**작품 요약서**

* **Basic Data**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **작품명** | **Doctor. AI** | | | **지원분야** | Software |
| **개발인원** | 1 명 | **본인역할** | 소 면 준 / 팀장 | **개발기간** | 2013.07  ~2013.08 |
|  |  |  |  |  |  |
| **개발환경** | Window 7 | | | | |
| **개발언어** | JAVA / JSP | | | | |
| **개발툴** | 이클립스 / 톰켓 | | | | |
| **작품소개**  **(요약)** | 자신의 병증상을 검색하면 예상 병명을 찾아주면서 해당 병을 조사해 주는프로그램 입니다. 주말 혹은 심야 시간, 병원 방문이 어려운 상황에서 간단한 질병 같은 경우는 해당 어플리케이션으로 병명을 추측, 간단한 민간요법을 시행할 수 있습니다. | | | | |
| **작품내용** | 1. 서비스 사용자가 질문 (ex)머리가 아파요,허리가 아파요) 을 검색 2. 질문을 네이버 지식인에 “의사답변” 필터링으로 검색한 후 해당 답변들을 전부 형태소 분석기에 명사형 추출 3. 추출된 명사 리스트를 DiseaseList 에 있는 병명에 메치시켜 병명 리스트를 작성 4. 작성 된 병명 리스트의 병명을 웹 서비스에서 제공되는 병 데이터를 추출 5. 웹 페이지에 병 리스트, 해당 병 클릭 시 병 상세 데이터를 출력 | | | | |
| **개발내용**  **(본인 구현부분)** | 1. JSP/HTML 출력을 위해 form 형식 데이터 넘기기,table 작성 및 JSP 프로그래밍을 사용한 search, main 페이지 구성 2. 형태소 분석에서는 꼬꼬마, 한나눔, lucene 3가지 형태소 분석기중 명사 추출기능만 사용하므로 가장 가벼운 lucene 을 사용 하였으며 추출한 명사 리스트를 DB를 사용하지 않고 Doctor AI 가 가지고 있는 diseaseList 에 메치 시켜서 병명 리스트를 만들었습니다. 3. 병명 판단 방법에서는 diseaseList 에 가나다 순으로 폴더를 만든 다음 각 폴더에는 폴더 명과 같은 초성을 가진 단어의 txt 파일을 만들고 해당 txt 파일 안에는 파일 명으로 시작하는 병명 리스트를 담고 있습니다. 명사가 병명인지 아닌지 판별하기 위해서 명사의 초성을 추출 후 가나다 폴더 초성과 메치 후 메치시에는 폴더 안에 txt 파일의 중성을 메치 또다시 메치 된 다면 해당 txt 파일을 open 후에 명사와 txt 파일안의 병명리스트 중에 명사가 있다면 해당 명사를 병명으로 판단 하였습니다.   해당 구조는 26 index를 가지는 Trie(트라이)구조와 매우 유사합니다.   1. 네이버 OPEN API 를 사용하여 지식인 검색을 할려 했으나 데이터 신뢰도를 올리기 위해서 ‘전문가 답변’ 만을 가져 오려고 시도 하였으니 OPEN API 필터링에는 지식인 이라는 대분류 밖에 지원이 되지 않습니다. 그리하여 일반 웹 페이지에서 get 방식으로 지식인 검색시 추가 필터링으로 전문가 검색이 가능하여 속도가 좀 느리더라도 일반적인 get 방식 검색을 사용 하여 지식인 답변을 추출 하였습니다 | | | | |
| **기타** | 프로젝트 개발 당시에는 자연어처리 SIG활동과 공부를 목적으로 진행한 프로젝트로 공모전 출작은 생각하지 않았었지만 자연어 처리 나 정보처리 시스템 같은 경진 대회에 출작 가능 했으며 따로 공모전에 나가지는 않았습니다  해당 프로젝트를 개발 하고 나니 특정 질문에 대한 지식인 최다 빈도수 명사 추출 프로그램이 되어 병명에 대한 검색이 아니라 diseaseList를 다른 목적의 리스트 (ex) 휴대폰 리스트, 소프트웨어 리스트) 로 변경 하여 사용하면 Doctor. Ai 가 아니라 다른 정보 수집 소프트웨어로도 사용 가능 하다는 장점이 있습니다. | | | | |

* **System Architecture**

|  |
| --- |
| **System  Architecture** |
| **그림 1** |
| JSP WEB 페이지에서 검색어 창에 질문자가 질문을 작성 작성된 질문은 네이버 지식인에 던져 지고 해당 지식인 답변 중에 “의사답변”이 붙은 지식인의 답변만 추출 추출된 답변은 형태소 분석기를 통해서 명사만 추출, 이 명사들을 질병 데이터를 통해 병명판독 후 리스트에 저장 후 중복 병명이 발생시 빈도수 +1, 이런 방법으로 답변에서 발견된 병명 리스트를 추출 추출된 병명들의 데이터를 네이버 건강에서 병에 대한 데이터를 가져와 JSP 클라이언트에 병명 리스트 및 데이터를 보여줍니다. |

* **Software/Hardware Architecture**
* **Function Explanation (Applicant)**

|  |
| --- |
| *검색방식 비교* |
| **그림 6** |
| **지식인 검색 방법**  초기 프로젝트 계획에는 네이버 open-api 를 이용하여 빠른 지식인 검색을 사용 하려 했으나 병명을 탐색하려는 만큼 의사 답변에서 병명이 나올 확률이 높으며 질문답변의 신뢰도를 올릴 수 있도록 의사 답변을 추출하기로 계획했습니다 하지만 open-api 에서는 세부적인 검색 필터링은 지원하지 않고 단순히 지식인 검색만 지원하여 해당 답변이 전문의 답변인지  일반인 답변인지 찾아내기가 힘들어서 일반적으로 URL 에 조건을 달아서 검색하는 GET방식을 선택. 호출된 HTML 문서에서 “alt=의사답변” 을 탐색하여서 해당 링크의 지식인의 답변을 가져오도록 하였습니다. 하지만 api를 사용하지 않은 만큼 답변을 가져오는데 질문 1건당 약 0.5 초 정도 시간이 소요된다는 점의 문제 점이 발생하였습니다 하지만 첫 프로잭트 작품인 만큼 속도면 보다는 결과물 정확도에 큰 비중을 주며 네이버에서는 초당 페이지 호출제한이 걸려있지만 지식인 검색시 의사답변은 많아야 5~60 개정도의 답변이 검색되므로 페이지 호출제한에 걸리지 않으므로 GET 방식 검색을 선택하였습니다. |
| **DiseaseList** |
| **그림 7** |
| **병명 판명 방법**  형태소 분석에서 나온 명사가 병명인지 판명 방법으로는 DiseaseList 파일 안에 가,나,다..순의 폴더 안에 각 병명의 첫 단어로 만들어진 txt 파일을 생성 해당 txt 파일에 파일명으로 시작하는 병명 리스트를 작성 합니다 (해당 병명들은 네이버 health 에서 제공하는 질병의료 서비스에서 병명을 수작업으로 만들었습니다)  판명해야 할 명사의 첫 번째 단어를 초성 중성 종성으로 분류 한 뒤 해당 초성과 같은 가나다 순 폴더로 들어간 뒤 txt 파일 명의 중성과 명사에서 추출한 중성과 같다면 해당 txt 파일을 OPEN 후 명사와 파일 안의 데이터를 하나 씩 비교해가면서 질병을 탐색 해 갑니다 여기서 명사와 병명이 메치 된다면 해당 명사는 병명으로 판명합니다. |
| **건강** |
| **그림 8** |
| **병명 데이터 크롤링**  병명에 대한 데이터를 크롤링 하는데 있어서 OPEN-API에서는 건강 서비스가 베타 서비스 라서 제공이 안되는 것 같습니다 그래서 병명 데이터를 크롤링 하는 작업도 지식인에서 가져오는 작업 과 같이 GET 방식으로 질병/의료 정보에서 각기병을 검색후 “의학정보” 에 있는 각기병 URL 가져와서 진료과 데이터를 추출후 각기병 URL에 있는 정의,원인,증상 등의 질병 데이터를 크롤링 합니다.  해당 작업 중 질병마다 정의,원인,증상 등 질병 데이터의 정보가 각각 달라서 “맨위로” 버튼을 기준으로 각 목차 별 데이터 카운트를 체크 후 해당 질병 정보를 가져옵니다. |
| **형태소 분석** |
| **그림 9** |
| **형태소 분석기**  크롤링한 의사 답변들을 형태소 분석 후 결과물을 명사만 추출할 수 있도록 코드를 수정 하여 결과물을 ArrayList에 저장 합니다 |