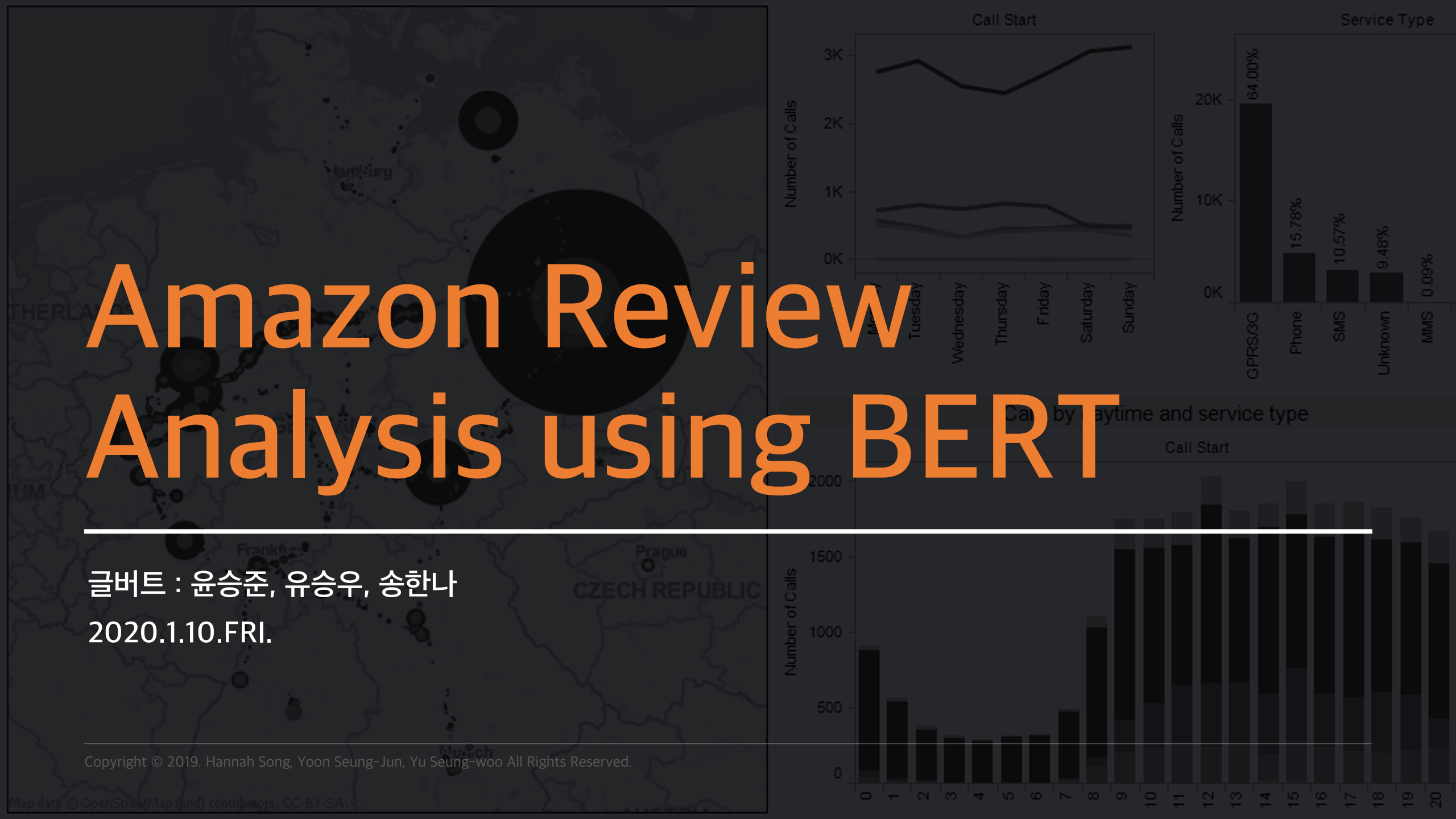
Amazon Review Analysis using BERT

글버트 : 윤승준, 유승우, 송한나

2020.1.10.FRI.



팀 구성원

글버트(Geul-BERT) 팀

윤승준

웹 서비스 /
머신 러닝 /
딥 러닝

유승우

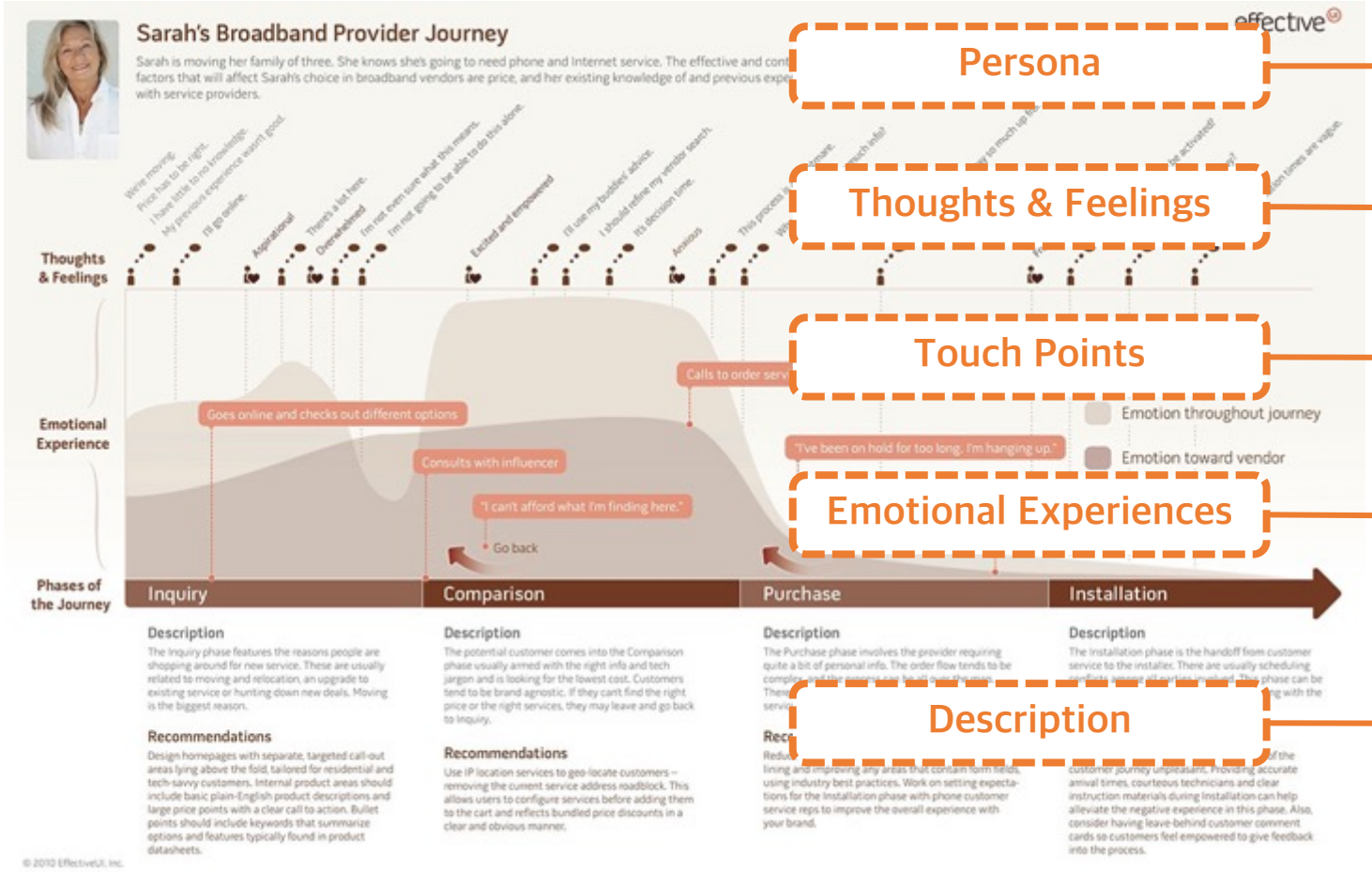
머신 러닝 /
딥 러닝

송한나

머신 러닝 /
딥 러닝 /
기획 및 디자인

프로젝트 주제 | BERT를 활용한 아마존 리뷰 분석

인공지능을 활용한 아마존 리뷰 분석



Persona → Clustering

Thoughts & Feelings → Topic modeling

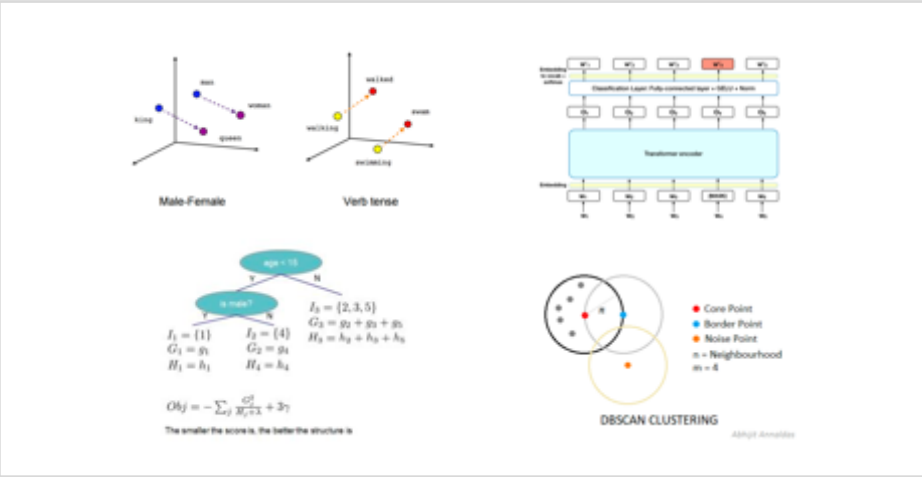
Touch Points → Word embedding

Emotional Experiences → Sentiment analysis

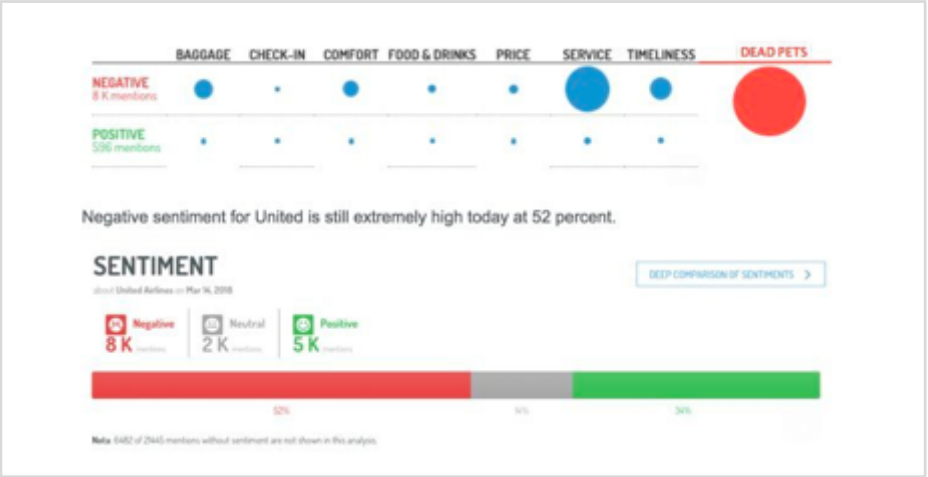
Description → Text generation

결과물 목표

Text Analysis and Generation using ML/DL/NLP

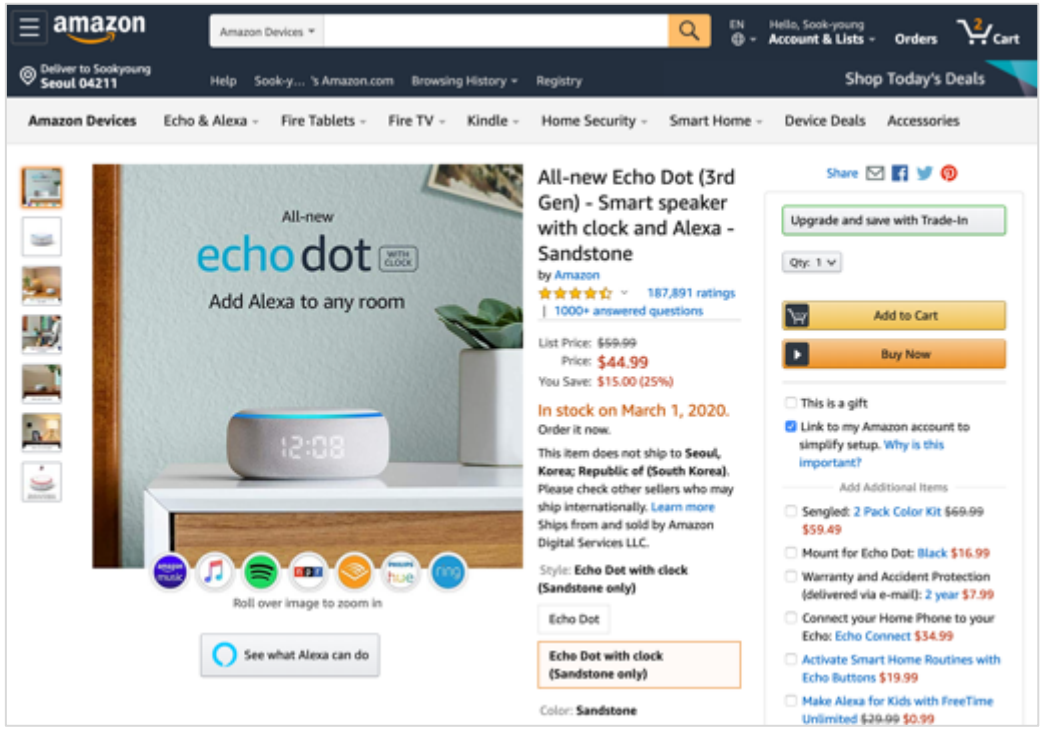


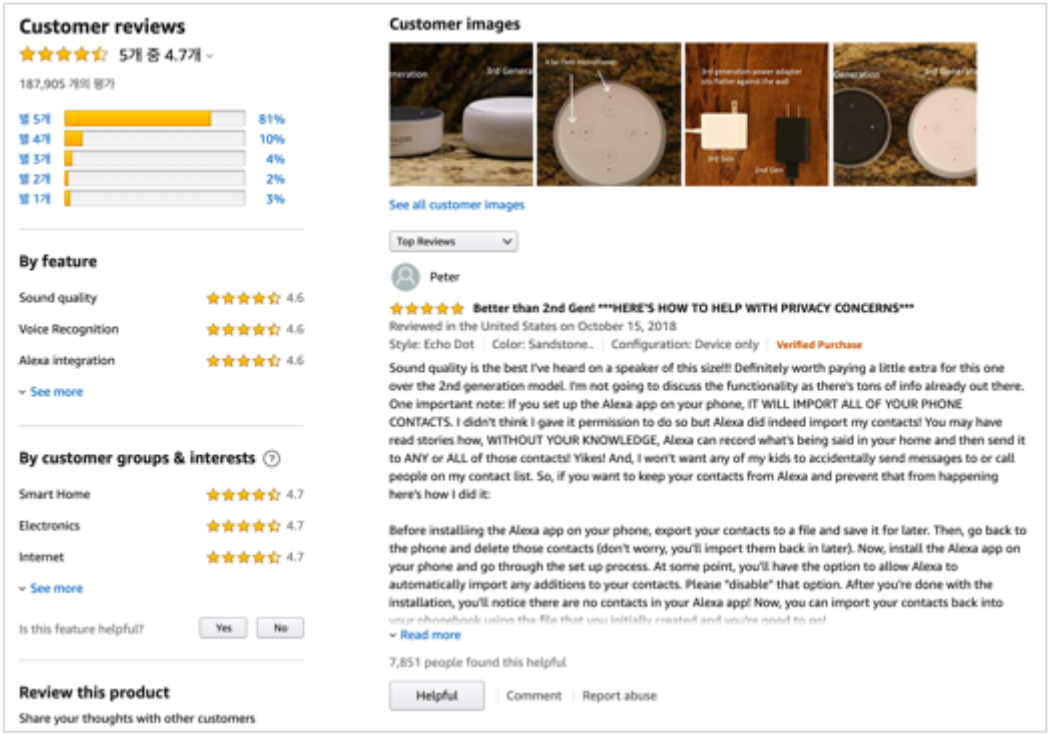
Data Visualization on Web Service



리뷰 데이터 수집

- **Amazon Online Review** : 스마트 스피커 Amazon Echo 제품에 관한 리뷰 데이터 수집.
- title, author, date of creation, product ratings, product, review text, verification





사용 기술

Pre-processing

데이터 정제 및 전처리

- 토큰화(Tokenization → 정제(cleaning) → 정규화(normalization) → 표제어 추출(Lemmatization) → 어간 추출(Steaming)

워드 임베딩

- Word2Vec

Clustering

문서 유사도

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

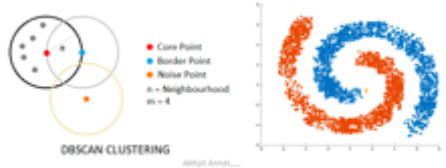
TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

- TF-IDF(Term Frequency-Inverse Document Frequency
- 코사인 유사도(cosine similarity)

군집 알고리즘



- DBSCAN(Density-Based Spatial Clustering of Applications with Noise)

Sentiment Analysis

감성 분석

- 주관성 탐지 (Subjectivity Detection) : (긍정, 부정, 중립, 객관 등의 지표 사용).
- 극성 탐지(Polarity Detection) : 자료가 긍정인지 부정인지 파악

분류 알고리즘

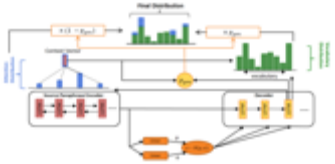
- XGBoost(eXtreme Gradient Boosting)

Text Summarization

언어 모델

- BERT(Bidirectional Encoder Representations from Transformers)

문서 요약



- VAE-PGN based on Bert Sum (Bert-based Summarization Model)

일정계획

1일차

- 팀 빌딩
- 주제 선정
- 웹 스크레이핑

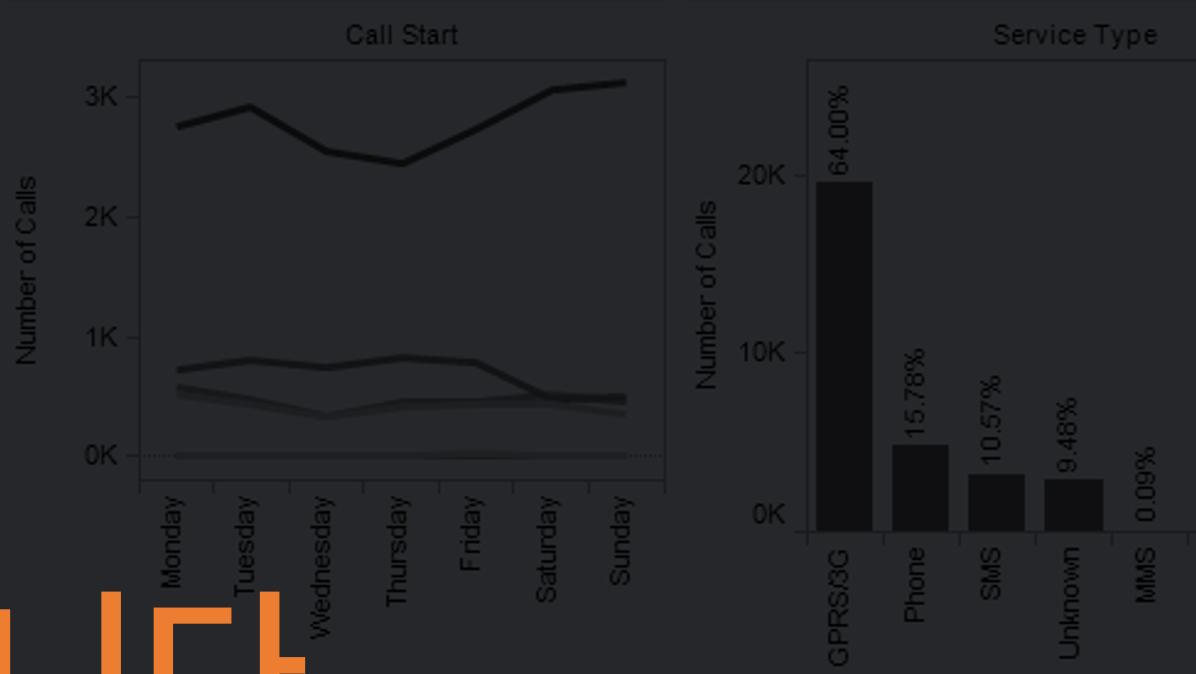
2일차

- 데이터 전처리
- 자연어처리
- 언어 모델 학습

3일차

- 감성 분석
- 시각화

감사합니다.



Calls by daytime and service type