 2 1/2 hours long, but I could have sat there another 2 1/2 hours and not noticed. Thoroughly entertaining, and I love how the directors weren't afraid to take chances. I've read a lot of other user reviews that claim that there's no plot. Unless you're mentally handicapped or not paying attention because you're on your phone the entire movie, the plot is pretty clear, and decent in my opinion.

318 out of 587 found this helpful. Was this review helpful? [Sign in](#) to vote.

[Permalink](#)



1

Introduction to Text Mining

Sentiment Analysis



NAVER 영화

★★★★★ 1 중간 중간 지루하고 결말은 그지같고 마블 히어로들은 약해 빠졌고...

2019.02.05 08:34 | 신고

공감 1 비공감 44



★★★★★ 10 마블 영화 중 가장 재밌는건 물론이고, 무엇보다 타노스라는 캐릭터가 단순히 빌런에서 안 그치고 상당히 매력적으로 묘사된 듯!

2019.04.20 08:26 | 신고

공감 1 비공감 0



1

Introduction to Text Mining

▪ Sentiment Analysis

긍정 상품평 BEST

오* [REDACTED] ★★★★★ 2019.03.26

원더스리빙 공기청정기 퓨리킹 WA360 33m³



가격파괴급 공기청정기입니다
배송은 당연히 로켓배송이니까 만족해요
하루만에 받았어요

작은방(투룸)에 쓸 공기청정기 찾아보는데 공청기들이 비싸도 너무 비싸더라고요
그래서 좀 더 가성비 좋은 공청기를 찾았는데 과장해서 100번정도 고민하다가
결국 원더스리빙꺼 구매했어요

1. 가격이 착한가
2. H13헤파필터 유무(다른필터 고려X)…

[더보기 >](#)

39명에게 도움 됨 [도움이 돼요](#) [도움 안 돼요](#)

신고하기

비판 상품평 BEST

한* [REDACTED] ★★★★★ 2019.04.14

원더스리빙 공기청정기 퓨리킹 WA360 33m³

쿠팡배송 엄망이네요
새벽배송 공동현관 길거리에 버리고가셨네요
상식적으로 이런상품은 문이잠겨있으면 전화라도해야는거 아닌가요?길거리에 버리고갔으면서 사과한마디없이
다음부터 공동현관비밀번호를 알려달라니 어의가없네요
오래된빌라여서 그런거없는데 어찌나요
전화한통화해서 배송하는게 그렇게어렵나요??
사과도없고 지들맘대로 배송완료해버리고 뭐하는건지

5명에게 도움 됨 [도움이 돼요](#) [도움 안 돼요](#)

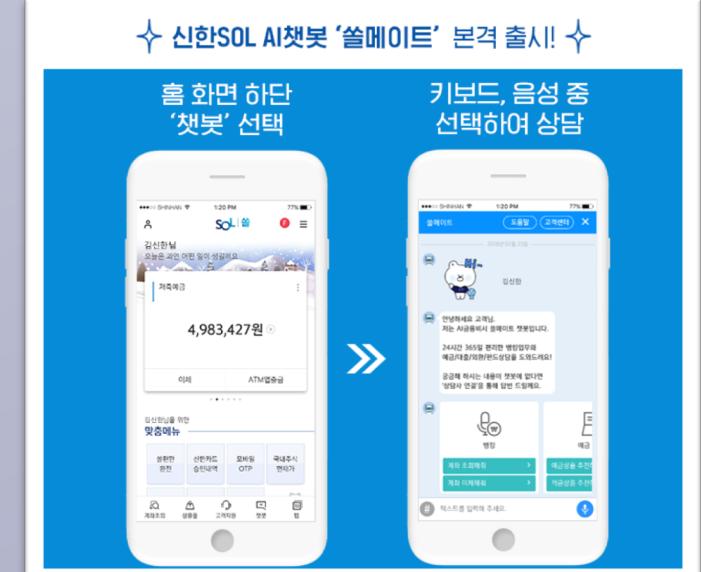
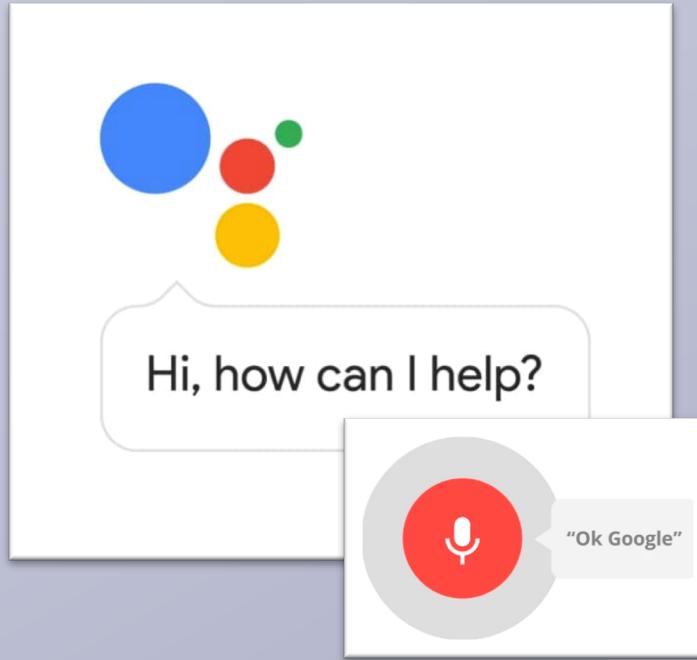
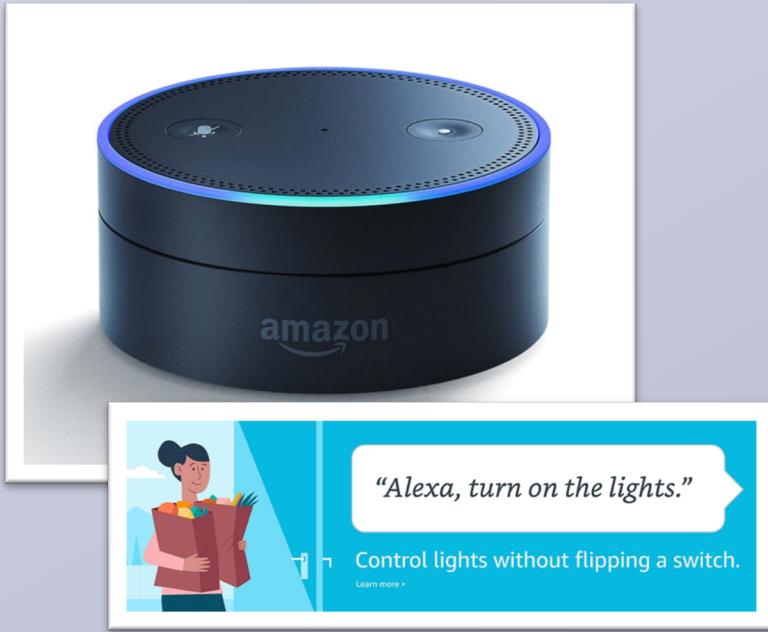
신고하기

coupang

Human Factors Psychology | Seoul National University

1 | Introduction to Text Mining

- Natural Language Understanding

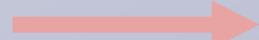


1

Introduction to Text Mining

▪ Recommendation

“청소기”
검색 결과



휴스톰의 다른 상품들

휴스톰 듀얼스핀 물걸레청소기, HS-10000
188,550 원 로켓배송
★★★★★ (23)

휴스톰 무선 듀얼스핀 물걸레 청소기 HS-9000
153,000 원 로켓배송
★★★★★ (2,690)

휴스톰 듀얼리튬 물걸레청소기 HS-12000W
338,000 원 로켓배송
★★★★★ (27)

휴스톰 유선 듀얼스핀 물걸레 청소기 HS-8000
95,200 원 로켓배송
★★★★★ (562)

HUSTORM
총 145 개
[브랜드샵 구경할까요?](#)

다른 고객이 함께 구매한 상품

휴스톰 물걸레 청소기 1회용 패드 40p + 청소포 ...
18,000 원 로켓배송
★★★★★ (9)

LG전자 슈퍼 싸이킹 3 싸이클론 진공청소기,...
241,020 원 로켓배송
★★★★★ (471)

유니맥스 강력분사 핸디 스팀 청소기 UVC-....
23,880 원 로켓배송
★★★★★ (21)

휴스톰 물걸레청소기용 극세사 찌든때 패드, 단...
6,500 원 로켓배송
★★★★★ (662)

아토케어 침구청소기 EP-880
92,650 원 로켓배송
★★★★★ (506)

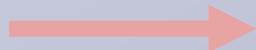
휴스톰 물걸레 청소기 드라이 패드 2p + 1회용 패드 ...
19,200 원 로켓배송
★★★★★ (2)

1

Introduction to Text Mining

▪ Recommendation

“Python Book”
검색 결과



Your recently viewed items and featured recommendations

Related to items you viewed

Aurélien Géron
[Hands-On Machine Learning with Scikit-Learn & TensorFlow](#)
by Aurélien Géron
Paperback \$29.49

Steven J. Luck
[An Introduction to the Event-Related...
POTENTIAL TECHNIQUE](#)
by Steven J. Luck
Paperback \$35.70

Steven Bird
[Natural Language Processing with Python](#)
by Steven Bird
Paperback \$34.06

2 | Preprocessing

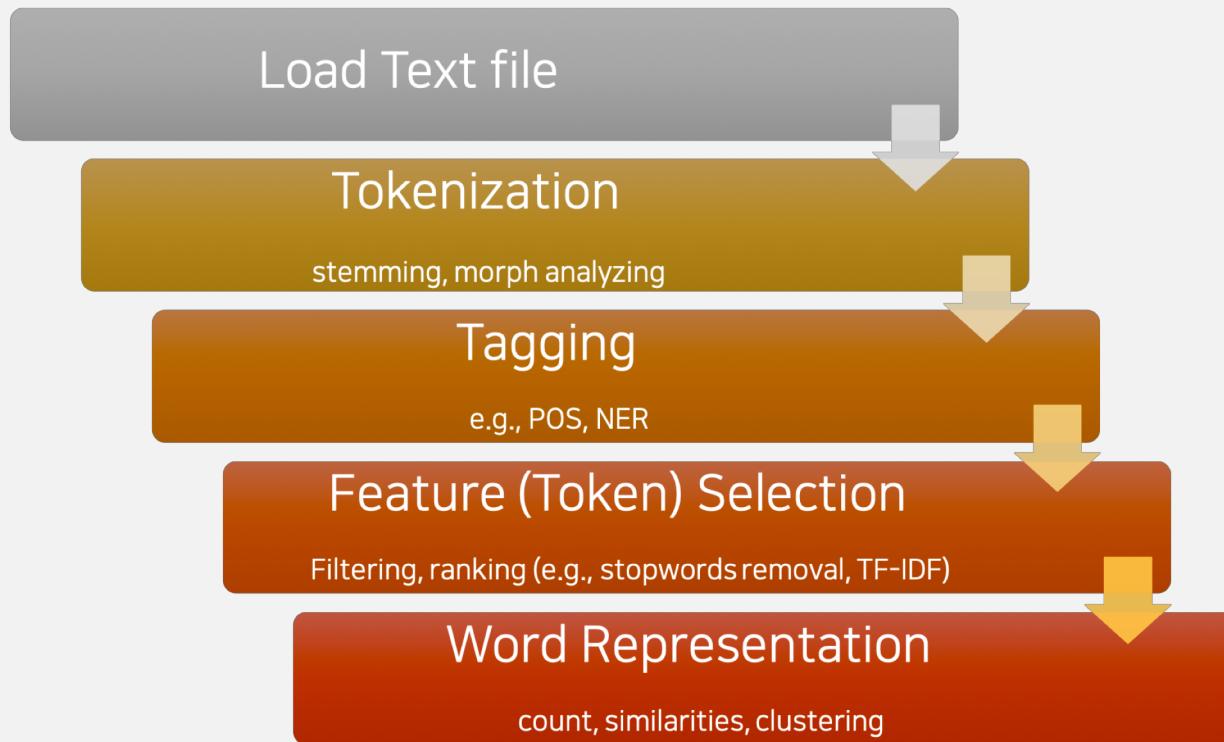
Collect

Preprocess

Feature
Extraction

Evaluation

2 | Preprocessing



3 Korean Morphological Analysis

한국어 형태소 분석기 소개

KoNLPy

국민 청원 데이터로 Wordcloud 만들기



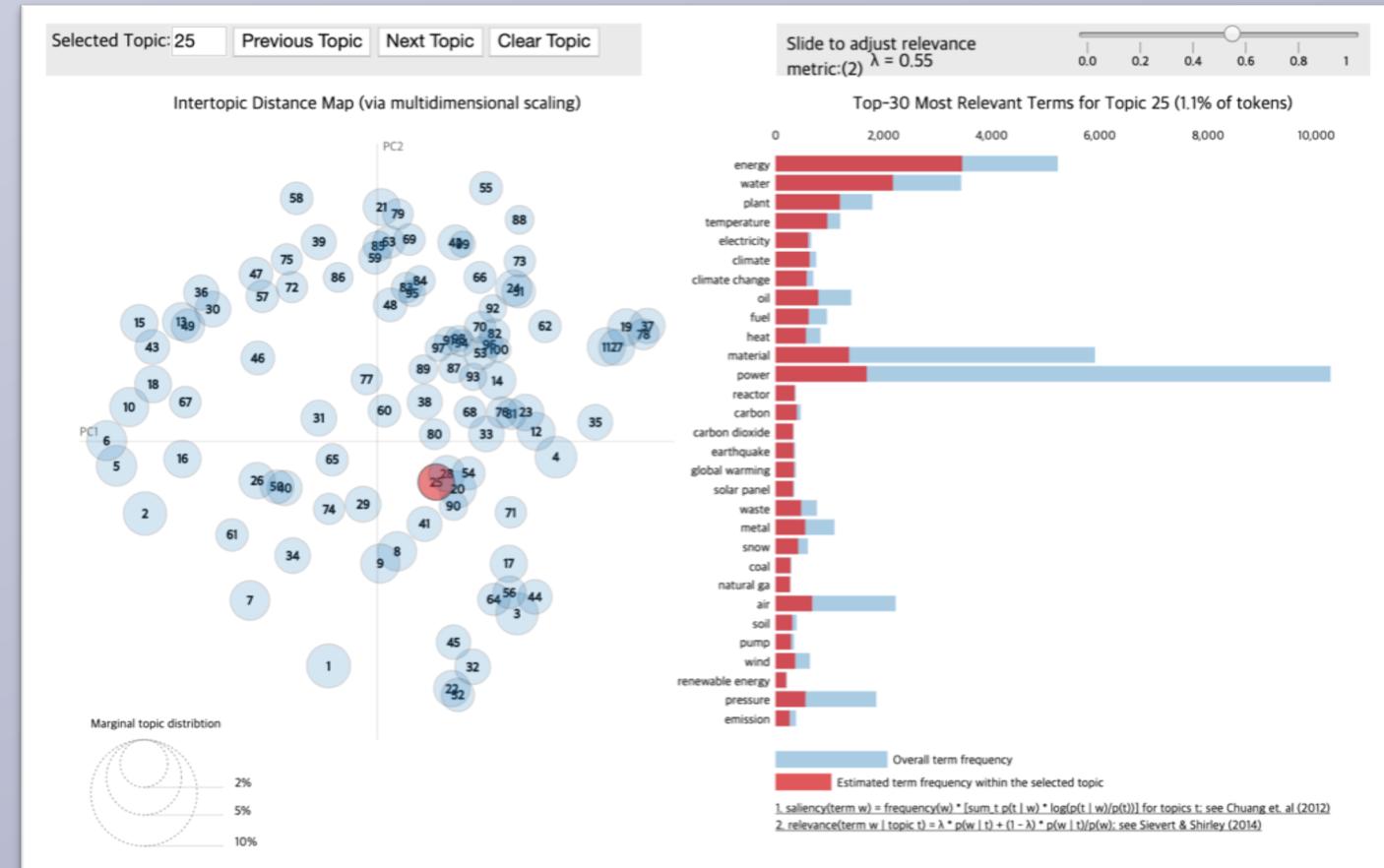
4 Vector Space Model

- Vector Space Model
- Term Document Matrix (TDM)
- TF-IDF (Term Frequency – Inverse Document Frequency)

5 Topic Modeling

Gensim

PyLDAVis 시각화



실습

깃헙 저장소:

<https://github.com/yoonlee78/textmining>