2017년 한이음 ICT멘토링
프로젝트 결과보고서

프로젝트명	R을 활용한 기업 성장성 분석

요 약 본

프로젝트 정보	
주제영역	□ 건강 ■ 생산성 □ 생활 □ 안전 □ 엔터테인먼트
기술분야	□ 이동통신 □ 방송스마트미디어 □ 기반SW·컴퓨팅 □ 디지털콘텐츠 □ 융합서비스 □ 네트워크 □ 전파·위성 ■ SW □ 정보보호 □ ICT 디바이스
달성성과	■ 논문게재 및 포스터발표 □ 앱등록 □ 프로그램등록 □ 특허 □ 기술이전 □ 실용화 □ 공모전(*공모전명*) □ 기타( )
프로젝트명	R을 활용한 기업 성장성 분석
프로젝트 소개	기업의 성장성과 가치를 매출, 매출원가,영업이익율 등의 정형데이터와 경제, 경영관련 뉴스 등 비정형 데이터를 기반으로 다양한 알고리즘을 활용해 분석하고, 각 알고리즘이 출력하는 그 결과의 유의성을 검증한다. 오픈소스 도구인 R을 활용해 정형데이터 분석에는 주성분회귀분석, 인공신경망 기법을 사용하고 비정형데이터 분석에는 나이브 베이지안 분류자, 긍/부정사전분석 모델을 사용해 분석한다. 분석된 결과를 검토하여 각 분석모델별 성능을 확인하고, 기업 성장성 예측을 위해 가장 적합한 모델과 필요한 데이터를 제시한다.
개발배경 및 필요성	기업의 성장성은 해당 기업의 이해관계자뿐만 아니라 잠재적 투자자, 일반 대중에 이르기까지 다양한 계층에게 매우 큰 관심사항이라고 할 수 있다. 이렇게 지대한 관심을 가지고 있음에도 불구하고 증권사나 펀드사와 같은 기업체가 아닌 경우, 다양한 기업관련 데이터를 분석하고 해당 기업의 성장성에 대해 의사결정을 내리는 것은 일반 투자자에게는 매우 어려운 일이다. 그러나 최근에는 빅데이터 분석, 공공데이터 개방 트렌드 의 확산과 각종 표준 API 의 배포 등으로 막대한 양의 기업 관련 데이터를 활용할 수 있어 적절한 분석기법과 모델이 제공된다면 누구나 정확한 기업 데이터를 기반으로 신뢰성 높은 의사결정을 진행할 수 있는 시대가 되었다. 데이터 분석을 위한 도구 측면에서도 오픈소스 기반의 데이터 분석 도구인 R 등이 대중화되면서 간단한 스크립트 언어인 파이썬(Python)을 기반으로 여러 가지 복잡한 통계 분석을 비교적 쉽게 활용할 수 있다.
프로젝트 주요기능	오픈 소스인 R과 파이썬을 활용하여 웹 크롤링을 진행하였다. 크롤링을 통해 얻은 재무제표 데이터와 주가 데이터를 활용해 정형 데이터 분석을 진행하였고, 거기에 더해 뉴스데이터를 활용하여 비정형 데이터 분석을 추가적으로 시행하였다. 결론적으로 정형 데이터 분석과 함께 비정형 데이터 분석을 시행하여 다양한 방면으로 기업의 성장성을 분석해보았다.
작품의 기대효과 및 활용분야	1)인사이트 도출역량 - 전혀 관계가 없어 보이는 데이터들 간의 상관관계를 분석하여 인사이트를 도출 2)프로세스 개선 - 기업 프로세스 내 문제가 있는 부분의 정확한 진단 및 개선포인트 도출 3)분석도구 활용역량 - R, 파이썬 등 빅데이터 분석도구의 활용역량 강화

(본문) 프로젝트 결과보고서

Ⅰ. 프로젝트 개요

가. 프로젝트 소개

ㅇ기업의 성장성과 가치를 매출, 매출원가, 영업이익율 등의 정형데이터와 경제, 경영관련 뉴스 등 비정형 데이터를 기반으로 다양한 알고리즘을 활용해 분석하고, 각 알고리즘이 출력하는 그 결과의 유의성을 검증한다. 오픈소스 도구인 R을 활용해 정형데이터 분석에는 주성분회귀분석, 인공신경망 기법을 사용하고 비정형데이터 분석에는 나이브 베이지안 분류자, 긍/부정사전분석 모델을 사용해 분석한다. 분석된 결과를 검토하여 각 분석모델별 성능을 확인하고, 기업 성장성 예측을 위해 가장 적합한 모델과 필요한 데이터를 제시한다.

나. 개발배경 및 필요성

ㅇ기업의 성장성은 해당 기업의 이해관계자뿐만 아니라 잠재적 투자자, 일반 대중에 이르기까지 다양한 계층에게 매우 큰 관심사항이라고 할 수 있다. 이렇게 지대한 관심을 가지고 있음에도 불구하고 증권사나 펀드사와 같은 기업체가 아닌 경우, 다양한 기업관련 데이터를 분석하고 해당 기업의 성장성에 대해 의사결정을 내리는 것은 일반 투자자에게는 매우 어려운 일이다. 그러나 최근에는 빅데이터 분석, 공공데이터 개방 트렌드 의 확산과 각종 표준 API 의 배포 등으로 막대한 양의 기업 관련 데이터를 활용할 수 있어 적절한 분석기법과 모델이 제공된다면 누구나 정확한 기업 데이터를 기반으로 신뢰성 높은 의사결정을 진행할 수 있는 시대가 되었다. 데이터 분석을 위한 도구 측면에서도 오픈소스 기반의 데이터 분석 도구인 R 등이 대중화되면서 간단한 스크립트 언어인 파이썬(Python)을 기반으로 여러 가지 복잡한 통계 분석을 비교적 쉽게 활용할 수 있다.

다. 작품 구성도

Step1 : 정형데이터 분석 - 주성분 회귀분석(PrincipleComponentAnalysis)

- 5년간 분기별 재무제표 데이터를 웹 크롤링(Crawling)을 통해 수집
- 각 변수별 값에 대한 결측치 처리, 스케일링 등 전처리 수행
- 영업이익률, ROA, ROE, 자산총계, 부채총계 등을 주성분으로 범주화
- 컬럼명 변환, 단위 통일 등 분석을 위한 표준화(Nomalization)
- 주성분 회귀분석을 활용하여 모델 적합

Step2 : 정형데이터 분석 - 인공신경망(ArtficialNeuralNetwork)

- 7년간 일별 주가 데이터를 웹 크롤링(Crawling)을 통해 수집
- 약 1000개의 학습데이터를 훈련집합(TrainingSet)과 검증집합(TestSet)으로 구분
- 훈련집합에의 함수 처리를 통하여 인공신경망 모델 학습
- predict를 호출하여 주가 예측 결과 도출

Step3 : 비정형데이터 분석 - 나이브 베이지안 분류자(NaïveBayesianClassifier)

- 1년간 '00홈쇼핑' 뉴스기사를 웹 크롤링(Crawling)을 통해 수집
- 각 변수별 값에 대한 전처리 수행
- 뉴스기사를 명사 단위로 잘라 10번이상 나온 단어들을 이용해 희소행렬로 변환
- 나이브 베이지안 분류자를 활용하여 모델 적합

Step4 : 비정형데이터 분석-긍/부정 사전(Positive/NegativeSentimentDictionary)

- 영어 긍부정 사전을 번역한 한글 긍부정 사전을 이용하여 희소행렬로 변환
- 파생변수 생성 - 기사별 긍정단어개수, 부정단어개수
- 긍정단어 개수가 많은 기사를 긍정기사라고 분류
- 긍정기사가 많이 나온 날은 주가가 오를 것이라 가설을 세우고 모델 적합

라. 작품의 특징 및 장점

ㅇ 기업을 성장성을 분석하는 연구에 정형 데이터 분석 뿐만 아니라 뉴스 데이터를 활용한 비정형 데이터 분석을 진행하였다는 점에서 다른 연구와는 다른 차별점이 있다. 또한, 오픈소스인 파이썬과 R을 사용하여 분석에 활용하였다는 점에서 특징이 있다.

Ⅱ. 프로젝트 수행결과

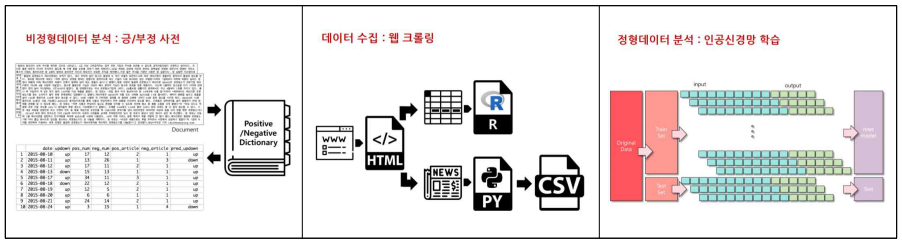
가. 주요기능

구분	기능	설명
S/W	주성분 회귀분석	재무제표 데이터를 활용한 정형 데이터 분석
	인공신경망	주가 데이터를 활용한 정형 데이터 분석
	나이브 베이지안 분류자	뉴스를 활용한 비정형 데이터 분석
	긍/부정 사전	뉴스를 활용하여 가설을 토대로 모델 적합

나. 프로젝트 개발환경

구분		항목	비고
S/W 개발환경	OS	XOS / Windows	
	개발환경(IDE)	MAC / Windows	
	개발도구	R studio, Jupyter notebook	
	개발언어	R, Python	

다. 결과물 상세 이미지



바. 달성성과

■ 논문게재 및 포스터발표	게재(발표)자명	논문(포스터)명	게재(발표)처	게재(발표)일자
	류지승	R과 분석 알고리즘을 활용한 기업의 성장성 예측에 관한 연구	한국정보처리학회	2017. 12. 03.

Ⅲ. 프로젝트 수행방법

가. 업무분담

번호	성명	역할	담당업무
1	강희석	멘 토	- 인사이트 제공, 전체적인 프로젝트 관리 (일정 관리)
2	류지승	팀 장	- 웹크롤링, 비정형 데이터 분석, 포스터 발표, 보고서
3	김민정	팀 원2	- 비정형 데이터 분석
4	김경수	팀 원3	- 정형 데이터 분석
5	이가연	팀 원4	- 정형 데이터 분석

나. 프로젝트 수행일정

구분	추진내용	수행일정								
		3월	4월	5월	6월	7월	8월	9월	10월	11월
계획	R을 활용한 기업 성장성 분석 연구를 위한 브레인 스토밍	●	●	●						
크롤링	파이썬을 활용하여 뉴스를 크롤링하고, R을 활용하여 재무제표와 주가 데이터를 크롤링				●	●	●			
분석	정형 데이터 분석						●	●	●	
	비정형 데이터 분석						●	●	●	
논문 작성	한국정보처리학회에 논문 투고								●	
포스터	한국정보처리학회, 한이음 발표를 위한 포스터 작성									●
오프라인 미팅	프로젝트 회의		●	●			●	●	●	

Ⅳ. 기대효과 및 활용분야

ㅇ 본 연구에서는 기업의 성장성을 예측하기 위해 오픈 소스 통계분석 도구인 R을 활용해 여러 가지 분석 기법들을 실제 적용해 보면서 그 성능을 확인해 보았다. 정형데이터를 기반으로 일반적인 회귀분석을 수행했을 경우,유의한 결론을 얻기 어렵다는 점을 실제 수행을 통해 확인하였으며 인공신경망을 기반으로 데이터를 학습시키고 분석하는 기법이 좀 더 나은 의사결정을 지원할 수 있음을 확인하였다. 또한 여기에 추가로 비정형 데이터인 뉴스의 긍/부정 뉘앙스를 분석하여 의사결정 시에 참고한다면 보다 안정적인 수준의 결론에 도달할 수 있음을 확인할 수 있었다. 향후에는 인공신경망의 정확성을 더욱 보강하기 위하여 보다 많은 훈련 데이터(TrainingSet)가 필요할 것으로 판단되며 훈련시에 과적합(Overfit)이 발생하지 않도록 유의해야 할 것으로 판단된다. 또한 비정형데이터는 보다 다양한 반응을 분석하기 위해 SNS의 반응에 대한 강도(Strength)분석이 추가로 필요할 것으로 보이며 문장과 전체적인 맥락을 인식하기 위한 연구를 차후 진행할 계획이다.

Ⅴ. 참고자료

[1] 범주형 자료분석 개론. Alan Agresti(2009.08)
[2] 빅데이터 분석을 위한 데이터마이닝 방법론. 강현철, 한상태, 최종후, 이성건, 김은석(2014.03)
[3] Probability and Statistical Inference, Global Edition. RRobert Hogg , Elliot Tanis , Dale Zimmerman(2014.12)
[4] 통계학. 류근관(2013.02)
[5] Statistics 4th Edition. David Freedman(2015.08)
[6] Python Programming: An Introduction to Computer Science. John Zelle(2014.09)
[7] 긍부전 사전 github 자료 : https://github.com

한이음 ICT멘토링 프로젝트 산출물

1. **한국정보처리학회 논문**

(R과 분석 알고리즘을 활용한 기업의 성장성 예측에 관한 연구) ·····

2. **포스터** (R과 분석 알고리즘을 활용한 기업의 성장성 예측에 관한 연구) ·····

3. **python, R 소스코드** ·····