교육
과정
소개서.

텍스트 분석을 위한 머신러닝 CAMP

Fast campus

코스요약

코스명	텍스트 분석을 위한 머신러닝 CAMP
기간	2019. 12. 7 - 2020. 2. 22 (12/28, 1/25 휴강)
일정	매주 토요일 14:00 - 17:00
장소	패스트캠퍼스 강남강의장
준비물	개인 노트북
담당자	02-501-9396 / help-ds@fastcampus.co.kr
수강료	1,400,000
상세페이지 url	fastcampus.co.kr/data_camp_mlfortext

코스목표

벡터 표현법을 기준으로 머신러닝 기법들의 작동 원리에 대해 이해하고, 실제 분석에 이용하는 실습을 통해, 효과적으로 텍스트 분석 모델링을 할 수 있습니다.

코스정보

머신러닝 알고리즘에 기반한 텍스트 분석을 배우고 싶으신 분들에게 추천 드리는 코스이며 다음과 같은 수강생 선수 지식이 필요한 강의입니다.

① 파이썬 학습 경험이 있어 코드를 보면 의미를 이해할 수 있다.
② regression, classification 등 머신러닝 기본 이론을 학습한 경험이 있다.

** 나에게 알맞은 강의인지 잘 모르시겠다면? 담당 매니저와 유선 상담이 가능합니다!
(문의: 02-517-0641)

코스특징

머신러닝, 원리부터 이해하고 직관적으로 배우자!

본 강의는 머신러닝 알고리즘의 핵심 개념을 이해하고, 이를 활용하여 '효율적인' 텍스트 분석을 하는 것을 목표로 직관적인 접근 방식을 지향합니다. 어려운 방법론, 겉핥기식 실습이 아닌 효율적이고 직관적인 텍스트 분석 기법을 배워 보세요.

SOYNLP 라이브러리를 개발한 강사님의 오프라인 직강!

한국어 텍스트 분석을 위한 SOYNLP 라이브러리 개발, PYCON 발표, 텍스트 분석 관련 강의 등 많은 지식과 경험을 보유한 김현중 강사님이 직접 머신러닝과 텍스트 분석의 접목에 대해 친절하고 자세하게 전달합니다.

수업시간 외 다양한 방법으로 자료 제공 및 커뮤니케이션

수업시간으로 끝나지 않습니다. 언제라도 텍스트 분석에 대해 궁금한 점을 해결하실 수 있게 수강생 분들만을 위한 온라인 Q&A가 운영되고, 블로그, 논문, 참고 서적 등 많은 양의 레퍼런스를 제공해 드립니다.

커리큘럼

1회차 토크나이징과 벡터라이징

문서를 벡터로 표현하는 방법 중 하나인 Bag-of-Words Model에 대하여 알아봅니다. 특히 문서를 단어열로 분해하는 방법 중 하나인 품사 판별과 형태소 분석에 대하여 알아봅니다. 이 과정에서 발생하는 미등록단어 문제의 원인을 알아보고 이를 해결하기 위한 여러 방법들에 대하여 알아봅니다.

- 문서를 벡터로 표현하는 과정: 토크나이징, 품사 판별, 형태소 분석, 그리고 KoNLPy
- 미등록단어 문제를 해결하기 위한 비지도기반 토크나이저들
- Bag of Words Model을 이용한 문서 표현

2회차 문서 군집화 (CLUSTERING)

비슷한 문서를 하나로 묶는 방법으로써 문서 군집화 기법이 이용될 수 있습니다. 이는 주제 수준에서의 문서의 벡터 표현과도 같습니다. 다양한 군집화 기법들이 존재하지만, 텍스트 분석에서는 여러 이유로 k-means가 문서 군집화 과업에 효율적이고 효과적입니다. 다른 군집화 기법들은 왜 문서 군집화에 적합하지 않은지, k-means는 왜 문서 군집화에 적합한지, 그리고 잠재적인 위험성은 어떤 것들이 있는지 알아봅니다. 마지막으로 간단한 연관어 / 키워드 추출 기법을 이용하여 군집 별로 레이블링을 하는 방법도 알아봅니다.

- Co-occurrence를 이용한 연관어 / 키워드 추출, PMI / n-gram 추출
- k-means와 그 외의 군집화 기법들을 이용한 문서 군집화
- 키워드 추출 방법을 이용한 군집화 결과 해석

3회차 문서 분류 (CLASSIFIER)

분류기를 이용하여 문서나 문장의 긍/부정, 혹은 그 종류를 분류하는 것은 대표적인 텍스트 분석 과업 중 하나입니다. 이를 이용하여 다양한 분류기들이 이용될 수 있습니다. 특히 Logistic Regression과 Naive Bayes 판별기는 문서 분류를 위한 baseline으로 자주 이용됩니다. 어떠한 특징 때문에 이들이 문서 분류의 기본이 되는지 알아보고, 그 외의 분류 기법들은 문서 분류에 적절한지 알아봅니다.

- Logistic Regression과 L1/L2 Regularization
- Feed Forward Network, Support Vector Machine, Naive Bayes, Decision Tree
- Random Forest, XGBoost, Evaluation Metrics

4회차 임베딩을 이용한 단어/문서의 벡터 표현과 시각화

객체의 특징을 벡터로 더 잘 표현할수록 각 과업의 성능은 향상됩니다. 단어/문서 임베딩은 단어의 문맥이나 형태 혹은 문서의 주제를 보존하는 벡터로 이들을 표현합니다. 이를 위한 다양한 단어/문서 임베딩 방법을 알아봅니다. 또한 임베딩은 고차원의 벡터를 2차원으로 표현하여 시각적으로 고차원 공간을 살펴보는데도 이용될 수

- Word2Vec, Doc2Vec, FastText, PMI + SVD방법에 대하여 알아봅니다.
- t-SNE, PCA, MDS, ISOMAP, LLE, UMAP

커리큘럼

5회차 토픽 모델링을 통한 유사 문서 탐색

Word2Vec과 같은 단어 임베딩 방법은 단어에 대한 문맥 정보를 보존하는 벡터 표현 방법입니다. 토픽 모델링은 이와 비슷한 방법으로 단어와 문서의 주제 정보를 보존하는 벡터 표현 방법입니다. 때로는 이 벡터를 확률 형식으로 만들수도 있습니다. 토픽 모델링에 이용되는 다양한 방법들에 대하여 알아봅니다.

- LSI, pLSI, LDA, NMF를 이용한 문서의 토픽 표현
- LDAvis를 이용한 토픽 모델의 시각화
- Sparse Matrix의 종류 및 이를 효율적으로 다루는 방법

6회차 그래프를 이용한 단어 분석 & 스크래핑을 이용한 데이터 수집

단어나 문서 간 유사도를 학습하는 방법으로 SimRank와 같은 그래프 기반 알고리즘을 이용할 수도 있습니다. 또한 추출 기반 키워드, 핵심 문장 선택을 위해서도 그래프 랭킹 알고리즘들이 이용되었습니다. 이러한 방법들의 원리에 대해 알아봅니다. 그리고 그래프 기반 모델링 방법들과 임베딩 기법들의 관계에 대해서도 알아봅니다.

- PageRank, TextRank, KR-WordRank를 이용한 그래프 랭킹 기반 키워드 탐색
- SimRank, Random Walk with Restart를 이용한 토픽 유사어 탐색
- 임베딩을 이용한 그래프의 시각화
- Beautiful Soup을 이용한 데이터 수집 실습

7회차 객체명 인식과 순차적 레이블링

순차적 레이블링 방법은 문장 내 단어들의 종류를 판별하기 위해 이용됩니다. 객체명 인식은 순차적 레이블링 기법을 이용하는 대표적인 과업입니다. Sparse Representation을 이용하는 대표적인 순차적 레이블링 기법인 CRF(Conditional Random Field)에 대하여 알아보고, Embedding Vector를 이용하는 순차적 레이블링 기법인 RNN(Recurrent Neural Network)과 그 후속 모델들에 대하여도 알아봅니다.

- CRF(Conditional Random Field)를 이용한 객체명 인식기
- RNN to LSTM-CRF

8회차 어텐션 방법을 이용한 개선된 문서 분류기

어텐션 기법은 입력 데이터에서 과업에 필요한 정보들을 선택적으로 강조함으로써 모델의 성능을 향상하는 방법입니다. 어텐션 기법의 발전 과정과 이를 이용하는 문서 분류 방법에 대하여 알아봅니다.

- Sequence to Sequence and Attention Mechanism
- Attention Based Sentence Classifier & HAN
- PyTorch

커리큘럼

9회차 ● **CNN (CONVOLUTIONAL NEURAL NETWORK)을 이용한 문서 분류기**

CNN(Convolutional Neural Network)은 Locality 정보를 활용하는 모델입니다. CNN의 원리를 알아보고, 텍스트 분석에서 이를 이용하는 문서 분류 방법에 대하여 알아봅니다.

- CNN(Convolutional Neural Network)
- Word/Char Level CNN for Sentence Classifier

10회차 ● **최인접이웃 검색과 오탈자 교정**

k-NN 기반 분류/회귀 모델은 가장 직관적인 머신러닝 기법이지만, 경우에 따라서는 이것만으로도 충분히 좋은 성능을 보여줍니다. 데이터의 개수에 비례한 검색 비용 문제를 해결하고 효율적으로 최인접이웃 기반 모델이 작동하게 만드는 방법에 대하여 살펴봅니다. 마지막으로 수업의 전체적인 내용들을 data representation 관점에서 review 합니다.

- 효율적인 k-NN 검색기, LSH, Tree Based Indexer, NN-descent
- Edit Distance 를 이용한 오탈자 교정
- review

강사소개

김현중

[약력]
-서울대학교 산업공학과 데이터마이닝 연구실 박사과정
-한국어 텍스트 분석을 위한 soynlp 라이브러리 개발
-PYCON KOREA 2017 '노가다 없는 텍스트 분석을 위한 한국어 NLP' 발표

수강환경

강남강의장

❖ 강의에 따라 강의장이 변경될 수 있습니다.