**빅데이터 보고서**

작성자 : 박성준

**각 분야별 작성시 고려사항 (빅데이터)**

* **주제와 관련한 데이터 수집이 적절하게 진행되었는가?**
* **수집된 데이터의 EDA와 전처리 기술이 잘 적용되었는가?**
* **주제와 관련한 데이터 분석 기술이 적절하게 적용되고 유의미한 결과를 도출하였는가?**
* **주제와 연관된 인사이트 도출을 위한 데이터 시각화가 잘 되었는가?**

**빅데이터 주요 업무 및 협력사안**

* Bigdata : 데이터 시각화를 기반으로 의미있는 인사이트 제공
* AI : 데이터 전처리를 통해 AI가 학습할 데이터 지원
* IoT : IoT와 협력하여 보험데이터 구축
* Cloud : 클라우드와 협력하여 모니터링 시각화 제공

**빅데이터 주요 업무 세부사항**

1. 배달데이터 시각화 및 인사이트 도출
2. 안전사고 데이터 전처리 및 시각화를 통한 인사이트 도출
3. 회귀분석을 통한 상관관계 파악 -> 없음
4. 연관분석을 통한 배달 및 안전의 연관성 분석
5. 키워드 분석을 통한 배달 및 안전 연관성 시각화(워드클라우드)
6. STT 인텐트 분석 기본 구조 설계
7. AI 학습을 위한 STT API활용 학습데이터 구축
8. 이상치 감지 구조 설계 및 라즈베리 파이 내 구축
9. WakeWord 성능 향상을 위한 STT 데이터 추출
10. 보험데이터 설계 (멘토링)
11. 보험데이터베이스 구축을 위한 주소 위경도 추출 및 파생변수 설계
12. 보험데이터를 통한 모니터링 페이지 시각화
13. 라이더 등급평가

**데이터 목록**

* **배달데이터(KT 빅데이터 플랫폼) :**

1. 배달상점 데이터 (42769 rows × 11 columns)
2. 시간지역별 배달 소요시간 평균 (375220 rows × 5 columns)
3. 시간 지역별 배달 주문건수 (381068 rows × 5 columns)
4. 시간 지역별 배달 평균주문금액 (374561 rows × 5 columns)
5. 업종 지역별 배달 주문건수 (197192 rows × 5 columns)
6. 업종 지역별 평균배달소요시간 (195986 rows × 5 columns)
7. 업종 지역별 평균주문금액(191306 rows × 5 columns)
8. 주문지역 인구 특성 (157374 rows × 8 columns)

* **연관분석 :** 구글 기사 크롤링 (185개)
* **AI학습데이터 :** 녹음파일 (660개)
* **보험데이터 :** 자이로센서 (자전거 – 6723 rows × 7 columns)

**참고사이트**

1. **텍스트 연관분석**

* <https://needjarvis.tistory.com/59>
* <https://lemontia.tistory.com/903>

1. **군집분석**

* <https://tariat.tistory.com/819>
* <https://m.blog.naver.com/samsjang/221017639342>

1. **이상치 감지**

* <http://www.databaser.net/moniwiki/wiki.php/%EC%9D%B4%EC%83%81%EC%B9%98%EC%A0%9C%EA%B1%B0%EB%B0%A9%EB%B2%95>
* <https://sosomemo.tistory.com/34>

1. **지도시각화**

* <https://blog.naver.com/kcchang61/221350672356>

**느낀 점 및 보완할 점**