**[프로젝트 보고서]**

|  |  |
| --- | --- |
| **프로젝트 명** | [오늘의 영화] 영어 컨텐츠 영어 난이도 분석 서비스 |
| **팀 명** | 오늘의 고민은? |
| **기간** | 2018/11/04 ~ 2018/11/09 |

|  |  |  |  |
| --- | --- | --- | --- |
| 팀원 | | | |
| NO. | 성명 | 연락처 | E-mail |
| 1 | 이영준 | 010-9402-1277 | dudwns411@naver.com |
| 2 | 차혜령 | 010-5042-7956 | csl7956@naver.com |
| 3 | 심재나 | 010-4108-4737 | shim.jaena@gmail.com |
| 4 | 최승권 | 010-3185-3303 | kach7100@gmail.com |

**1.팀 소개**

오늘의 고민은? 조 :

다들 고민이 많은 사람들로 고민만 하지 말고 조이름으로 정하자는 뜻에서 정하게 되었습니다.

**역할**

전체 : 데이터(영화 스크립트) 수집

이영준(조장): 웹 크롤링(인기 영화 제목, 장르)/ R, Python - Oracle DB 연동/ 쿼리문 작성/ 데이터(영화 스크립트) 분석/ 회귀 분석/ 회의록, 보고서 작성/ 전체 코드 수정

차혜령: 데이터(영어 단어) 수집 / R, Python - Oracle DB 연동/ 쿼리문 작성/ 데이터(영화 스크립트) 분석/ 데이터 시각화(R)/ R로 웹 생성

최승권: 주제 선정/ 전략 수립/ 데이터(영어 단어) 수집 / 데이터 정제(단어 난이도 분류, 이상데이터 처리, 자연어처리)/ 보고서 작성

심재나: 주제 선정/ 전략 수립/ 데이터 정제(영화 스크립트 자연어 처리) / 웹서버 구축/ 회의록, 보고서 작성/ 자료 정리

**2. 프로젝트 개요**

* **주제**

좀 더 쉬운 영어학습을 위한 난이도별 영어 콘텐츠 (영화) 추천

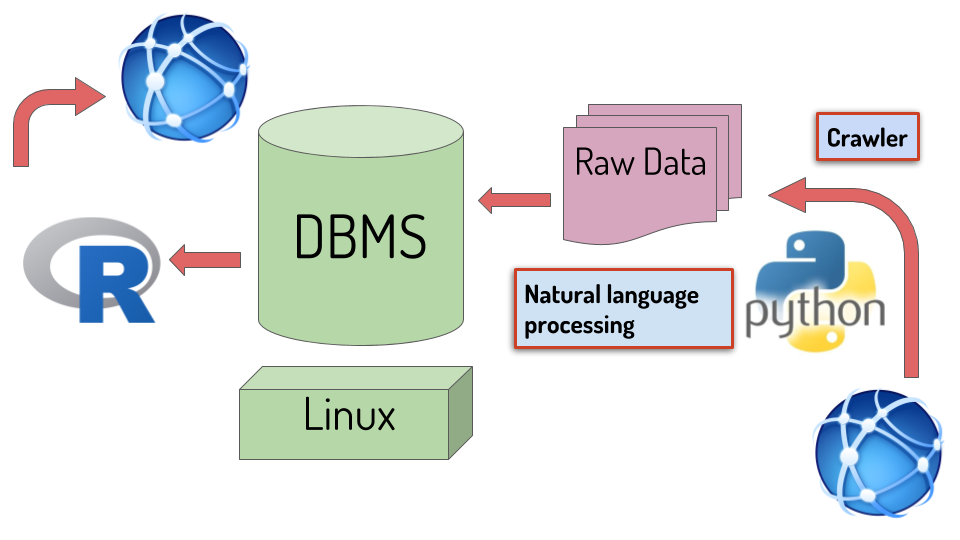
* **이유**

이전보다 토익, 토플 점수보다는 토스, 오픽이 중요해지고 있는 만큼, 단순히 책으로 단어를 공부하는 것 보다 쉽고 재밌게 회화실력을 키우기 위하여 영화, 드라마 등의 콘텐츠 시청이 늘고 있다.

하지만 자신의 수준에 맞는 프로그램을 선택하기가 어렵기에 이를 추천하기 위한 서비스를 구상하였다.

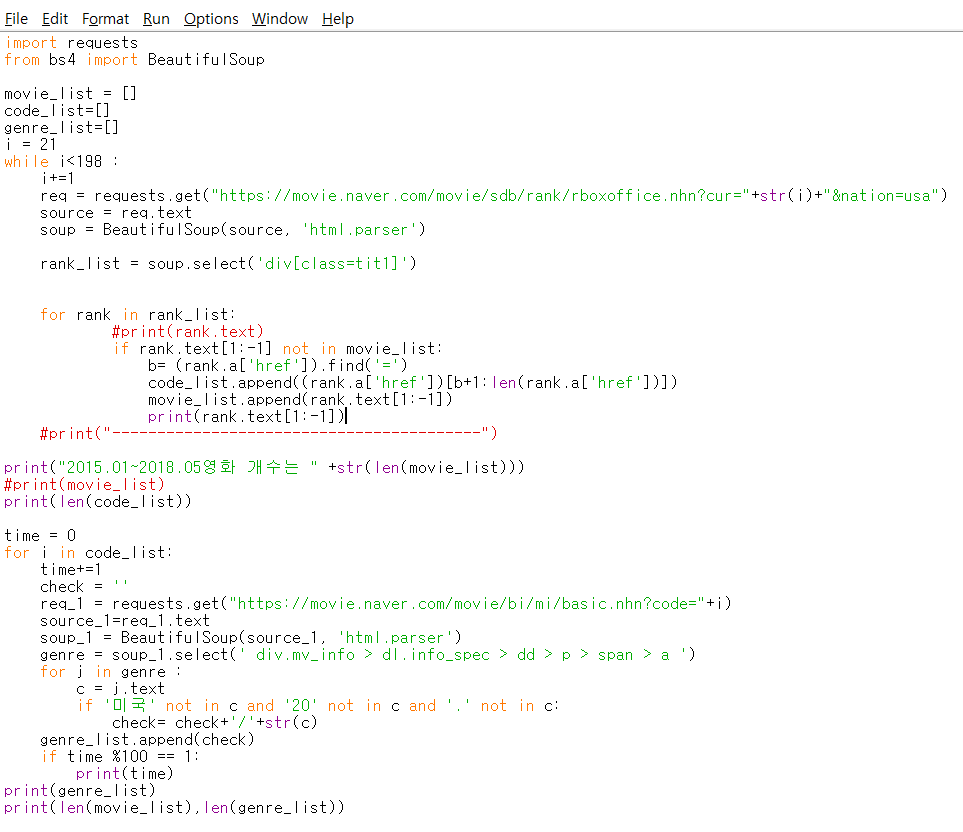
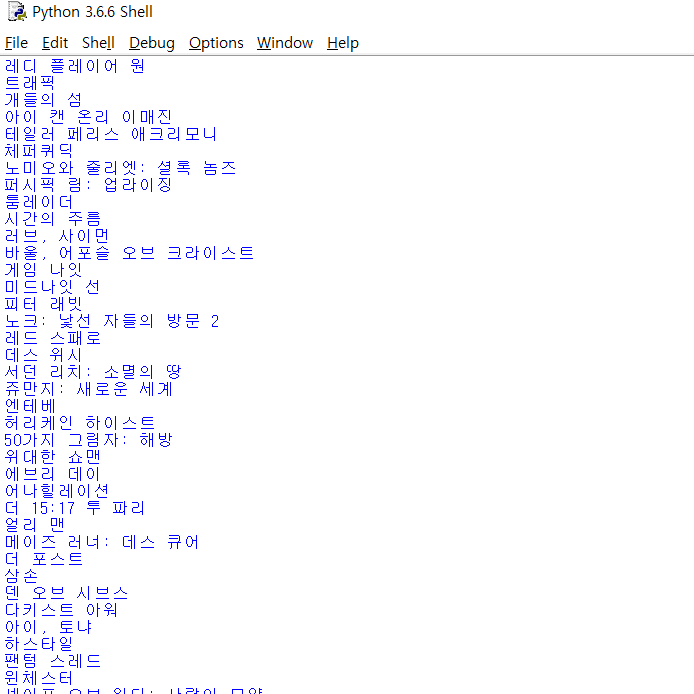
**3.개발 내용**

**-시스템구성**

****

**- 사용기술**

- 네이버 영화에서 2015년부터 2018년 상반기까지 박스오피스에서 흥행한 영화들을 Python 으로 크롤링(requests, BeautifulSoup)



- 그 목록에 영화들의 자막들을 Python의 자연어처리 패키지로 동사원형 형태로 변환.

(nltk : PorterStemmer, Lancaster Stemmer, WordNetLemmatizer)

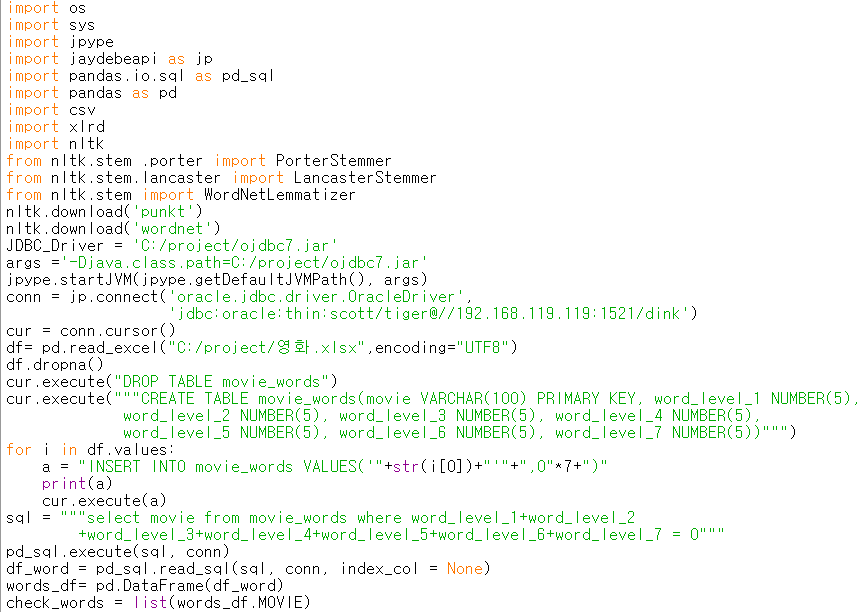


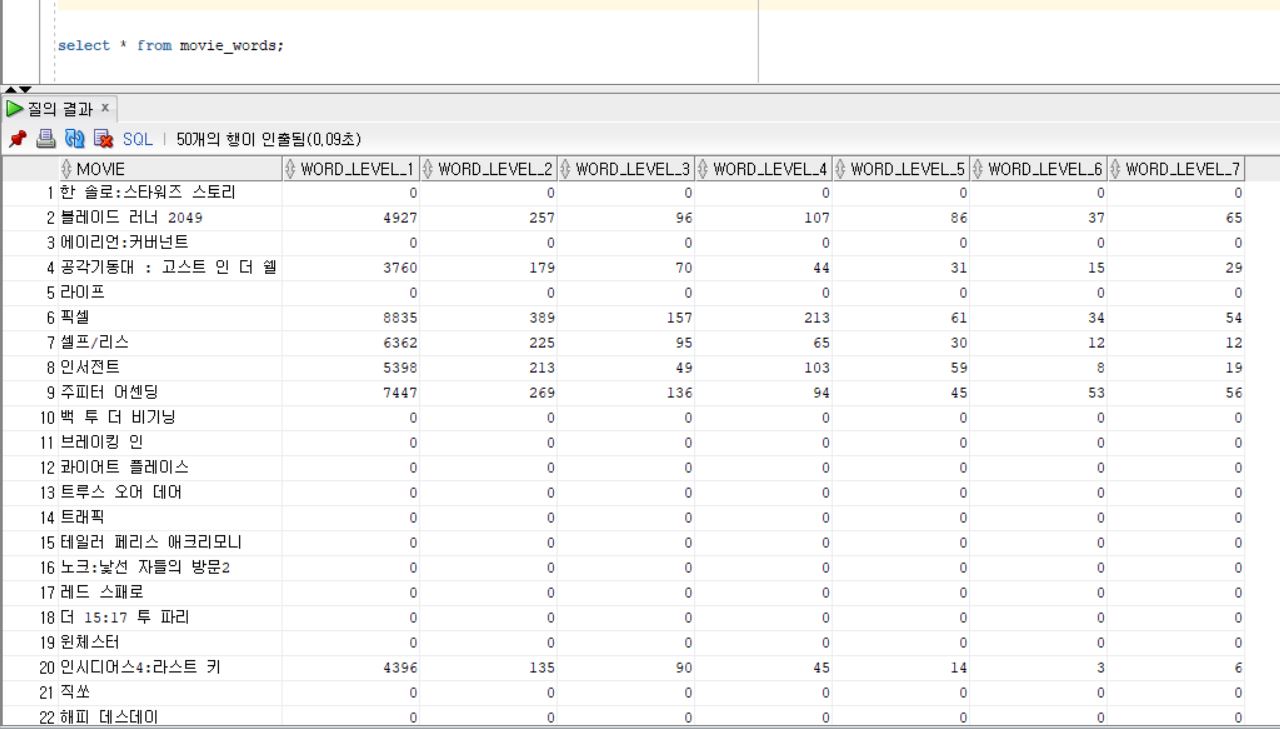




* 전처리 과정중 하나인 자연어처리에서 추가적으로 중복값 제거와
* she's 등의 단어가 ['she', 's'] 로 리스트에 두가지 단어로 처리가 되는 등의 오류를 개선
* It she he, we're 등 유효단어로 카운트하기에 부적절한 워드는 절삭처리

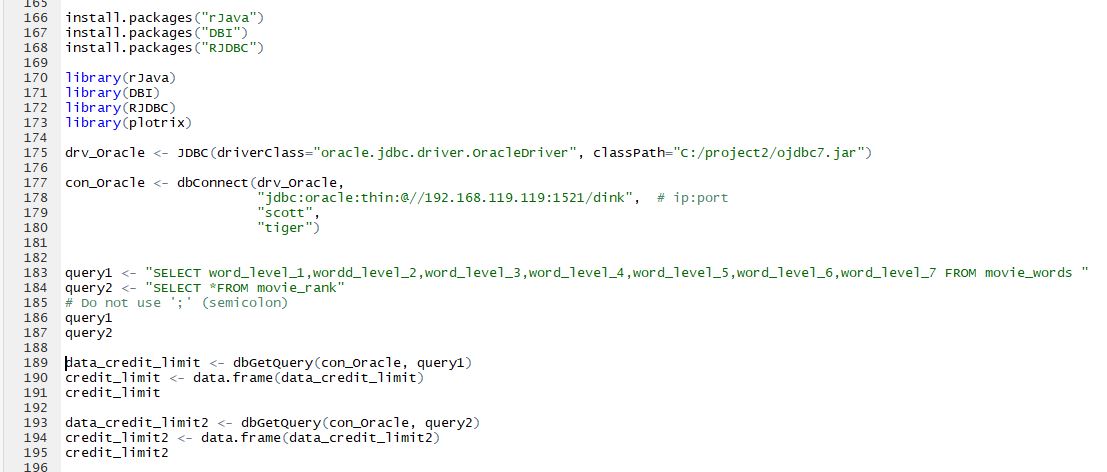
- 파이썬과 Oracle 연동, 단어와 자막을 분석 후(7단계 단어 수준) 영화를 난이도 5단계로 나눔.

-> 데이터 베이스에 저장(jpype, jaydebeapi)

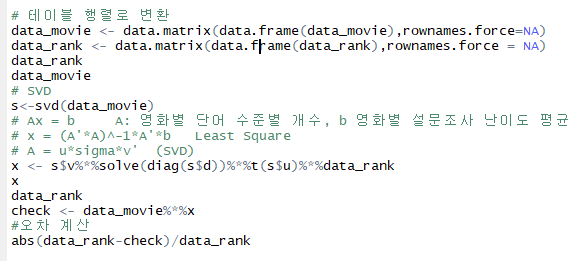


\* R과 Oracle 연동 및 분석, 시각화

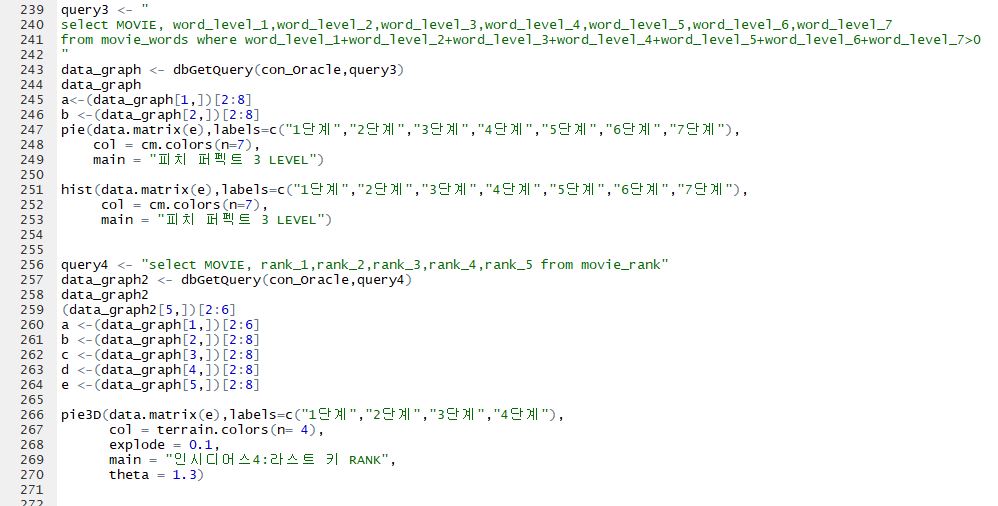
- 데이터베이스의 데이터를 활용하여 최소자승법(Least Square using Singular Value Decomposition) 으로 예측 모델 모델링

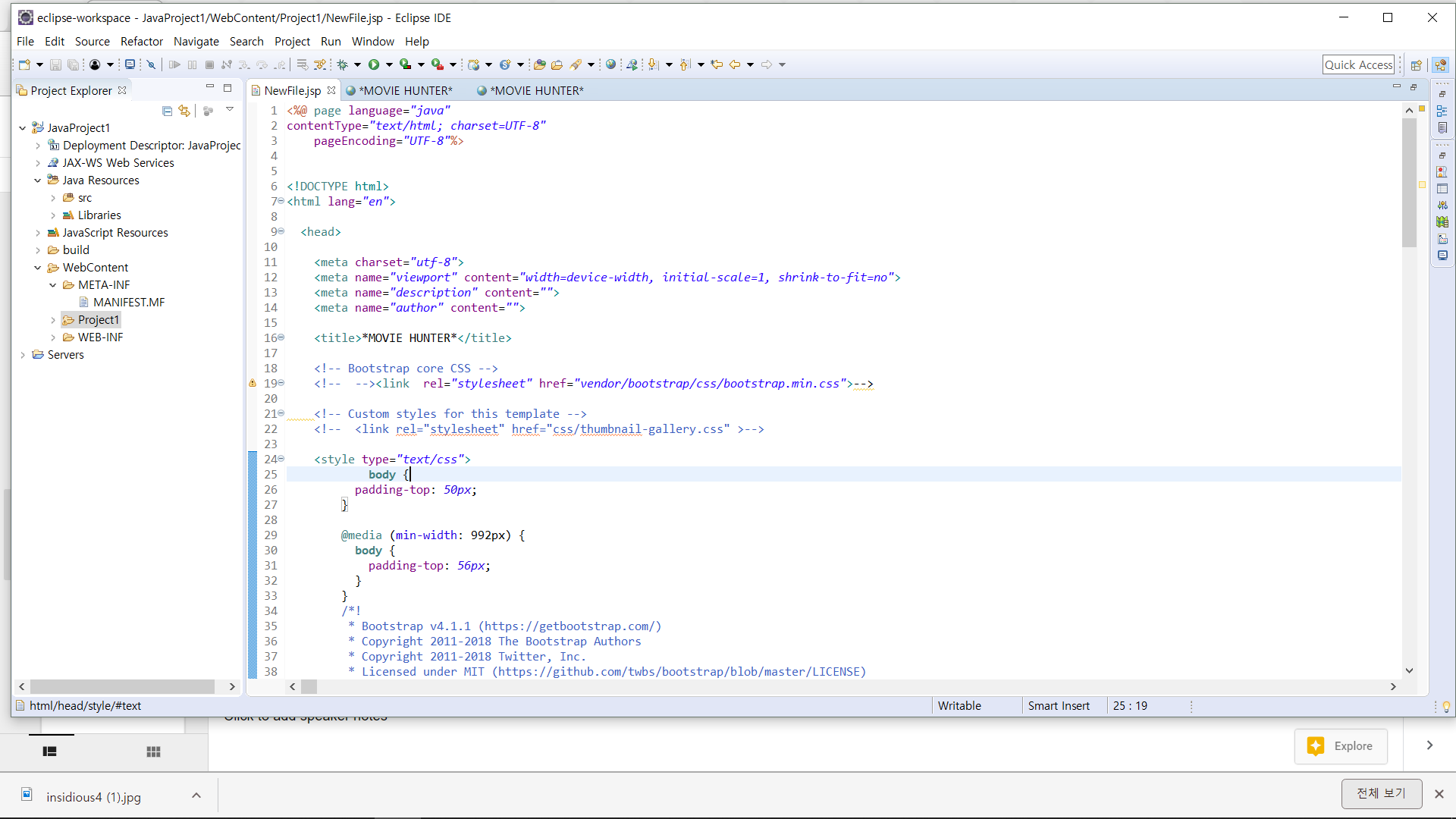


Least Square

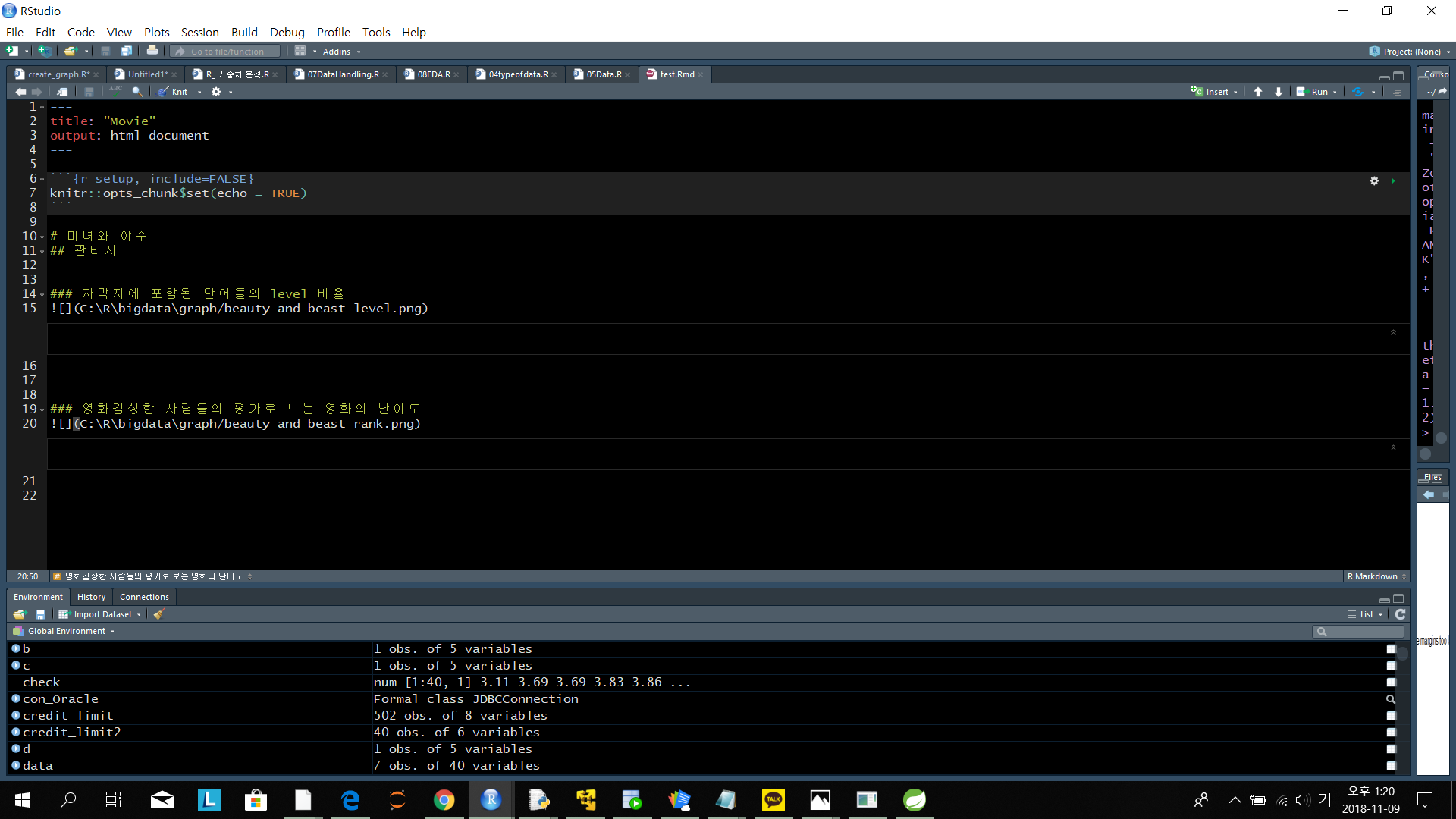


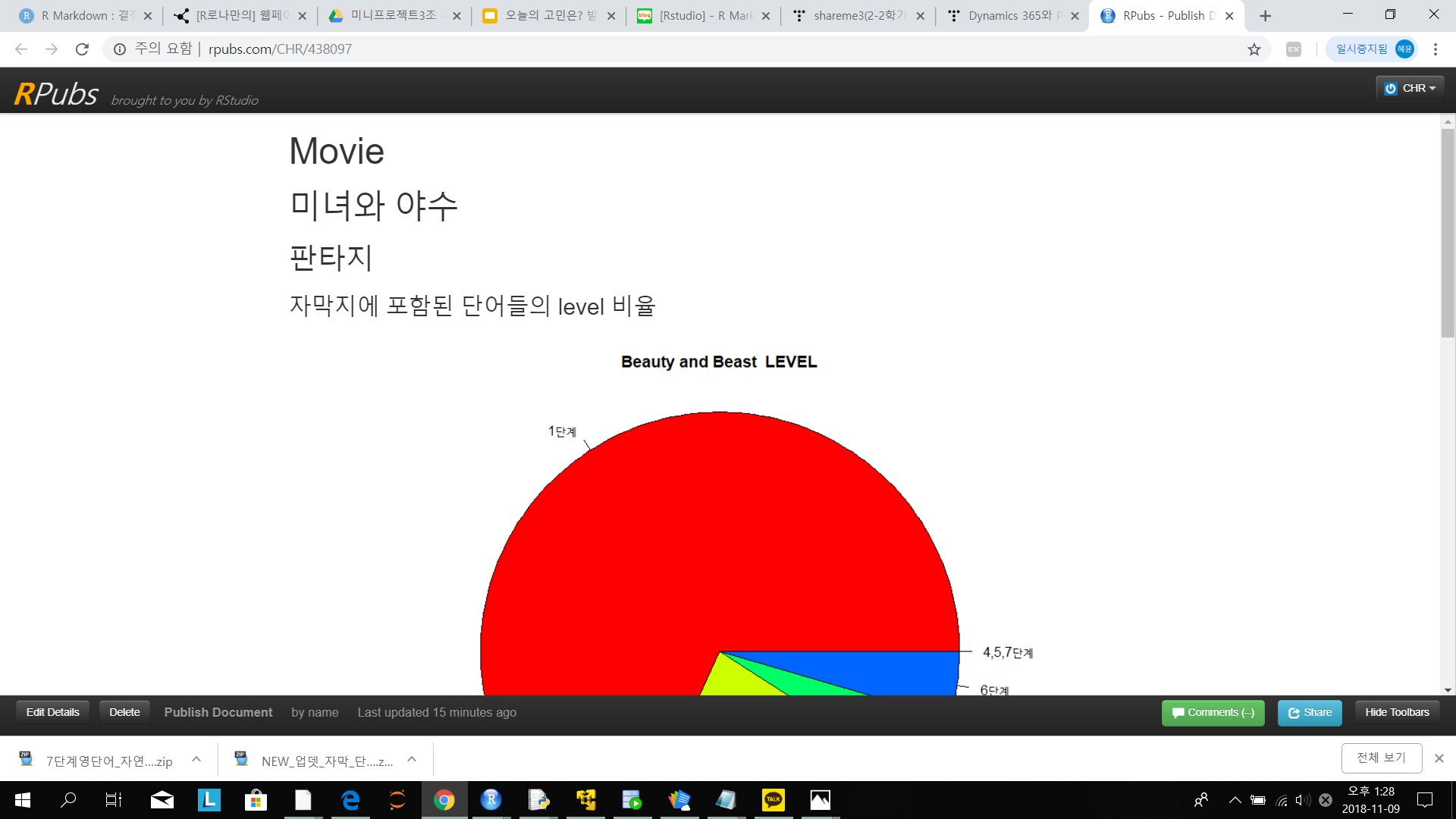
시각화



- 웹서버 구축(JDK 8(WEB) + Tomcat 8(WAS) + Java EE Eclipse / HTML + CSS)

- R로 웹 페이지 생성

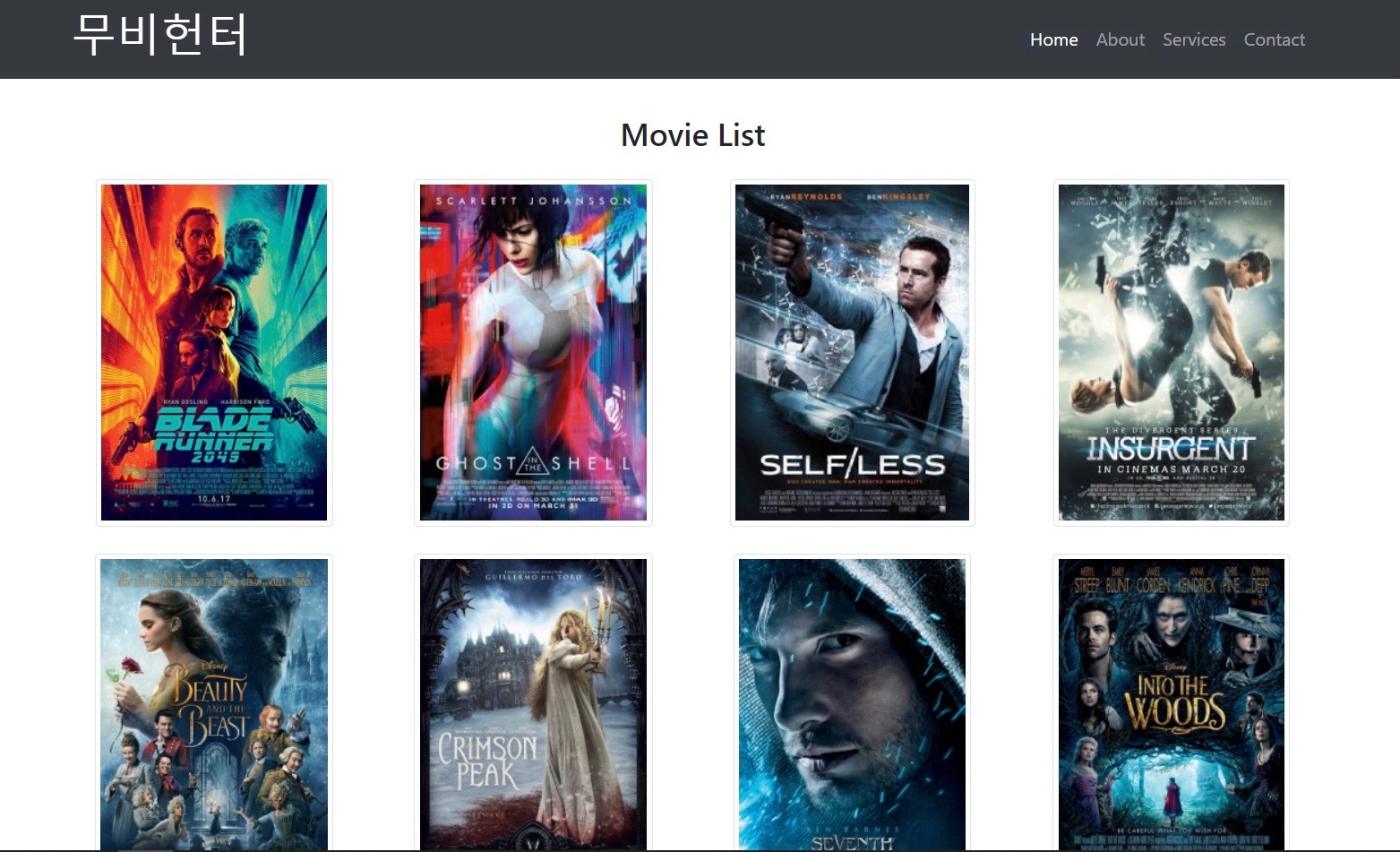


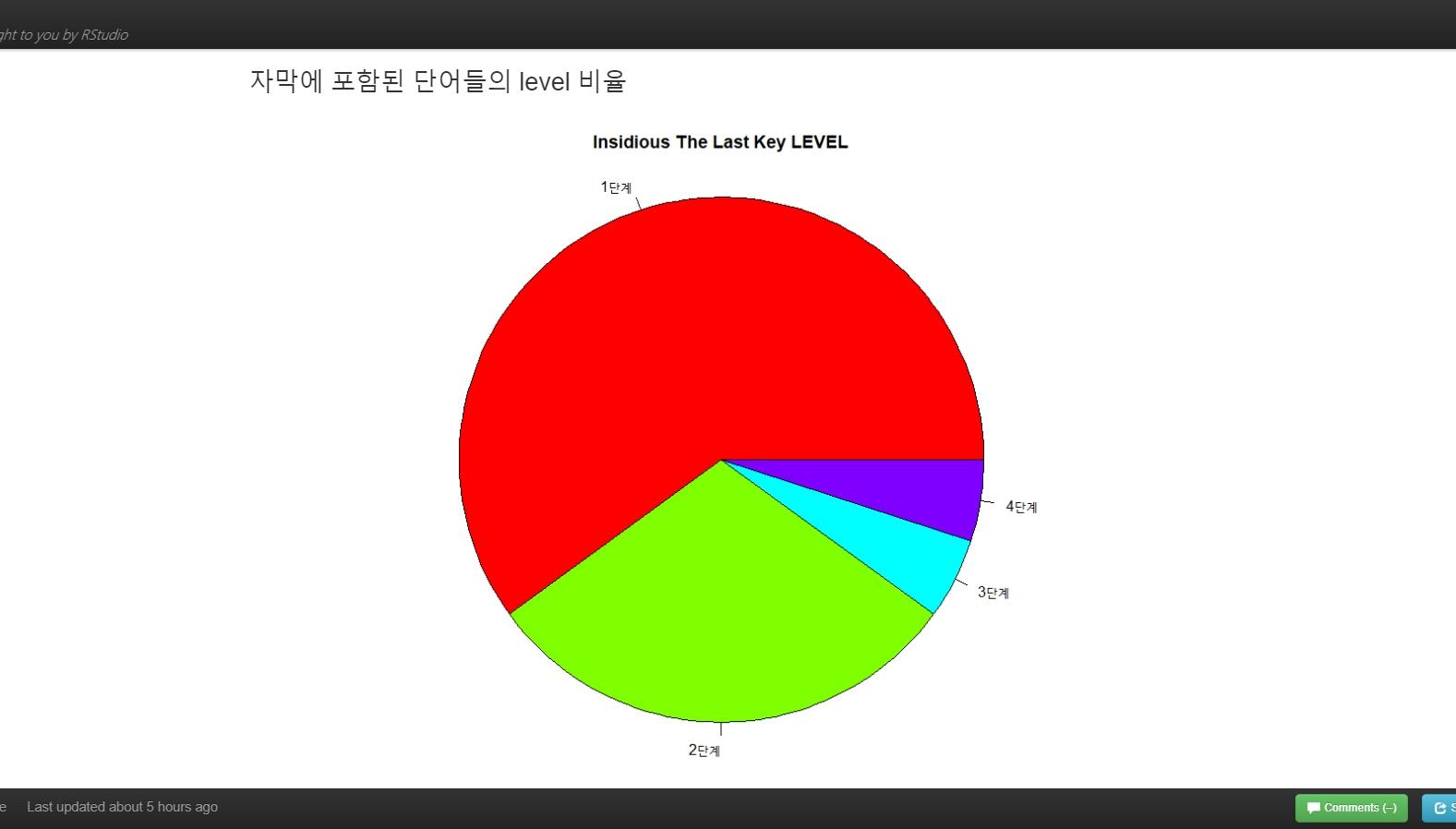


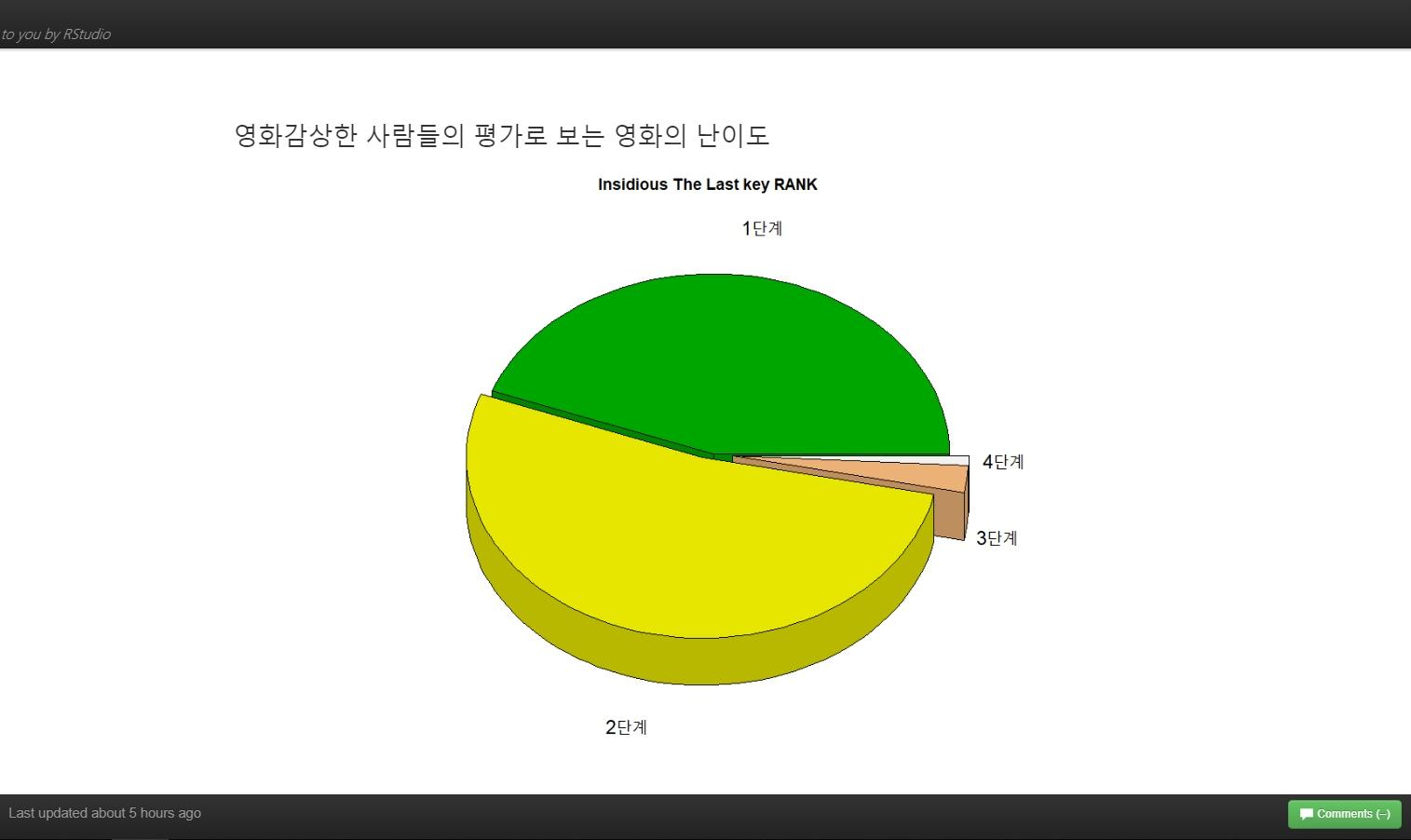
* **데이터 설계**

|  |  |
| --- | --- |
| 기준 데이터 | 영화에 자주 쓰이는 단어들을 7단계로 나눈 Database 테이블 |
| 보조 데이터 | 설문조사한 데이터(임의) |
| 결과 데이터 | 시각화 및 웹 서비스 |

**4. 결과물**

****

****

****

* 사용기술

|  |  |  |  |
| --- | --- | --- | --- |
| **구분** | **이름** | **버전** | **목적** |
| 사용툴 | Eclipse Java EE IDE for Web Developers |  | 자바환경개발, 웹서버 구축 |
| Python | 3.6.6 | 데이터 전처리 |
| RStudio | 3.5.1 | 분석 및 시각화, 웹 연결 |
| sqldeveloper |  | 오라클 데이터베이스 개발 환경 |
| 서버통신 | Apache Tomcat | 8.0 | 웹서버 connector |
| 파이썬 라이브러리 | BeautifulSoup |  | Html 파싱 |
| requests |  | Html 문서 가져오기 |
| nltk |  | 자연어처리 패키지 |
| PorterStemmer, Lancaster Stemmer | 어간추출, 형태소분석 |
| WordNetLemmatizer | 음소표기법 |
| jpype |  | Java class library 접근 |
| jaydebeapi |  | Java JDBC 접근 |
| pandas.io.sql |  | 쿼리문 실행 후 DataFrame으로 변환 |
| os |  | 운영체제 제공기능 수행 |
| sys |  | 인터프리터와 관련된 정보와 기능을 제공 |
| pandas |  | 데이터 분석 지원 |
| csv |  | Csv 파일 읽기 쓰기 |
| xlrd |  | 엑셀에서 데이터 추출 |
| R 라이브러리 | rJava |  | Java 접근 |
| DBI |  | 데이터베이스 인터페이스 |
| RJDBC |  | JDBC 접근 |
|
|
|
|
|

5. 문제점 및 개선사항 개선사항

**문제점**

큰 오차율

[원인]

설문조사 난이도 랜덤설정

난이도 예측 변수 부족

영어단어 난이도 분류

부정확한 예측 모델

**개선사항**

영화 스크립트 추가

영어단어 수준의 분류 개선

예측 정확성을 높이기 위한 모델 개선

영어 컨텐츠 난이도 설정을 위한 설문조사 진행

변수 추가(억양 및 대화속도 등)

웹서버 개선

시각화 추가 및 개선