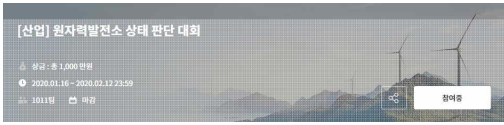
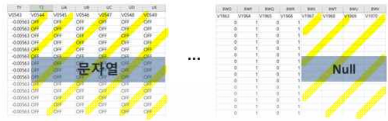


프로젝트 요약서 (1/4)

주관	Dacon
참여	팀원
주제명	원자력발전소 상태 판단 대회

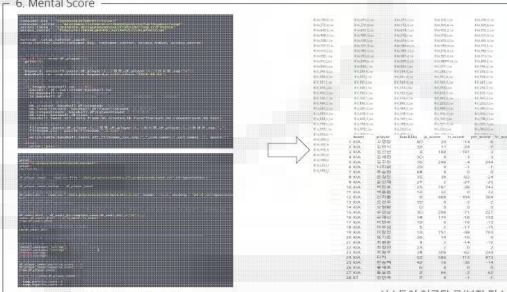
구분	프로젝트 상세내용
프로젝트명	원자력발전소 상태 판단 대회
구현목표	198개의 상태 클래스 판단
구현내용	<div>1. 주제</div> <div>  </div>
	<div>2. 데이터 전처리 (1/2)</div> <div> <p><원본데이터></p>  <p>문자열 (Bad, ON, OFF, CLOSE, Open equip fail, Normal, No Data, System char[], MID POSITION, Priority 3, Configure)과 Null 값에 대해서 다른 데이터와 비교하여 데이터의 특성에 맞게 최빈값 전처리</p> <pre>1 # None 값을 최빈값 최빈값으로 채워줌 2 with tqdm(total=len(train.columns)) as pbar: 3 for col in train.columns: 4 train[col] = train[col].fillna(train[col].mode()[0]) 5 pbar.update(1)</pre> </div>
상세 설명	<div>3. 데이터 전처리 (2/2)</div> <div> <pre>1 real_yver = [] # 정답 변수 2 with tqdm(total=len(train.columns)) as pbar: 3 for i in train: 4 if len(train[i].unique()) != 1: 5 real_yver.append(i) 6 pbar.update(1)</pre> <p>목적: 변별력이 없는 변수를 제거하여 정확도와 속도 개선</p> <p>수행내용: 특정 변수의 데이터가 모두 동일하다면 (원소의 수가 하나라면) 상태판단에 변별력이 없을 것임. 따라서 모든 데이터가 동일한 변수를 제외</p> <p>결과: 5122개의 변수 중 3513개의 변수 선택 train과 test셋에 적용 시켜준 뒤, 재사용을 위해 저장</p> </div>
	<div>4. 모델 구축 및 검증</div> <div> <p><모델구축parameter값></p> <pre>1 # 학습용 데이터를 로드함 2 param = {'train': train, 3 'boosting_type': 'gbdt', 4 'objective': 'multi:softmax', 5 'num_class': 198, 6 'metric': 'multi_logloss', 7 'learning_rate': 0.00235, 8 'num_threads': 7, 9 'num_leaves': 75, 10 'feature_fraction': 0.4, # 특징 샘플링 11 'bagging_fraction': 0.4, # 부트스트랩 12 'bagging_freq': 10} # 데이터의 일부를 무작위</pre> <p>learning_rate, max_depth, num_leaves, fraction 값들을 조정하면서 학습 실행</p> <pre>1 # 학습용 데이터에 대해 1000 epoch씩 학습시킴 2 # validation 데이터에 대해 1000 epoch씩 검증시킴 3 # validation 데이터에 대해 1000 epoch씩 검증시킴 4 # validation 데이터에 대해 1000 epoch씩 검증시킴 5 # validation 데이터에 대해 1000 epoch씩 검증시킴 6 # validation 데이터에 대해 1000 epoch씩 검증시킴 7 # validation 데이터에 대해 1000 epoch씩 검증시킴 8 # validation 데이터에 대해 1000 epoch씩 검증시킴 9 # validation 데이터에 대해 1000 epoch씩 검증시킴 10 # validation 데이터에 대해 1000 epoch씩 검증시킴</pre> <p>1000 epoch씩 학습시킴하며 모델을 저장하고 log_loss값 확인</p> <p>Did not meet early stopping. Best iteration is: (4000) valid_0's multi_logloss: 0.229422</p> <p>Did not meet early stopping. Best iteration is: (6000) valid_0's multi_logloss: 0.213872</p> <p>총 6000번의 학습을 진행하여 모델을 구축</p> </div>
상세 설명	<div>1. 주제 선정/개요</div> <div>1-1) 안전한 원자력발전을 위해 한국수력원자력에서 제공한 모의 운전 및 실제 데이터를 기반으로 하는 AI 알고리즘 개발</div> <div>2. 분석 및 기능 설계</div> <div>2-1) 데이터 수집 : 제공 데이터 사용 및 외부 데이터 사용 불가</div> <div>2-2) 데이터 전처리 - 문자열 처리 : 문자열 (Bad, ON, OFF, CLOSE, Open equip fail, Normal, No Data, System char[], MID POSITION, Priority 3, Configure)과 Null 값에 대해서 다른 데이터와 비교하여 데이터의 특성에 맞게 최빈값 전처리</div> <div>2-3) 데이터 전처리 - 변수 선택 : 변별력이 없는 변수(원소의 수 하나)를 제거하여 정확도와 속도 개선</div> <div>3. 분석 기법</div> <div>3-1) lightGBM : 트리기반의 gradient boosting 알고리즘</div> <div>4. 분석도구(R/파이썬-(머신러닝/딥러닝) 등)</div> <div>파이썬 (pandas, lightGBM(머신러닝))</div>

프로젝트 요약서 (2/4)		
주관	한국경제신문	
참여	팀원	
주제명	국립공원 활성화를 위한 빅데이터 분석 아이디어	

구분	프로젝트 상세내용
프로젝트명 구현목표	국립공원 활성화 를 위한 SNS 마케팅 (스냅촬영 서비스) 데이터 분석 을 통하여 국립공원 활성화를 위한 아이디어 도출
구현내용	<div> <div> 1. 기획배경 <p>인스타그램은 주목도가 관심을 가질 가능성이 높은 콘텐츠를 노출하는 알고리즘이기 때문에 이용자가 게시물을 보면 볼수록, 좋아소를 누를수록, 데이터가 축적되어 그 정확도가 높아진다.</p> <p>오늘 인스타그램을 통한 홍보는 필수가 된 만큼 식당, 카페, 별칭까지 인스타그램 계정이 있다. 실제 방문객을 가늠할 수 있는 한 남성은 인스타그램 계정을 만들고 홍보하면서 승인이 있을 것이다. "이"라는 어떤 방문객에게 인스타그램 감성이라는 하나의 장르가 생겨난 것이 아닐까 싶다" 고 말했다.</p> </div> <div> 2. 데이터 전처리 과정 </div> </div> <div> <div> 3. 모델링 모형 <p>t-SNE 는 시각화를 위한 기계 학습 알고리즘이다.</p> <p>이 비선형 차원 축소는 두세 차원의 저 차원 공간에서 고차원 데이터 원본을 시각화에 적합 기법이다.</p> <p>구체적으로, 유사한 물체가 근처 상에 의해 모여있고 다른 물체가 먼 상에 의해 높은 확률로 모여있는 방식으로 각 고차원 물체를 2 차원 또는 3 차원 점으로 모아놓는다.</p> <p>원래 알고리즘은 객체 간의 유클리드 거리를 유사성 매트릭스의 기준으로 사용한다.</p> </div> <div> 4. 서비스 구현 </div> </div>
상세 설명	<div> <div> 1. 주제 선정/개요 <p>1-1) 선정 배경 : 국립공원에 대한 낮은 인식률과 지속적인 방문객 수의 저하</p> <p>1-2) 솔루션 : 국립공원에 진행 중인 콘텐츠도 있지만, 새로운 참여형 콘텐츠의 필요성</p> </div> <div> 2. 분석 및 기능 설계 <p>2-1) 데이터 수집 및 전처리 : 트위터와 인스타그램에서 22개의 국립공원 및 포토존 관련 데이터 수집</p> <p>2-2) 텍스트 분석 : 연관어 분석, t-SNE, 이미지 태그 분석</p> </div> </div> <div> <div> 3. 서비스 소개 <p>3-1) 기술의 특징점</p> <ul style="list-style-type: none"> 국립공원의 창의적인 마케팅 전략으로 스냅촬영을 실시 국립공원별 포토존 활성화 및 사진작가들만의 이색촬영기법 활용 사진기반 SNS 플랫폼을 통해 국립공원에 대한 인식률을 높이고, 방문객수를 높이고자 함 </div> <div> 4. 활용방안 및 기대효과 <p>일시적인 콘텐츠가 아닌, 지속적으로 진행될 수 있는 참여형 콘텐츠이기 때문에 국립공원에 대한 인식을 개선시켜 방문객 증가를 기대할 수 있다. 또한 여행업 활성화를 통해 서비스 일자리 창출 효과를 기대할 수 있다.</p> </div> </div> <div> <div> 5. 분석도구(R/파이썬-(머신러닝/딥러닝) 등) <p>파이썬 (pandas, konlpy, t-SNE(머신러닝))</p> </div> </div>

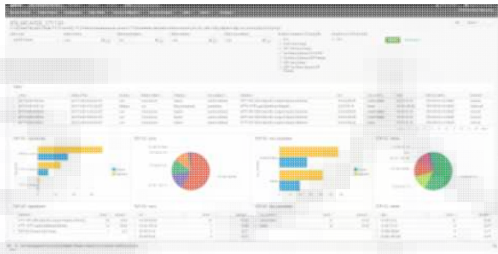
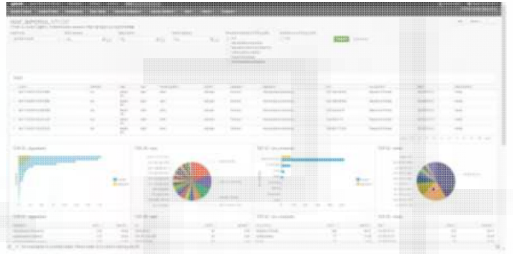
프로젝트 요약서 [2/4]

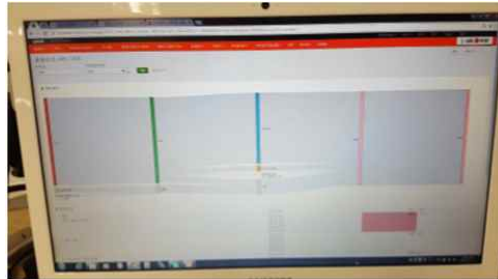
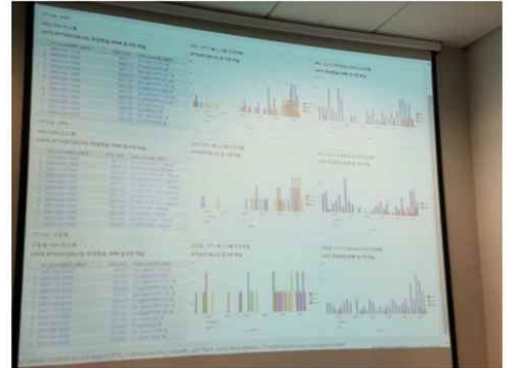
주관	한국경제신문
팀원	팀원
주제명	빅데이터 분석 아이디어

구분	프로젝트 상세내용
프로젝트명	Baseball Score Model (야구 성적 예측 서비스)
구현목표	야구 데이터 분석을 통하여 야구선수의 경기력을 예측하는 지표 생성
구현내용	<div>1. 기획배경</div>  <p>다들 야구라는 스포츠에 대해 이해를 못하고 있어요. 그로 인해 메이저리그 팀을 운영하는 사람들이 선수들을 오만, 빔을 잘못 이르고 있고.</p> <p>팀 운영자들은 선수를 사는 일만 신경쓰죠.</p> <p>중요한건 선수가 아닌 승리를 사는 거예요.</p> <p>승리하려면 독점할 선수를 사야죠</p> <p>- 미니멀 대사중 -</p>
	<div>2. 데이터 전처리 과정</div>  <p>200+ Player files 7 score</p>
구현내용	<div>3. 모델링 모형</div>  <p>선수들이 언급된 트윗을 수집</p> <p>선수들이 언급된 긍/부정 점수</p>
	<div>4. 서비스 구현</div>  <p>7. Result</p> <p>멘탈등급: LEVEL 4 예상 AVG: 0.283 스카우트 고려 사항</p> <p>1. 높은 멘탈 등급 2. 압박 스트레스 컨트롤 가능 3. 환경 변화로 인한 성적 변동이 낮을 것으로 예상</p> <p>스카우트 추천 지수:</p>
상세 설명	<div>1. 주제 선정/개요</div> <p>1-1) 선정 배경 : 프로야구 각 구단은 우승을 목표로 양질의 선수를 구하기 위해 힘쓰고 있으나, 객관적 타격 지표를 분석하여 선수를 선발해도 향후 성적을 예측하기 어려움</p> <p>1-2) 솔루션</p> <ul style="list-style-type: none"> - 정규분포를 따르는 타격지표를 선정하여 예측의 신뢰도를 높임 - 트위터 크롤링 데이터를 활용한 감성지수를 분석하여 선수들의 긍/부정 평가점수 도출 - 타격지표와 긍/부정 평가점수를 결합하여 선수의 경기력을 예측하는 지표(멘탈지수)로 사용 <div>2. 분석 및 기능 설계</div> <p>2-1) 데이터 수집 및 전처리 : KBO 웹사이트에서 경기기록 수집, 트위터에서 선수 언급 데이터 수집</p> <p>2-2) 경기기록 데이터 분석 : EDA, Random Forest, Light GBM</p> <p>2-3) 텍스트 분석 : 긍/부정 사전을 통해 선수별 평가점수 계산</p> <div>3. 서비스 소개</div> <p>3-1) 기술의 특징점 : 과거 공격 성적을 바탕으로 향후 성적을 예측하여 선수 스카우트에 활용</p> <div>4. 활용방안 및 기대효과</div> <p>'멘탈지수'를 새로 개발하여 선수의 경기력에 영향을 미치는 요인을 분석하여, 보다 양질의 선수를 스카우트할 수 있는 가능성을 기대할 수 있음</p> <div>5. 분석도구(R/파이썬-(머신러닝/딥러닝) 등)</div> <p>R, 파이썬 (pandas, Light GBM(머신러닝), Random Forest(머신러닝))</p>

프로젝트 요약서 (3/4)


회 사	가이온
부 서	빅데이터 연구소
주 제 명	빅데이터 분석 프로젝트 수행

구분	프로젝트 상세내용	
프로젝트명	보안로그 수집/가공/분석	
프로젝트 진행일	2017.11. ~ 2018.02.	
주사용 기술	splunk	
역할	고객사 ○○○ 의 보안 로그 분석 대시보드 생성	
구현내용	1. 시각화 (대시보드)	2. 시각화 (대시보드)
		

구분	프로젝트 상세내용	
프로젝트명	○○홈쇼핑 번호분리 프로젝트	
프로젝트 진행일	2017.10. ~ 2017.11.	
주사용 기술	splunk	
역할	매체에 따라 고객에게 노출되는 번호가 분리됨에 따라 고객경험분석 프로젝트의 대시보드 수정 및 추가 개발	
구현내용	1. 시각화 (대시보드)	2. 시각화 - 프로젝트 오픈 모니터링 용도
		

프로젝트 요약서 [3/4]

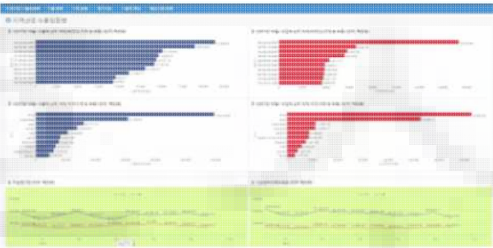
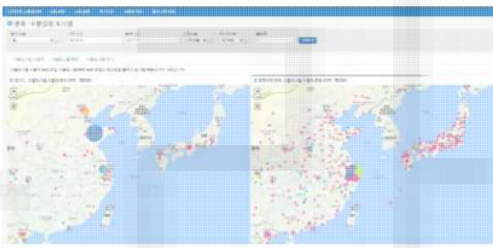
회 사	가이온
부 서	빅데이터 연구소
주 제 명	빅데이터 분석 프로젝트 수행

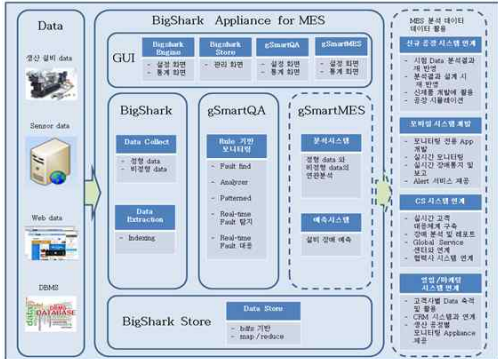
구분	프로젝트 상세내용	
프로젝트명	○○홈쇼핑 상담원 데스크탑 프로젝트	
프로젝트 진행일	2017.09. ~ 2017.10.	
주사용 기술	splunk	
역할	고객이 ○○홈쇼핑에서 겪은 경험들을 일목요연하게 정리하여 고객과 상담원의 의사소통을 도와주는 화면을 구축하는 프로젝트에서, splunk로 데이터를 가공하는 업무를 맡음	
구현내용	1. 웹페이지에 가공한 데이터 표출	
		

구분	프로젝트 상세내용	
프로젝트명	물류차량 차륜/차축 베어링 상태 모니터링 시스템	
프로젝트 진행일	2017.09. ~ 2019.02.	
주사용 기술	splunk, html, javascript	
역할	열차의 wheel, bearing 의 상태를 모니터링하는 splunk 시스템 구축 및 유지보수	
구현내용	1. 시각화 (대시보드)	2. 시각화 (대시보드)
		

프로젝트 요약서 [3/4]

회 사	가이온
부 서	빅데이터 연구소
주 제 명	빅데이터 분석 프로젝트 수행

구분	프로젝트 상세내용	
프로젝트명	smarttrade (무역통계데이터 분석)	
프로젝트 진행일	2016.04. ~2019.02.	
주사용 기술	splunk, java(servlet), oracle	
역할	무역통계 데이터 수집/가공/분석/시각화	
구현내용	1. 시각화 (대시보드)	2. 시각화 (대시보드)
		

구분	프로젝트 상세내용	
프로젝트명	MES 연계형 실시간 분석/장애 탐지/예측이 가능한 빅데이터 Appliance 개발	
프로젝트 진행일	2014.12. ~ 2016.03.	
주사용 기술	java, spring framework, spark, nutch, hdfs, html, css, javascript, jsp, mysql	
역할	MES 시스템의 각 설비에서 발생하는 데이터를 수집/가공/분석/시각화	
구현내용	1. 구성도	2. ML predict 인터페이스 구현
		<pre>public interface WebService { public void setRequest(HttpServletRequest request); // spark1 - machine learning algorithm public ResultVO SVM(String hdfs, String inputFile, int poercentage, int iteration, String modelSaveTime); public ResultVO dt(String hdfs, String inputFile, int poercentage, String modelSaveTime); public ResultVO rb(String hdfs, String inputFile, int poercentage, String modelSaveTime); public ResultVO lr(String hdfs, String inputFile, int poercentage, String modelSaveTime); public String kmeans(String hdfs, String inputFile, int pCluster, int iteration); public String SVM(String hdfs, String inputFile); public String mes_classification(String hdfs, String inputFile, int poercentage, int iteration, String modelSaveTime); public String mes_predict(String hdfs); public String preProcessing(String mCode,String[] inputFile); } // spark1 - machine learning algorithm - predict public List<Tuple2<Object, Object>> SVMpredict(String hdfs, String inputFile, String modelPath, String outputFile); public List<Tuple2<Object, Object>> dtpredict(String hdfs, String inputFile, String modelPath, String outputFile); public List<Tuple2<Double, Double>> rbpredict(String hdfs, String inputFile, String modelPath, String outputFile); public List<Tuple2<Object, Object>> lrpredict(String hdfs, String inputFile, String modelPath, String outputFile);</pre>

프로젝트 요약서 (4/4)


학	교	성균관대학교	
연	구	실	정보 및 지능 시스템 연구실
주	제	명	소셜미디어 텍스트마이닝

논문명	Competitive Self-Training Technique for Sentiment Analysis in Mass Social Media
등재기관 (발행처)	Korean Institute of Intelligent Systems
작성일자	2014.12.3.
논문내용	To analyze user's emotion automatically by analyzing Twitter using "data without sentiment labels", not only "data with sentiment labels", to increase accuracy of sentiment analysis through an improved Self-Training, one of semi-supervised learning.

논문명	대용량 소셜 미디어 감성분석을 위한 반감독 학습 기법
등재기관 (발행처)	한국지능시스템학회
작성일자	2014.10.1.
논문내용	본 논문에서는 "감성 레이블이 있는 데이터"와 함께 "감성 레이블이 없는 데이터"도 활용하기 위해서 반감독 학습기법인 self-training 알고리즘을 적용하여 감성분석 모델을 생성한다.
추가사항	KCI 등재

논문명	대용량 소셜 미디어 감성분석을 위한 반감독 학습 기법
등재기관 (발행처)	한국지능시스템학회
작성일자	2014.4.18.
논문내용	본 논문에서는 기계학습 기법으로 감성분석 모델을 생성하기 위해서 각각의 트윗에 긍정 또는 부정의 레이블이 필요한데, 레이블이 없는 데이터를 사용하여 감성분석 모델을 생성하는 방법을 제안한다.
추가사항	춘계학술대회 우수논문상 수여

논문명	트위터에서 유용한 정보를 추출하는 방법
등재기관 (발행처)	한국지능시스템학회
작성일자	2012.11.9.
논문내용	본 논문에서는 트윗에서 의미있는 정보를 추출하기 위하여 개인적인 일상이나 느낌 등의 정보의 밀도가 낮은 트윗을 제거하므로 트윗의 정보밀도를 높이는 작업을 제안한다.

번호	성명 거주지 나이	생년월일 (만 나이)	출신학교 (전공)	졸업연도	빅데이터 기술능력
1	 홍소라 경기/수원시	1988.1.21. (32)	성균관 대학교 전기전자 컴퓨터공 학	2015	<p>· 활용 기술</p> <ol style="list-style-type: none"> 1) Python 2) SQL (oracle, mysql) 3) R 4) Java 5) splunk 6) spark 7) html, javascript, css 8) jsp 9) spring framework 10) nutch 11) hdfs <p>· 프로젝트 경험</p> <ol style="list-style-type: none"> 1) 원자력 발전소 상태 판단 대회 (2위) 2) 국립공원 활성화를 위한 빅데이터 분석 아이디어 공모전 (우수상 수여) 3) 야구예측모델 4) 보안로그 수집/가공/분석 5) ○○홈쇼핑 번호분리 프로젝트 6) ○○홈쇼핑 상담원 데스크탑 프로젝트 7) 물류차량 차륜/차축 베어링 상태 모니터링 시스템 8) 무역통계데이터 분석 (smartrade) 9) MES 연계형 실시간 분석/장애 탐지/예측이 가능한 빅데이터 Appliance 개발 10) 트위터 감성분석 (우수논문상 수여) <p>· 자격증 (컴퓨터 관련)</p> <ol style="list-style-type: none"> 1) 데이터분석준전문가 2) 정보처리기사 3) OCJP (SCJP) <p>· 산업체 실습/부서</p> <p>와이즈넷 / Data사업부 개발팀</p> <p>· 관련 교육 이수내역</p> <ol style="list-style-type: none"> 1) 한국경제신문/혁신성장청년인재양성과정 실무프로젝트 기반 빅데이터 전략 마에스트로과정 (960시간) 2) Java CBD Developer (500시간)