안녕하세요 반갑습니다.

엄인성 포트폴리오 2020

목차

사용가능 기술

0

프로그래밍 언어 및 분석 툴에 대한 숙련도에 대한 설명.

기술

Python

데이터 분석을 할때 가장 많이 사용한 언어이며, 분석에 필요한 다양한 라이브러리를 활용가능합니다.

PyTorch

기계학습 프레임워크이며 Keras와 함께 가장 많이 사용되는 기계학습 프레임워크이며, BERT모델을 구축하기 위해 사용하였습니다.

SQL(Oracle)

기본적인 DDL, DML, DCL을 어려움 없이 사용가능합니다.

Keras, Tensorflow

기계학습과 관련된 프로젝트에서 가장 많이 사용한 프레임워크이며, 다양한 모델(Seq2Seq, Fast RCNN 등)을 구현한 경험이 있습니다.

NoSQL(MongoDB)

우분투 클라우드서버에서 DB관리를 위해 설치하였고, 관리를 담당하였습니다.

Linux(Ubuntu)

기계학습과 관련한 프로젝트를 위해 사용한 운영체제이며, 기본적인 리눅스 명령어 사용에 어려움이 없습니다.

기술

Git

개인 프로젝트를 형상관리를 위해 해본 경험이 있으며, 관련된 명령어를 사용할 수 있습니다.

R

데이터 분석을 위한 기본 라이브러리 및 문법사용가능이 가능합니다.

Slack

업무진행을 위해 사용한 협업도구이며 약 2년간 사용한 경험이 있습니다.

Java

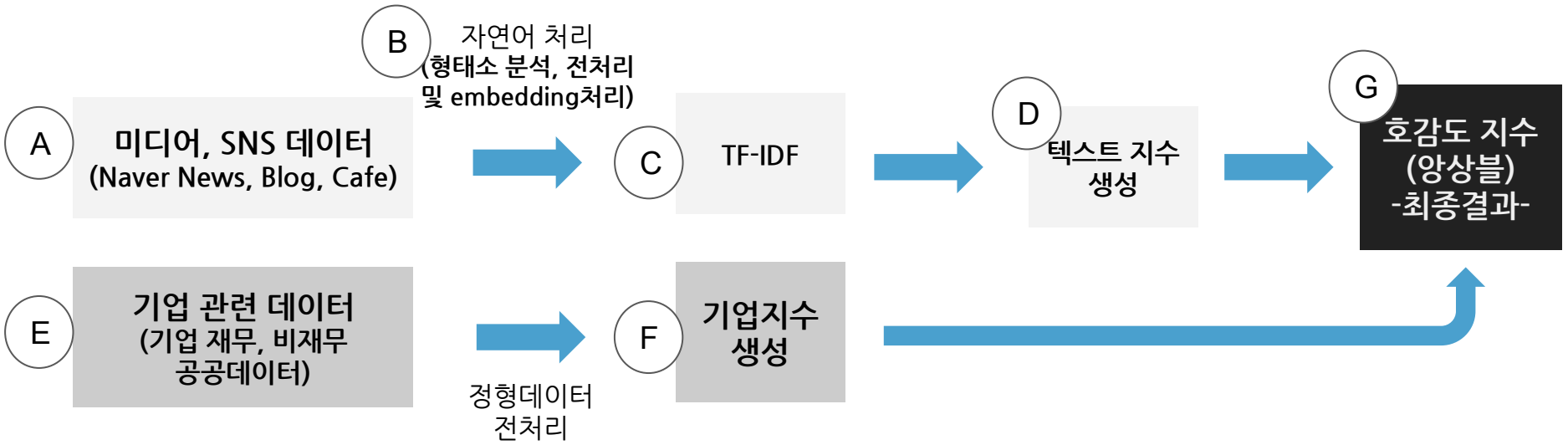
클래스와 인터페이스를 정의하는데 어려움이 없으며 기본적인 문법을 사용가능합니다.

LH공사 호감도 지수 생성

1

- 분석 및 개발 기간: 2019.02 - 2019.05*(3개월)*
- **챗봇 관련 기업 와이즈넛과 함께 진행한 LH공사에 대한 호감도를 생성하는 프로젝트**
- 운영체제: Windows 10
- 개발언어: Python 3.5
- 개발툴: Jupyter notebook
- **분석패키지: Gensim, Statemodels, Plotly, Pandas, WordCloud, PyLDAvis**
- **분석알고리즘: LDA, TF-IDF, Word2Vec, ARIMA, Ensemble**

분석 프로세스



A. w업체로 부터 전달 받은 언론, SNS 데이터
B. 전달 받은 데이터를 자연어 처리 진행.
   a. 형태소 분석 → 불용어 → 단음절 제거 → word to index
C. TF-IDF알고리즘을 이용하여 주요단어를 선별.
D. 텍스트 지수 생성
   a. 언론사: 언론사별 월별 가중치를 계산 후 지수를 생성.
   b. SNS: 광고 내역을 제거한후 지수 생성
E. 기업 관련 데이터
   a. 기업 재무, 사업 데이터, 공공데이터수집
F. 기업 지수 생성
G. 텍스트 지수와 기업지수를 앙상블 하여 최종 호감도 지수 생성.

주요 프로세스

〈 언론사별 뉴스 영향도 모델링 〉

* 일별 언론사 중요도 및 뉴스별 영향도 모델링을 위해 다음 가중치를 생성.

$$\text{언론사 가중치} = \frac{\text{언론사 문서 개수}}{\text{전체 문서 개수}}$$

$$\text{중요 키워드} = \text{언론사 가중치} \times \text{문서 단어별 } TF-IDF \times \text{문서 단어빈도} \times \text{단어별 감성 점수(감성사전기반)}$$

$$\text{문서별 영향도 지수} = \sum_{i=1}^{n} \text{문서별 중요 키워드}_i$$

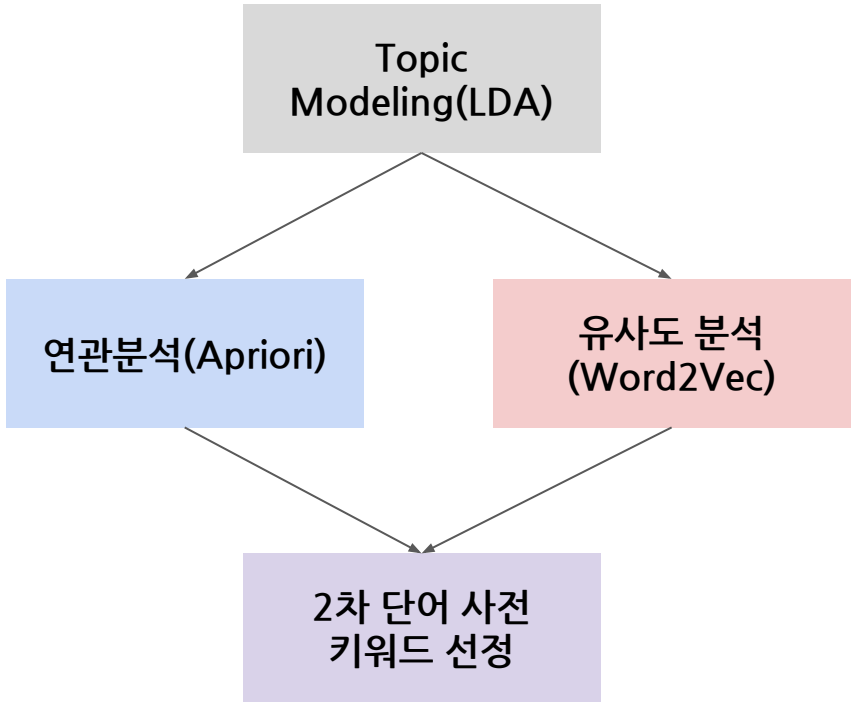
$$\text{일별 영향도 지수} = \sum_{i=1}^{n} \text{문서별 영향도 지수}_i$$

$$\text{일별 언론사 영향도 지수} = \sum_{i=1}^{n} \text{당일 동일 언론사 문서 영향도 지수}_i$$

$$\text{일별 언론사 중요도} = \frac{\text{일별 언론사 영향도 지수}}{\text{일별 영향도 지수}}$$

〈 감성 사전 구축 과정 〉

* 감성점수를 생성하기 위한 보충 감성 사전 생성.



문제발생및 해결

1. TF-IDF 계산 시간

문제 발생: TF-IDF를 Python으로 구현하고 분석하는 것에 대한 비용이 매우 컸음.
문제 극복: 스트링 값에 대한 연산비용이 매우 크기 때문에 단어들을 index map처리 하고 임계값을 적용하여 불필요한 연산처리 제어하도록 하였음.

2. 기업관련 데이터

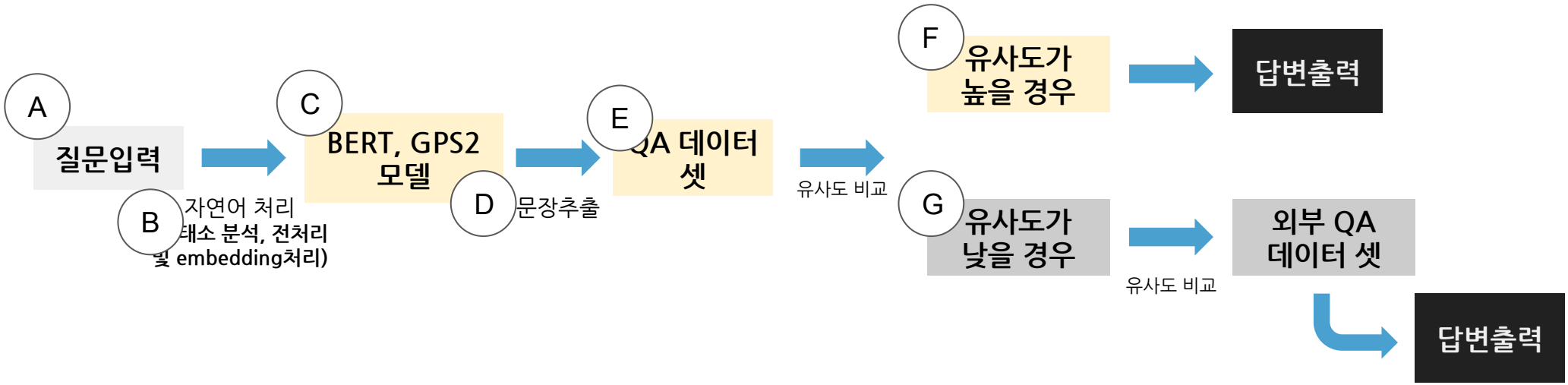
문제 발생: LH공사와 관련된 데이터를 찾을 수 없었음
문제 극복: 공공기관 경영정보 데이터 시스템을 통해 LH 한국토지주택공사에 관련된 데이터를 수집할 수 있었음

음원 추천 챗봇 시스템

2

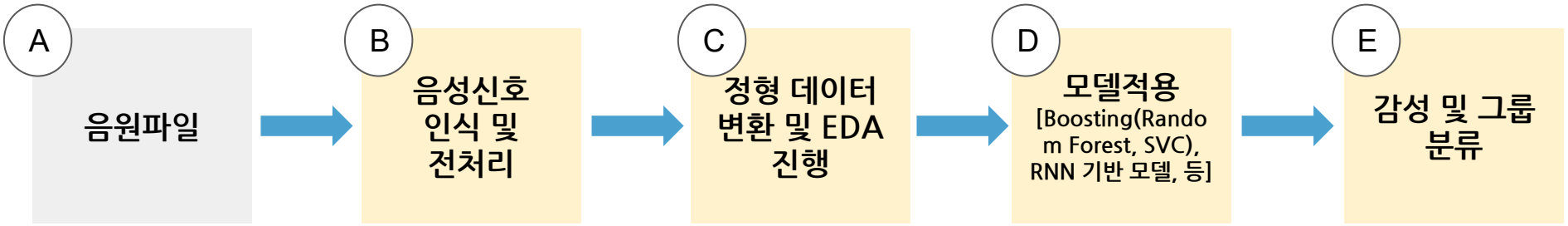
- 분석 및 개발 기간: 2020.,03 - 2020.05*(3개월)*
- **챗봇 모델을 생성하고 음원 장르를 분류하여 음원을 추천하는 챗봇 시스템을 개발한다.**
- 운영체제: Windows 10
- 개발언어: Python 3.5
- 개발툴: Jupyter notebook
- **분석패키지: Tensorflow, Keras, Pytorch, Numpy, BeautifulSoup, Selenium, librosa**
- **분석알고리즘: Sequence to Sequence, Attention, BERT, Boosting, Euclidean distance, Decision Tree**

챗봇 분석 프로세스



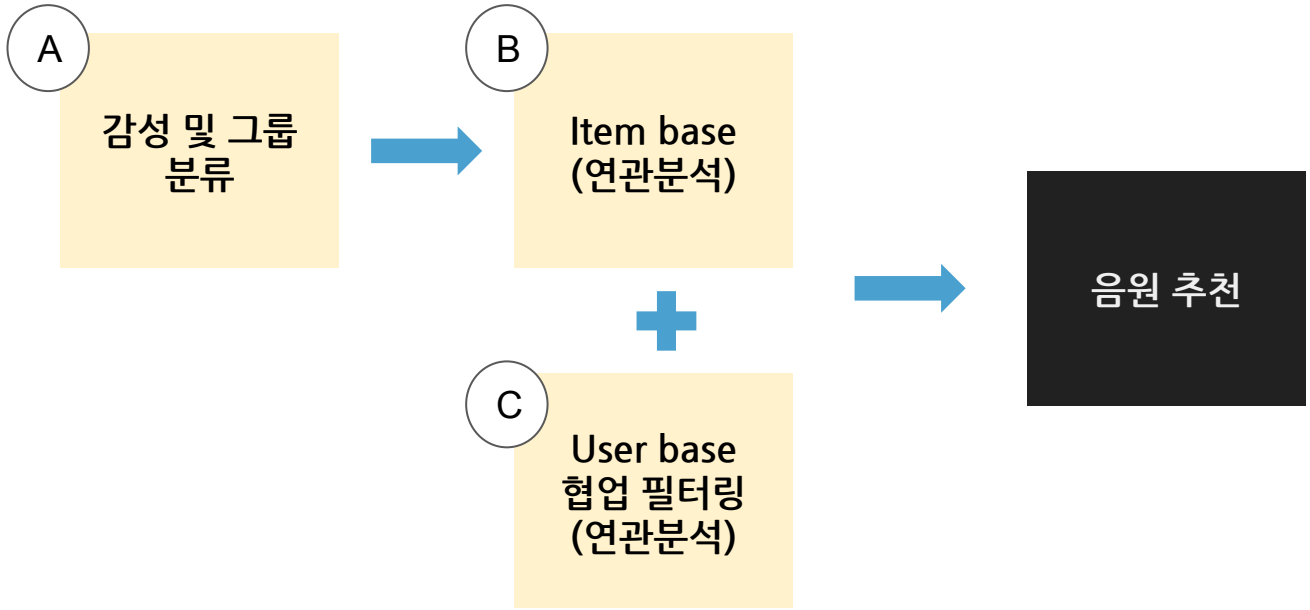
A. 챗봇에 질문을 입력.
B. 전달 받은 데이터를 자연어 처리 진행.
   a. 형태소 분석 → 불용어 → 단음절 제거 → word to index
C. 정제추출된 형태소를 BERT, GPS2 모델에 학습시킨다.
D. 학습된 모델을 통해 문장을 추출
E. 추출된 문장이 완성도 높은 문장이 아니 때문에 미리 존재된 Q/A데이터 셋이 유사도 분석을 통해 완성된 문장을 출력.
F. 유사도가 기준점 이상 일경우 최종 답변을 출력.
G. 유사도가 기준점 이하 일경우 외부에 존재한는 답변 출력.

음원 분류 분석 프로세스



A. 음원파일을 파싱.
B. 파싱된 음성에 대한 신호를 인식, 전처리 진행.
C. 정형데이터 변환 및 EDA를 통해 크리티컬 변수를 추출.
D. 여러 모델에 적용한후 Boosting 기법을 통해 융합모델을 생성.
    a. RandomForest / SVC / RNN
E. 최종적으로 감성 및 그룹을 분류.

추천 시스템 프로세스



A. 음원 분류 분석에서 추출된 유저별 최종 결과를 입력으로 사용한다.
B. item base 분석.
C. 협업 필터링을 통해 유사 그룹을 생성하고 음원을 추천한다.

문제발생과 극복

1. 한글과 관련한 챗봇 사전 학습 모델 부재
BERT나 GPT2 모델 자체는 생성할 수 있으나 실제 서비스에 적용하기 위해서는 대량의 데이터 학습된 모델이 필요하였으나 모델에 학습시킬 데이터가 매우 부족하였음.

따라서 기존 Q,A명사들의 유사단어를 추출하여 데이터 셋에 적용하는 방식으로 데이터에 대한 양을 증대시킴.

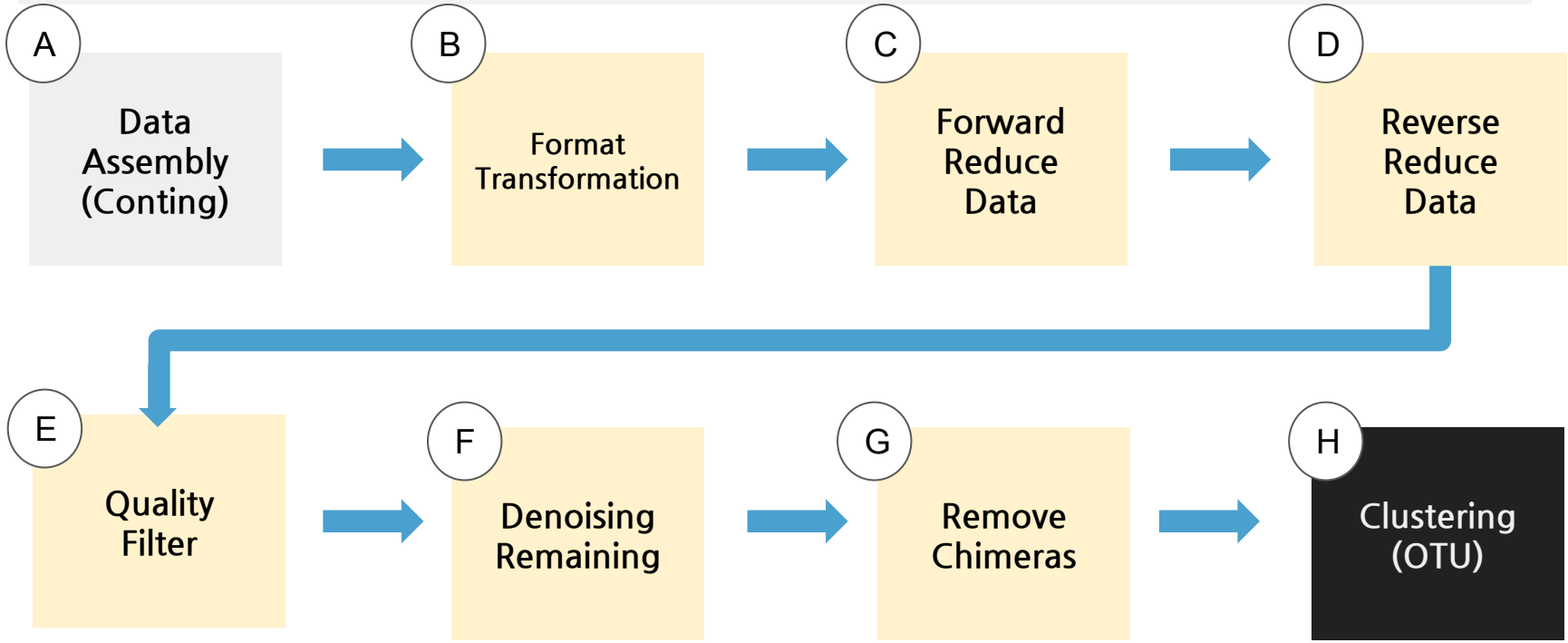
반려견 유전체 분석 관리시스템 (MMS)

3

- 분석 및 개발 기간: 2019.,02 - 2020.05*(3개월)*
- **반려견 유전체분석을 위한 관리시스템 구축.**
- 운영체제: Windows 10
- 개발언어: Python 3.5
- 개발툴: Jupyter notebook
- **분석패키지: Qiime2**
- **분석알고리즘: 유사도알고리즘(Jaccard distance)**

분석 시스템 프로세스

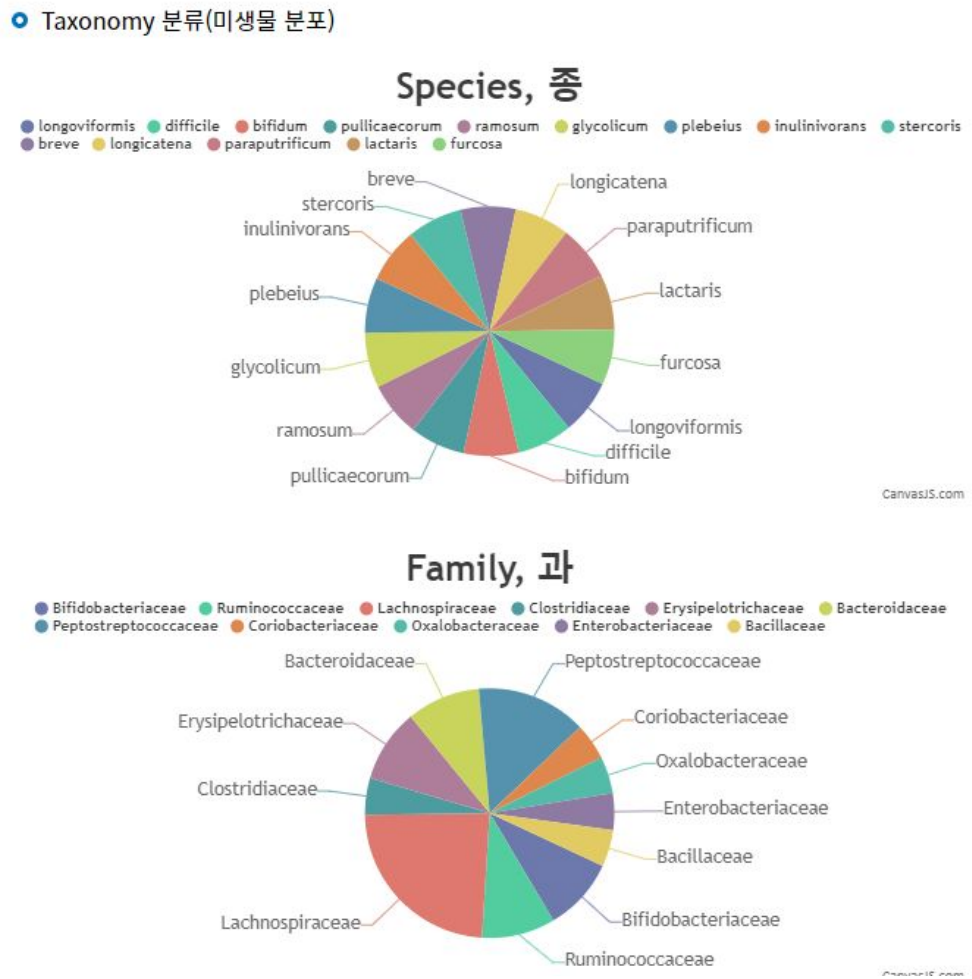
A. 실제 동물 병원에서 추출된 반려견 샘플을 유전 시퀀스 생성기업에 전달하고 시퀀스 데이터를 생성.
B. 분석에 사용될 포맷으로 변경 수정한다.
C. RNA의 유전체에서 Forward부분을 추출.
D. RNA의 유전체에서 Reverse부분을 추출.
E. 추출된 RNA 샘플에서 질낮은 데이터를 제거.
F. 노이즈를 제거.
G. Chimeras를 탐색하여 제거.
H. 최종적으로 존재하는 RNA 시퀀스에서 Taxonomy를 분류.(분류분석은 QIIME에서 사용된 패키지중 외부 패키지인 Sikit learn의 머신러닝을 이용하여 분류분석을 진행.)



결과

분석

전과정에 대한 분석은 QIIME라이브러리를 이용하여 진행하였음.



문제발생과 극복

처음 해보는 생물 관련 분석이라 어떤 순서로 분석을 진행할지에 대한 어려움이 많이 있었다.
분석을 진행하기 위해서 Qiime 라이브러리의 도서와 도큐먼트 웹페이지를 번역 분석하였고 기간은 약 3달정도 걸렸다.

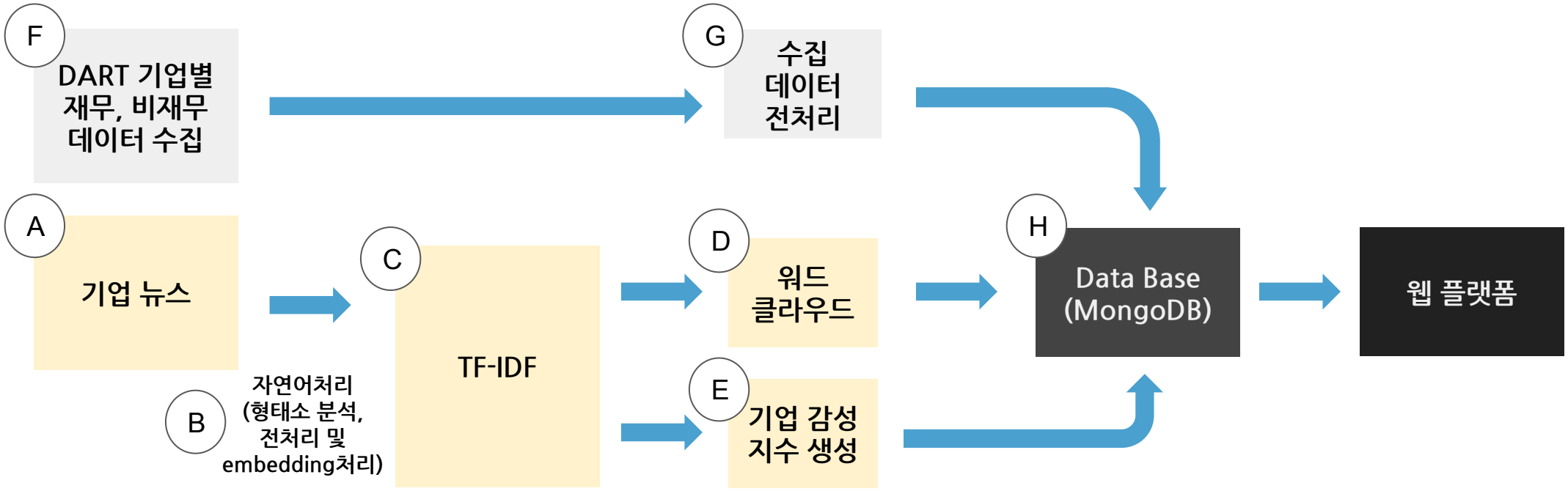
그 내용을 바탕으로 분석을 진행하였고 최종적으로 국민대 수의학과 연구실의 박사에게 어드바이스를 받으며 분석에 신빙성을 더 하였다.

그로인해 정부과제에서 합격 판정을 받아 프로젝트를 성공리에 마무리 하였다.

Nowmoment Company 4

- 분석 및 개발 기간: 2018.,05 - 2020.03*(1년 10개월)*
- **300여개의 상장기업의 성장성을 분석 및 예측하고 기업 보고서를 제공하는 플랫폼.**
- 운영체제: Windows 10
- 개발언어: Python 3.5
- 개발툴: Jupyter notebook
- **분석패키지: Python, keras**
- **분석알고리즘: LDA, TF-IDF, Word2Vec, ARIMA, Ensemble, RNN, LSTM**
- **데이터 베이스: NoSQL(Mongo DB)**

분석 프로세스



A. 기업관련 뉴스 수집(크롤링)
B. 전달 받은 데이터를 자연어 처리 진행.
   a. 형태소 분석 → 불용어 → 단음절 제거 → word to index
C. TF-IDF를 이용하여 주요 단어 추출
D. 핵심단어를 파악하기 위해 워드 클라우드를 생성.
E. 2 에서 처리된 자연어를 통해 기업에 대한 감성지수를 생성.
F. DART에서 기업별 재무, 비재무 데이터를 수집.
G. 수집된 데이터를 전처리하여 정리.
H. TF-IDF, 워드 클라우드 키워드, 기업 감성점수, 수집된 수치 데이터를 일별로 MonogoDB에 입력.

최종결과

전통산업 기업분석 리포트

[유통] BGF리테일 282330 (코스피) PDF Download 화면출력

주가정보

재무재표

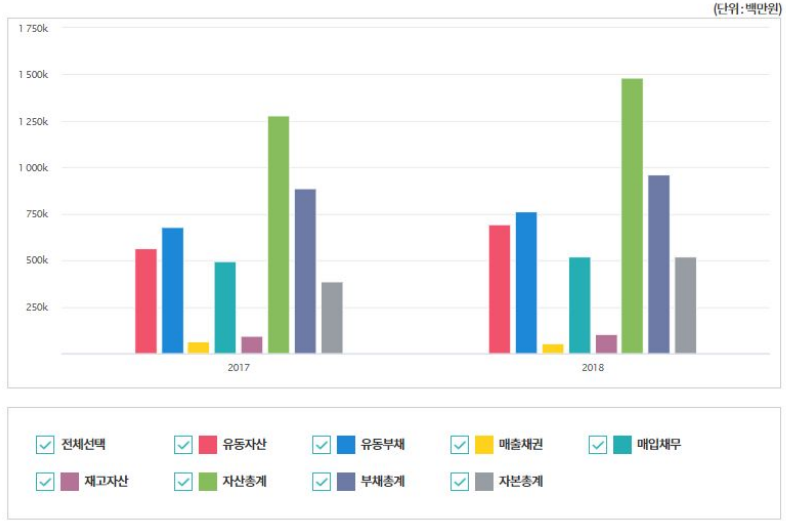
연간 분기

재무상태표 포괄손익계산서 현금흐름표

(단위 : 백만원)

		2020 1Q	2019 4Q	2019 3Q	2019 2Q	2019 1Q
매출액	백만원	1,393,052	5,946,068	1,582,745	1,516,513	1,349,820
영업이익	백만원	18,479	196,623	64,821	60,955	26,339
당기순이익	백만원	12,007	151,377	50,179	45,946	20,953

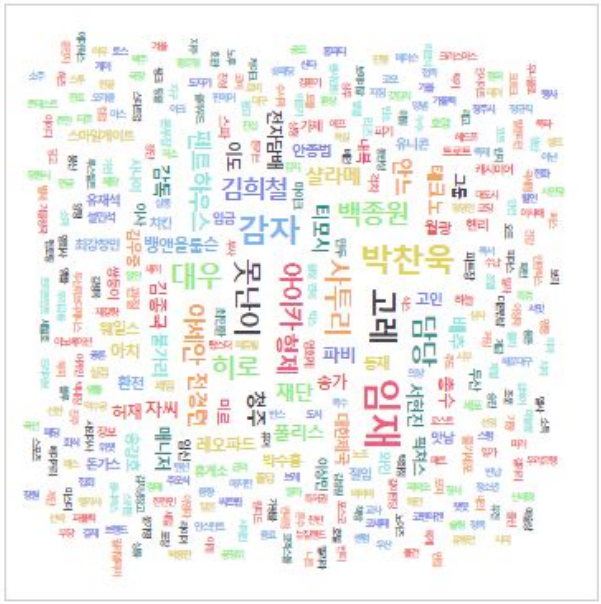
(단위 : 백만원)



최종결과

기업 중요 토픽

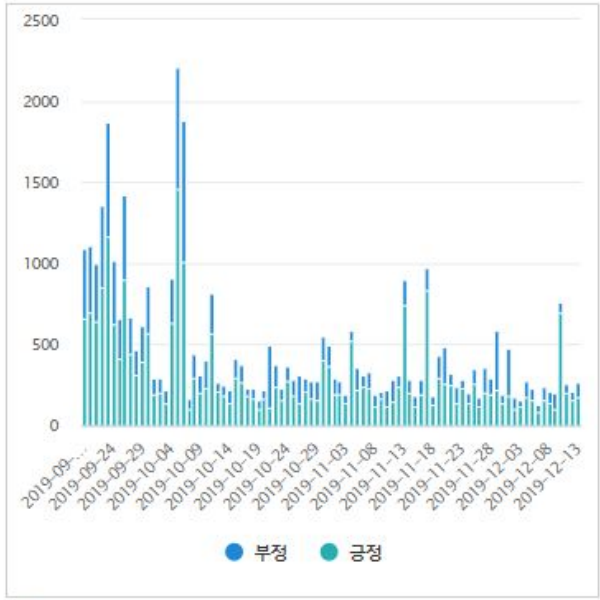
기간 : 2019.09 ~ 12



*출처 : 네이버 뉴스

기업 뉴스 감성분석

기간선택 2019.09.19 ~ 2020.06.05



Bar chart Pie chart

문제발생과 극복

DART API 문제와 각 기업별 회계장부가 상이한 문제가 있음으로 크롤링코드와 API를 적용하기 매우 어려웠다.

특이 케이스를 지닌 기업별로 맞춤형 크롤링 코드를 생성하였으며, 데이터 수집에서 많은 시간이 사용되었다.

Nowmoment company를 통해 기업분석 가능성을 확인 하였고 인터넷 언론사와 협약하여 언론 기사를 작성하였다.

http://cnews.thebigdata.co.kr/view.php?ud=2020032513404724d0a8833aad_23

감사합니다.

이력서:
https://github.com/NinCastle/resume/blob/master/pdf/resume.pdf
연락처: 010-9915-8508