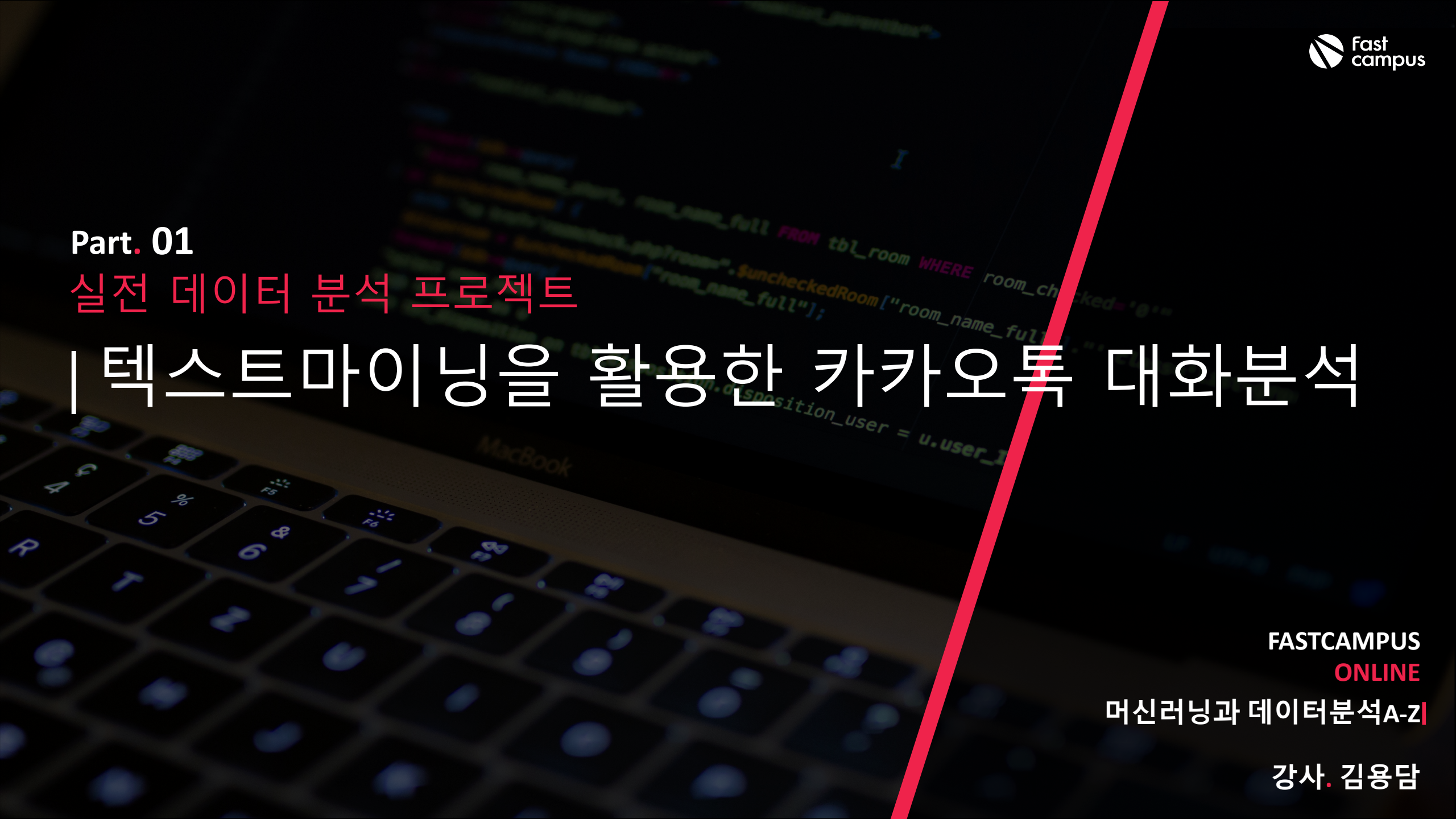


Part. 01

실전 데이터 분석 프로젝트

| 텍스트마이닝을 활용한 카카오톡 대화분석

FASTCAMPUS
ONLINE
머신러닝과 데이터분석A-Z|

강사. 김용담

I 텍스트 마이닝 심화

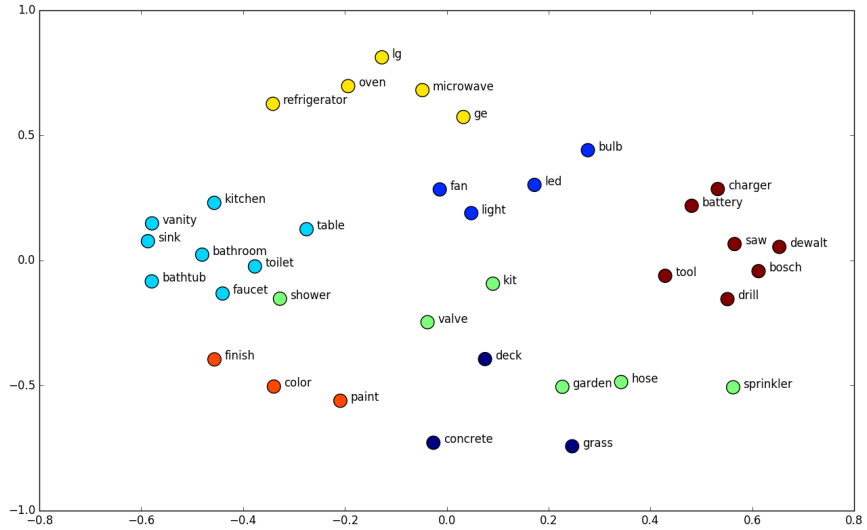
머신러닝 기법 LDA(토픽 모델링 기법), SVM(문서 분류 기법) 등

워드 임베딩(Word Embedding)

입력값

딥러닝 기법 RNN, LSTM, Transformer, BERT 등

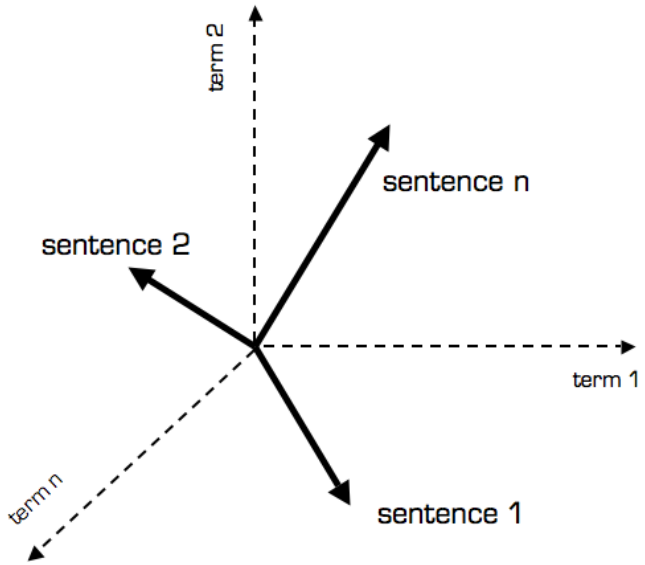
I 워드 임베딩이란?



Word Embedding

- 단어를 컴퓨터가 이해할 수 있는 벡터로 표현하는 방법
- Word Embedding = Word + Embedding
- Sparse Representation (BOW, TF-IDF)
- Dense Representation(word2vec, Glove 등)

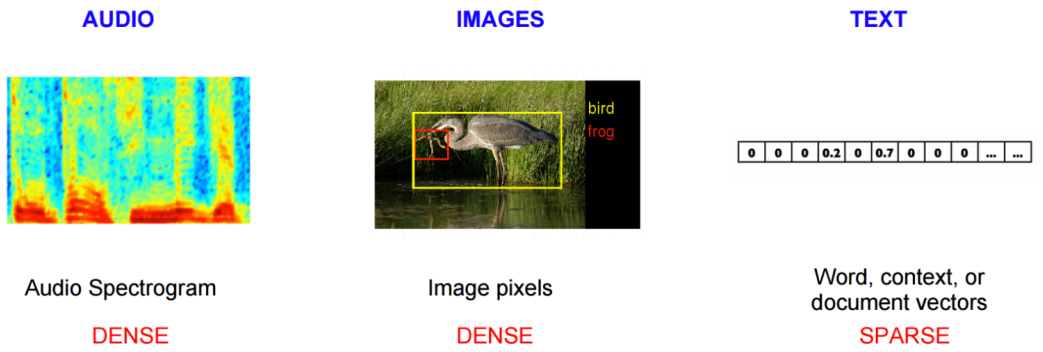
I 희소 표현(Sparse Representation)의 문제점



Vector Space Model

- 문서 데이터에 존재하는 모든 유니크한 단어 수가 벡터의 차원이 되어 고차원 공간이 됨
- 단어의 문맥 정보가 사라짐
  예) 문장 내 순서(word order), 문장 내 동시등장 (co-occurrence)
- 차원의 저주(Curse of dimensionality)로 인해 분석 기법의 성능이 악화됨

I 밀집 표현(Dense Representation)



Dense Representation

- 이미지나 오디오 데이터는 양질의 고차원 데이터로 표현됨 (dense representation)
- 기존 방법인 VSM은 단어를 discrete symbol로 표시하기 때문에 정보 전달력이 떨어짐
- 기존의 count-based method가 아닌 predictive model을 사용하여 단어의 주변 정보를 반영한 dense representation을 표현함

Reference : https://tensorflowkorea.gitbooks.io/tensorflow-kr/content/g3doc/tutorials/word2vec/

Part. 01

실전 데이터 분석 프로젝트

감사합니다.

FASTCAMPUS
ONLINE
머신러닝과 데이터분석A-Z

강사. 김용담