자연어 처리 교육과정 로드맵

자연어 전처리

토큰화
- 문장 토큰화
- 단어 토큰화

품사태깅
- 문장 토큰화
- 단어 토큰화

원형복원
- 어간 추출(Stemming)
- 표제어 추출 (Lemmatization)

불용어처리
- 불용품사 제거
- 불용어 제거

구문분석
- 구문분석

표현(Representation)

단어의 표현
- 원핫 인코딩

문서의 표현
- Bag of Word (BoW), TDM
- TF-IDF

Word embedding
- Word2Vec
- Globe

차원축소
- 특이값 분해 (SVD)
- 주성분 분석(PCA)

분석(Analysis)

핵심키워드 추출
- TF-IDF
- Text rank

문서요약
- Luhn summerizer
- Text rank

토픽모델링
- LSA
- LDA

분류
- 회귀모델, TF-IDF
- RNN, CNN
- Naive bayes classifier

감성분석
- 사전기반
- CNN

최신논문

BERT
- Seq2seq, Transfomer
- BERT

SOTA 모델
- XLNet
- GPT2

논문구현
- 텍스트마이닝 활용
  금융통화위원회 의사록 분석
- 고객 리뷰를 활용한 VoC
- CNN 활용한 뉴스 분류