2020년도

소프트웨어 설계 및 실험

프로젝트 제안서

11 조
201404163 정찬휘
201624615 한연수
201624522 우승빈

목차

1. 프로젝트 목표

사진으로 찍은 A4 1 장 분량의 영문 문서의 주요 문장을 3~4 줄로 요약하기

2. 개발배경 및 서비스 필요성

읽을 텍스트의 양은 많은데 시간은 부족하다. 특히 영어텍스트의 경우 한글텍스트보다 읽는데 더 많은 시간이 소모된다. 따라서 이 시간을 줄여주기 위해 OCR 기능이나 Textrank 와 같은 알고리즘, AWS Comprehend 와 같은 기능을 이용해 먼저 요약해줄 필요가 느껴짐

3. 주제 및 기능

A. 주제 : 사진처리, 자연어처리

B. 기능

1. OCR 처리
2. 문장 및 단어추출
3. 요약된 문장 열람
4. 예전에 요약한 원본텍스트/요약문 열람

4. 구현 내용 및 방법

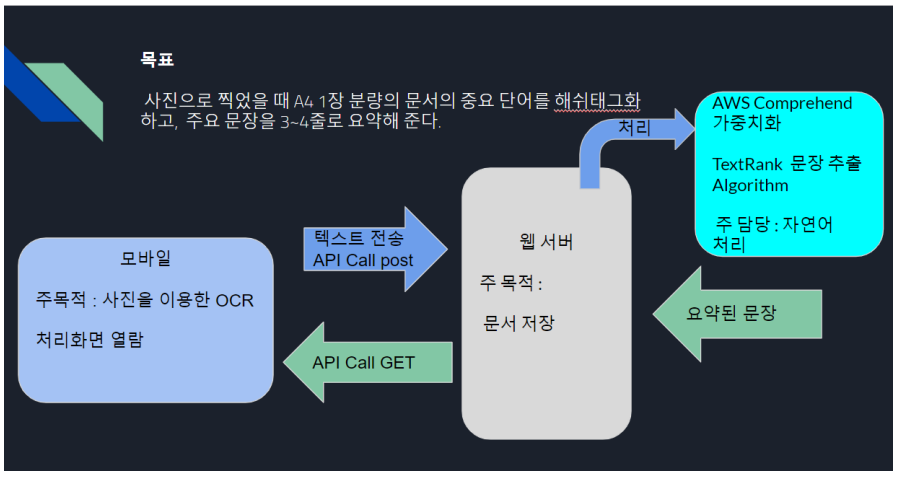
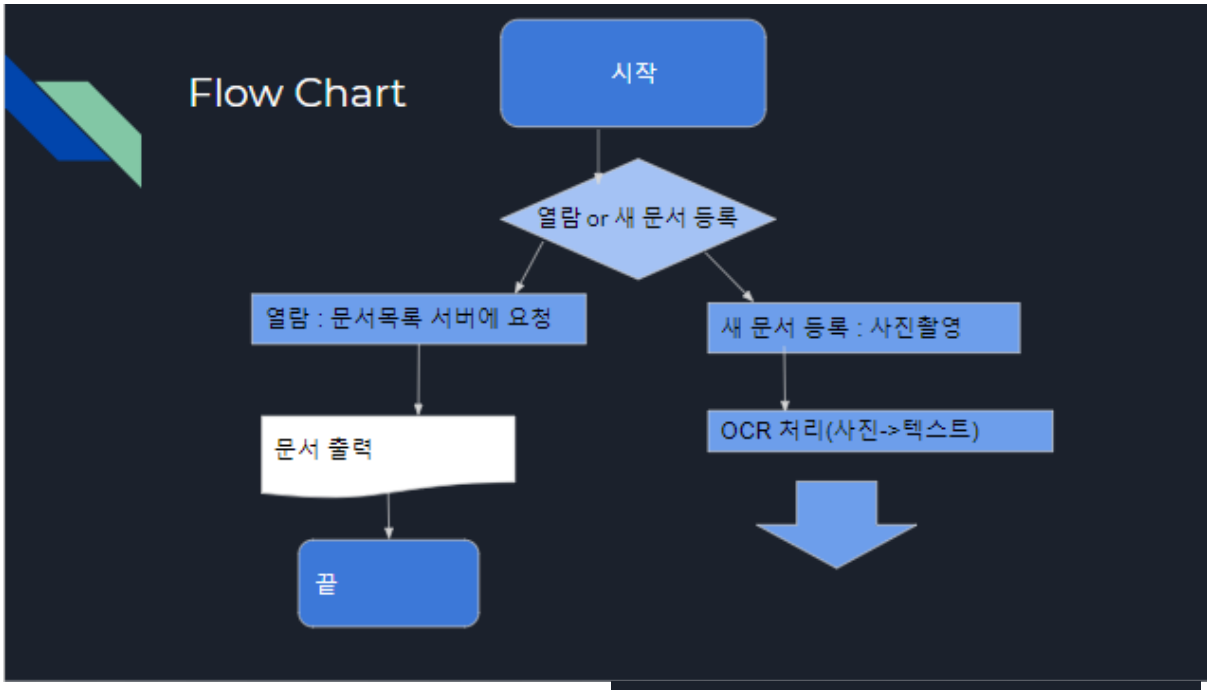


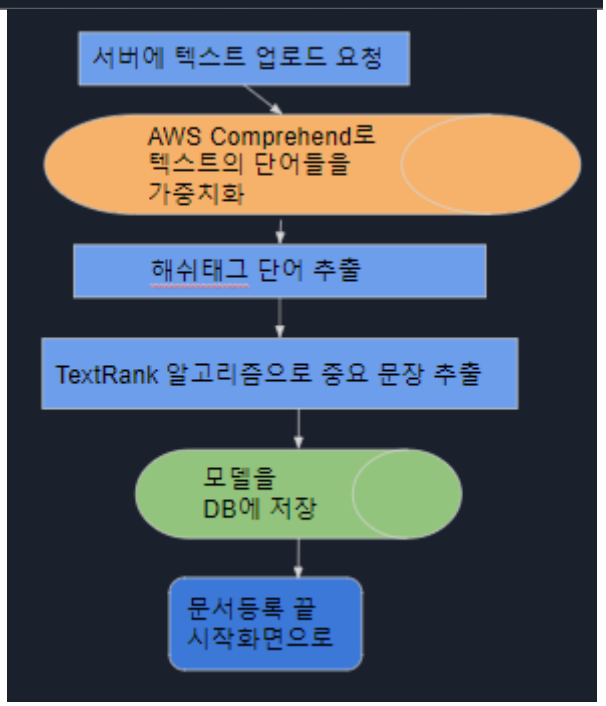
그림 1) 전반적인 구상도

4 - 0. 전반적 설명

안드로이드 단에서 사진을 찍어 OCR 로 텍스트를 추출해 서버로 보내면 서버는 이를 AWS 와 Textrank 로 처리하여 문장과 단어를 추출해 문장은 각각 요약문과 색인을 위한 해쉬태그로 사용하여 이를 열람할 수 있게 제공한다.

-Flow chart





4 – 1. Android Application

- 사진을 찍어 사진을 추출하고 이를 수정함
- 결과화면 열람/이전문서 검색 및 재열람

4 - 2. Web Server

- CRUD 서비스 구현, 처리된 텍스트 저장
- 해시태그를 이용하 탐색지원
- 장고 어드민을 이용한 문서 수정과 삭제 기능

4 - 3. TextRank Algorithm

- 구글이 만든 PageRank 알고리즘에서 파생된 알고리즘인 TextRank 는 노드가 될만한 text unit 을 추출하고 이들을 출현빈도에 따라 간선을 구성하는 알고리즘이다.

- Pagerank 알고리즘을 이용해 유사하게 복잡도를 계산해본다면 단어의 가중치 벡터 R 을 구하는 방법은 Power Method 를 이용하는 것이다. 초기 가중치 벡터 r 의 모든 값을 1/N 으로 설정하고, 간선에 따라 각 단어에 값들을 간선의 수만큼 r 을 나누어 다른 단어들에게 전달 한 후, 임의의 Surfer 값을 정하여 다른 단어들로부터 할당받은 값의 합 x surfer + 1/N x (1-surfer) 인 r 의 값들이 변동이 없이 수렴할때까지 아래와같이 전이행렬 P 를 계속해서 반복하여 구하면 된다. 그리고 이 단어들을 원래의 문장에 대입하고, 전체적으로 가중치가 높은 문장들을 추출하여 요약문을 작성한다.

- 이럴경우, 일반적으로 복잡도는 O(im), (i=반복횟수, m=그래프엣지의 총수)라 이야기되고 이 최악의 케이스는 O(iN)^2 정도로 마무리되기에 이 연산은 서버에서 소화할 수 있을것으로 예상된다.

4 - 4. AWS Comprehend

- 아마존의 AWS Comprehend API 는 자체적인 알고리즘에 의하여 요청받은 텍스트 문서를 각 단어별로 가중치를 할당하고, 이를 단어 - 할당치 표로 정리하여 출력하는 API 이다.

5. 개발일정

	Milestone1(4~7주차 예상)	Milestone2(7~11주차 예상)	Milestone3(최종)
개념	각단에서 최우선기능구현	이를 통신을 이용해 연결/ 알고리즘의 본격적인 활용	오류색출/개선점 개선
앱	*촬영에서 텍스트추출 전체적 UI개발	API활용	디버깅/기능추가
서버	DB에 저장할 모델설정 실제 활용할 함수 구현	*API를 이용해 앱과 통신 알고리즘을 서버에서 실사용	
기술	알고리즘의 완벽한 이해	알고리즘 재사용	*알고리즘 발전or 더 나은 대안제시

6. 역할분담

정찬휘	서버통신, 해쉬태그화
한연수	OCR, 서버통신
우승빈	안드로이드 UI, TextRank