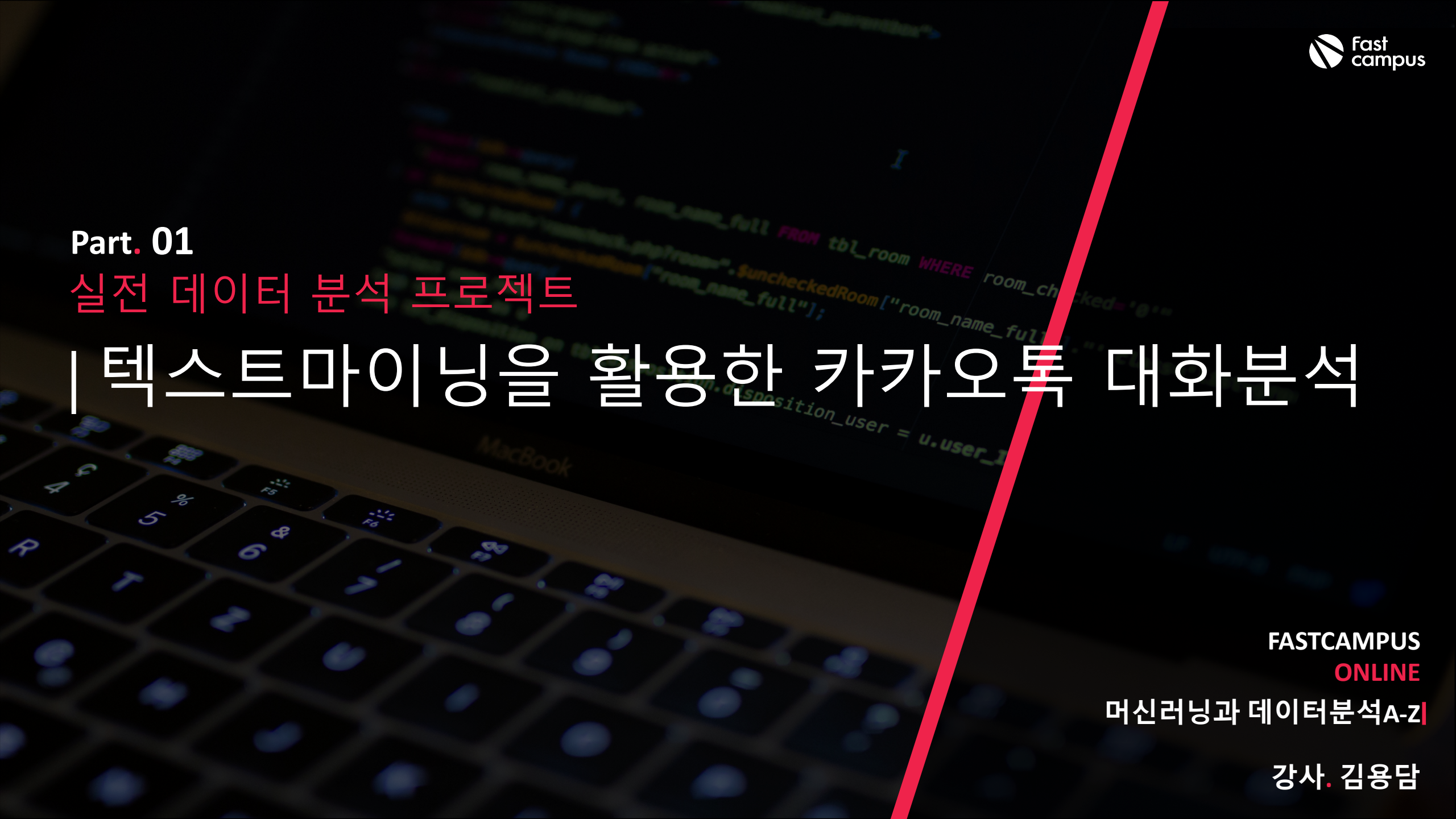


Part. 01

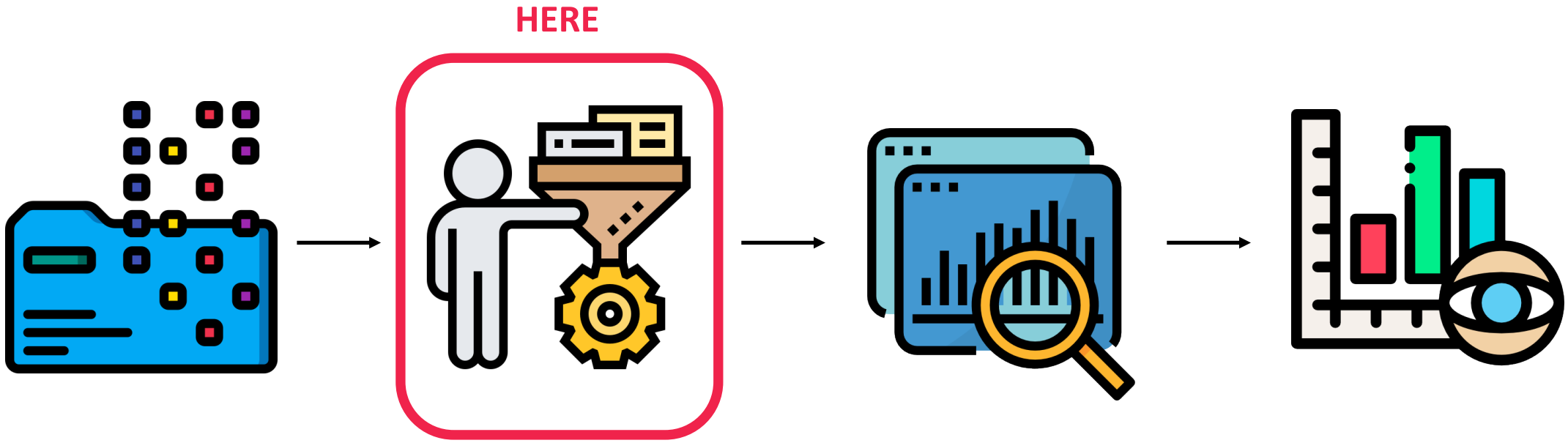
실전 데이터 분석 프로젝트

| 텍스트마이닝을 활용한 카카오톡 대화분석

FASTCAMPUS
ONLINE
머신러닝과 데이터분석A-Z|

강사. 김용담

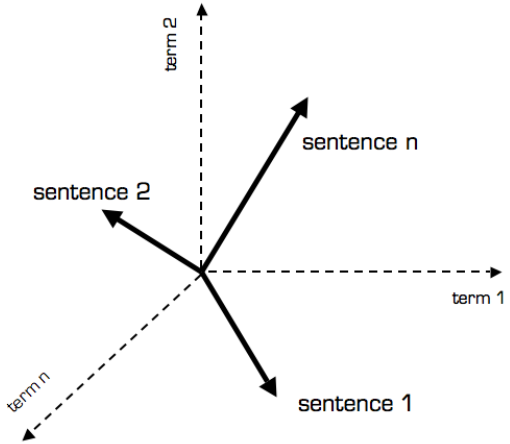
텍스트 마이닝 과정



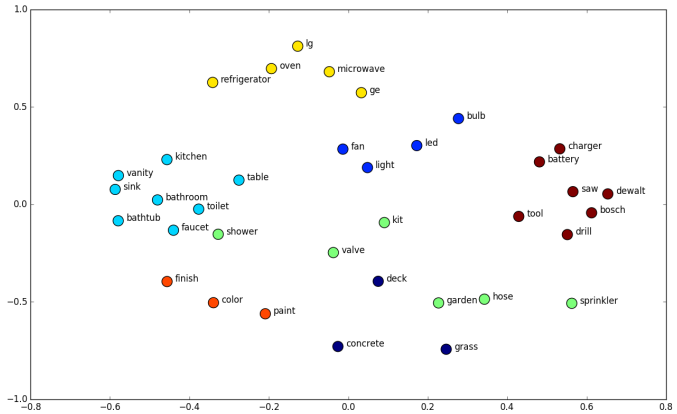
텍스트 전처리 과정 이해하기

텍스트 인코딩(Text Encoding)
텍스트를 벡터로 표현하기

paint / picture / days / young



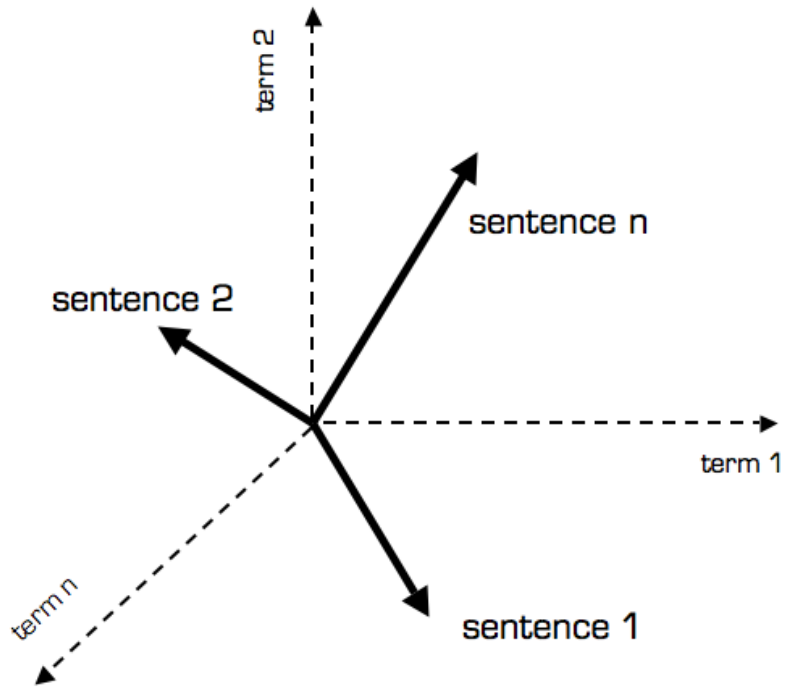
Vector Space Model



Word Embedding

FAST CAMPUS
ONLINE
김용담 강사.

I 벡터 공간 모델(Vector Space Model)



- 문서를 벡터로 표현으로 나타내는 방법
- 벡터의 특정 차원은 하나의 단어를 의미
- 대표적인 방법으로 Bag-of-Word(BOW)와 Term Frequency – Inverse Document Frequency(TF-IDF)
- 정보 검색이나 문서 랭킹 시스템에서 주로 사용

I Bag-of-Words(BOW)

문서 1: 나는 강아지를 좋아한다

문서 2: 나는 강아지와 산책을 좋아한다

문서 3: 산책하는 것은 나의 취미이다

	나	강아지	산책	취미
문서 1				
문서 2				
문서 3				

I Term Frequency – Inverse Document Frequency

- $v_d = (w_{1,d}, w_{2,d}, \ldots, w_{N,d})$
- $tf(t,d) = f_{t,d}$
- $idf(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$
- $tfidf(t,d,D) = tf(t,d) \cdot idf(t,D)$

- $tf(t,d)$ 는 문서 d안에 있는 각 단어 t의 빈도
- $idf(t,D)$는 단어 t가 전체 문서 D에서 등장한 문서의 수의 역수
- $tfidf(t,d,D)$는 문서 D안에 있는 특정 문서 d안에 있는 각 단어 t의 $tf(t,d) \cdot idf(t,D)$의 값

I Term Frequency – Inverse Document Frequency(TF-IDF)

문서 1: 나는 강아지를 좋아한다

문서 2: 나는 강아지와 산책을 좋아한다

문서 3: 산책하는 것은 나의 취미이다

	나	강아지	산책	취미
문서 1				
문서 2				
문서 3				

Part. 01

실전 데이터 분석 프로젝트

감사합니다.

FASTCAMPUS
ONLINE
머신러닝과 데이터분석A-Z

강사. 김용담