사 업 계 획 서

1. 서비스 개요

1-가. 서비스 명

Cloud News : 자연어 처리 및 OCR을 이용한 군 신문 빅데이터 분석 서비스

=> https://cloudnews.run.goorm.io/ (본 서비스는 Chrome에 최적화되어있습니다.)

1-나. 서비스 요약

군 신문의 특성상 타 신문들보다 접하기 어렵고, 여러 포털들처럼 다양한 종류의 신문을 한군데 모아놓는 서비스를 제공하지 않아 접근성이 떨어집니다. 저희 서비스는 군 장병들의 신문에 대한 진입장벽을 낮추고, 빅데이터 분석기법을 이용해 해당 월 자에 대한 시각적인 자료, 키워드 등의 분석을 통한 편의성 증대를 목적으로 하고 있습니다.

2. 서비스의 적절성

2-1. 활용 공공데이터

- 국방 공공데이터 – 국방부_공사신문.zip, 국방부_육사신문.zip
- 육군 사관학교 – 육사신보 (2021.01 / 2021.03 / 2021.05)
- 공군 사관학교 – 공사신보 (2021.03 / 2021.05)
- 국가 보훈처 – 나라사랑신문 (2021.01 ~ 2021.05)
- 국방 홍보원 – 국방일보 (2021.01 ~ 2021.06)

2-2. 서비스 개발 배경 및 활용 적절성

< 국방 및 안보에 대한 관심 증대의 필요성 >

군과 관련된 다양한 컨텐츠 중, 군 장병들에게 소외되고 있으면서도 중요한 정보들을 전달해주는 '신문'에 대한 인식 제고와 정보전달의 필요성을 느끼게 되었습니다. 신문은 안보 및 국방과 관련하여 핵심적인 정보들을 갖고 있어, 이를 활용해 장병들이 쉽고 빠르게 접근할 수 있는 계기를 마련하고 군 소식에 더욱더 관심을 갖게 하는 부분에 기여하고자 개발하게 되었습니다.

< 국방 데이터의 실용적인 가공 >

국방 공공데이터라고 정확히 명시된 데이터는 공군 사관학교 신문 및 육군 사관학교 신문을 사용하였으나, 그 외에 국방일보나 나라사랑신문 등 모두 국방과 관련된 홈페이지에서 얻을 수 있는 데이터들로 구성하였습니다. 서비스 소비층은 주로 군 관계자나 장병분들을 대상으로 타겟팅을 하여 최대한 국방과 관련된 데이터를 활용하고자 했습니다.

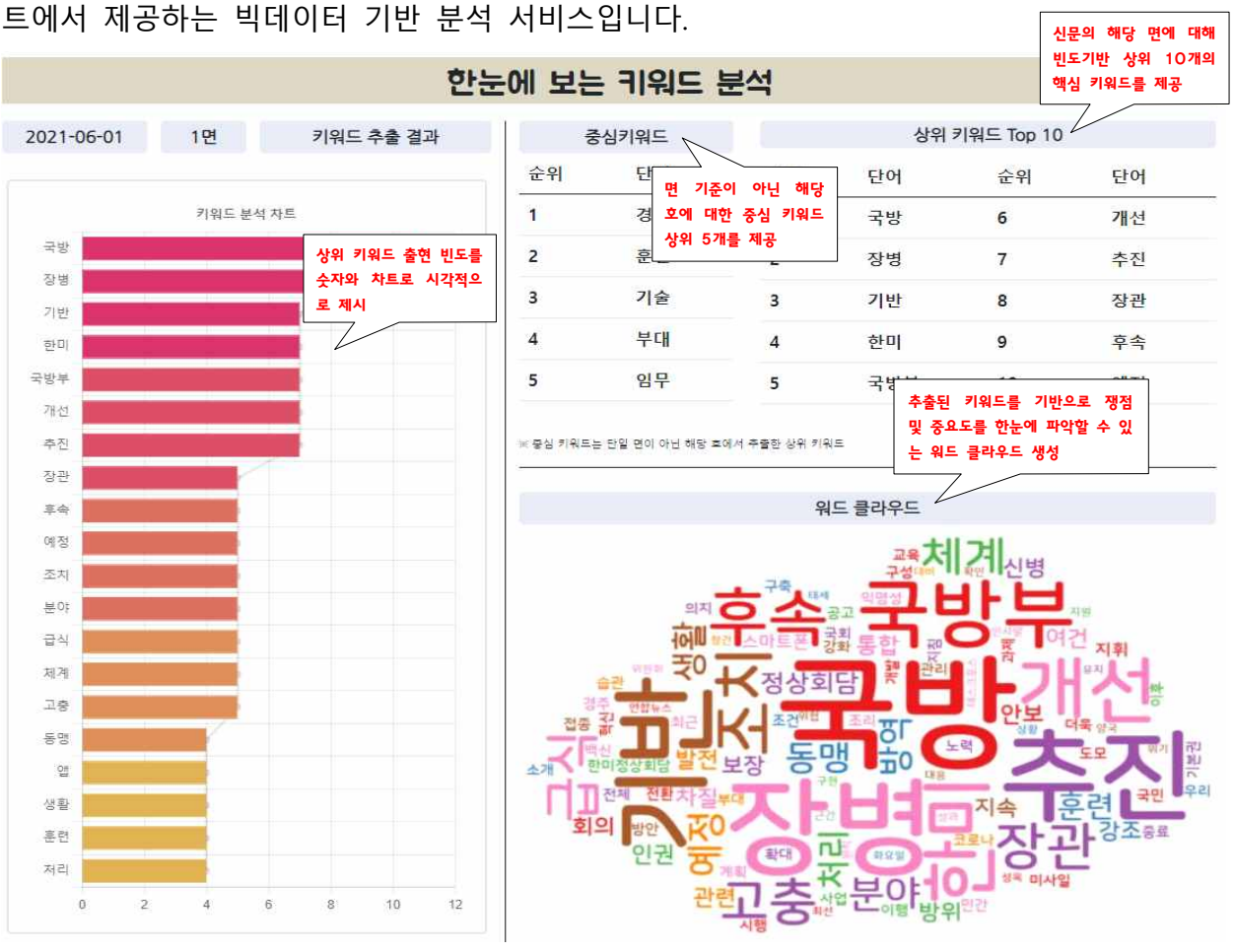
사용한 데이터들은 모두 직접 구성한 정형화된 파이프라인을 통해 정제 및 가공을 거쳤으며, 이렇게 정제된 데이터를 바탕으로 시각화를 통해 서비스 화면을 구성합니다. 내부적으로는 해당 서비스가 추후 실용화될 경우 구성한 파이프라인을 통해 자동으로 동작하는 서비스까지 연계가 가능하도록 개발하였습니다.

3. 서비스의 독창성

국방과 관련된 신문에 대한 원본 데이터만을 제공하는 기존의 국방 홍보원, 국가 보훈처 등의 사이트들과 달리, 저희 서비스는 **빅데이터 분석기법을 활용**해 뉴스라는 공공데이터가 가진 가치의 창출에 집중합니다. **AI기반의 OCR기술과 자연어 처리기법**을 통해 추출된 텍스트를 대상으로 키워드 추출 및 분석을 통해 뉴스들의 핵심 키워드들을 시각적으로 제공함으로써 장병들의 신문에 대한 거부감으로 인한 낮은 구독률을 개선하고 접근성을 높입니다.

특히 저희 서비스는 정형화된 Excel 또는 표 형태의 데이터를 사용하는 타 서비스들과 달리, **이미지나 PDF 형식의 비정형 데이터를 사용해 유의미한 인사이트를 제공**하고 있다는 점에서 타 서비스들과의 차별화를 두고 있습니다. 구체적으로, PDF형식의 데이터뿐만 아니라, 이미지로부터 텍스트를 인식해 추출하는 OCR 기법을 사용하여 데이터가 이미지인 경우에도 유의미한 분석 결과물을 만들어낼 수 있습니다. 또한, **자체적으로 개발한 데이터 정제 및 분석 파이프라인을 통해 새로 입력되는 데이터에 대해 분석결과 도출의 자동화가 가능하도록 연결이 용이**하다는 점도 큰 장점으로 작용합니다.

뉴스 데이터 외에도 이러한 서비스를 국방부와 연관된 **다양한 서비스에 접목**을 시킬 수 있어 컨텐츠적인 측면에서 높은 확장성과 잠재력을 내포한 서비스입니다. 아래는 저희 웹사이트에서 제공하는 빅데이터 기반 분석 서비스입니다.



(ex. 국방일보 2021.06.01. - 1면)

4. 서비스의 기술성

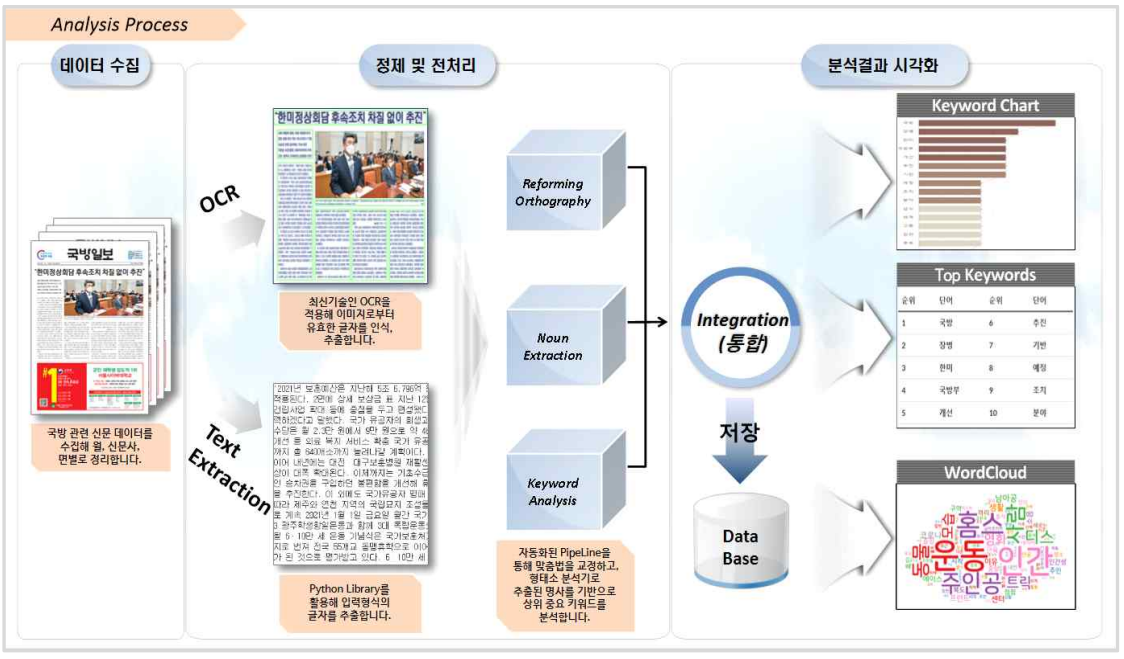
전반적인 서비스의 기술성은 크게 데이터 정제 및 분석, 개발 파이프라인, Web 구현으로 3가지 파트로 구성됩니다. 세부 내용은 다음과 같습니다.

1. 정제 및 분석

데이터는 뉴스 PDF와 뉴스 이미지, 두 가지를 사용합니다. 이미지 파일인 경우, 빛을 이용해 문자를 판독한 뒤, 그 반사 광선을 전기 신호로 바꾸어 글자를 인식하는 최신기술인 OCR을 적용해 유효한 글자들을 추출합니다. PDF인 경우는 Python의 라이브러리인 tika, PyPDF 등을 이용해 입력되어있는 글자들을 추출합니다. 이때 월, 신문사, 면별로 PDF 또는 이미지를 분할한 뒤, 추출된 글자들을 정리합니다. 동시에 신문별로 페이지 수도 다르므로 이 또한 파이썬 상에서 구현한 함수를 통해 자동으로 계산됩니다.

글자들은 단번에 깔끔하게 추출되지 않습니다. 맞춤법이나 띄어쓰기에 오류가 많아 정규표현식을 통해 특수문자, 기호 등의 무의미한 문자들을 걸러냅니다. 그 후 handspell 라이브러리를 통해 자동화된 띄어쓰기 및 맞춤법 교정으로 가독성을 높이고 유의미한 단어들이 잘 인식되도록 정제합니다.

이렇게 추출된 문자들에 대한 기본적인 정제가 끝나면 빅데이터 분석기법 중 사람의 발화 및 작문을 다루는 자연어 처리를 수행합니다. 형태소 분석기 중 하나인 KoNLP의 Okt 모듈을 사용하여 명사들을 추출하고 해당 명사들의 빈도수를 분석하여 데이터 프레임 형태의 CSV로 가공합니다. 이때 해당 과정에서 무의미한 단어들은 제외하는 불용어 제거과정이 수반됩니다. 추출된 명사 중 의미는 없는데 지나치게 자주 나오는 단어들로 판단되는 경우, 불용어로 지정해 키워드에서 제외합니다. 이렇게 선정된 키워드로 워드 클라우드를 생성해 이미지 형태로 제공합니다. 이렇게 산출된 키워드와 워드 클라우드는 파이프라인을 통해 Web에 연결된 데이터베이스에 저장되게 됩니다. 전체적인 과정은 다음과 같습니다.



2. 개발 파이프라인

- 파일 형식의 데이터를 가공하여 DB에 저장.
- 자동화된 데이터 처리 파이프라인의 기반으로 분석이 완료된 개별 CSV, PDF 등을 데이터베이스 스키마에 맞게 가공하여 테이블에 삽입. (파이프라인 구성 시 가볍고 빠르게 동작할 수 있도록 최대한 파이썬 내장 라이브러리만을 활용하여 자체 개발)
- 삽입된 데이터를 기반으로 웹사이트에서 동적 차트 구현 및 화면 세부 구성.
- 추후, 추가되는 신문데이터를 공공데이터 API로 받을 수 있게 되면 분석 코드와 융합하여 새로운 신문 및 키워드를 실시간으로 데이터베이스에 추가하는 파이프라인으로 확장이 용이하도록 개발. (API -> DataAnalysis code -> csv/pdf file -> PipeLine -> DB -> Web)

3. Web 기능 및 특징

- 4가지 신문에 대한 뉴스정보를 통합해 제공.
- 날짜 필터링과 원하는 키워드 검색기능.
- 뉴스 상위 키워드 도출 및 동적 차트와 워드 클라우드를 이용한 시각화.
- 신문 면 단위의 키워드와 해당 호의 중심 키워드를 같이 제공.
- 사용자가 시각적인 부담을 느끼지 않을 수 있는 색감을 고려해 화면 UI를 구성.
- 사용자 편의를 고려해 쉬운 화면 넘김, 로딩, 시각화 요소들을 최적화.

< 개발 환경 >

- 개발 및 배포용 클라우드 서버 goorm IDE 사용
- Javascript: nodeJs + ExpressJs
- Design Framework: Bootstrap + tailwind css
- Data Base: MYSQL
- 데이터 정제 및 분석: Python – Google Colab 사용

※ **Service web site** : https://cloudnews.run.goorm.io/ (본 서비스는 Chrome에 최적화되어있습니다.)

소스코드 : www.github.com/jangsus1/CloudNews

5. 서비스의 가능성

5-1. 서비스의 사업성

서비스의 사업성은 다음과 같습니다.

첫 번째로, 카카오 플러스와 연동한다면 사용자들이 신청할 경우 신청한 신문에 대해 매호마다 키워드를 보내주는 형식으로 뉴스들을 정리해주는 구독 서비스를 통해 수익을 창출할 수 있습니다.

두 번째로, 군 관계자와 장병들을 주요 소비층으로 타겟팅한 서비스로써 홍보를 원하는 국방과 연관된 채널이나 기관으로부터 금액을 제공받고 관련 데이터 또는 광고매체를 DB에 저장해 사용자들에게 새로운 컨텐츠를 제공할 수 있습니다. 구체적으로 예를 들자면 사기업의 측면에서는 군인들의 구매성이 높은 제품들의 광고나 앱이 될 수 있고, 국가 기관에서는 군 관련 지원금, 군인 혜택들을 홍보하는 좋은 수단이 될 수 있습니다. 구축된 웹사이트 특성상 배너광고 이식이 용이한 것이 장점으로 작용합니다.

5-2. 사회적 가치 창출

1. 산출된 키워드 데이터를 바탕으로 이를 축적하여 일별 주요 키워드 동향, 이슈의 흐름 파악 등 분석영역의 확장을 통해 '뉴스 대쉬보드 모니터링' 시스템의 구축이 가능해집니다.

2. 국방일보에서 발생한 뉴스 표기 오류사항에 대해 유연한 대처가 가능합니다. 실제로 2021년 5월 14일자 미사령관의 한국 이름을 잘못 표기하여 국방일보 14만부를 폐기한 사례가 있는데, 이와 같은 오류로 인한 추가적인 비용을 줄여줍니다.

3. 대부분 종이신문은 공급된 만큼의 수요를 갖고있지 않습니다. 이는 어느 정도의 수요가 필요한지, 독자층이 얼마나 되는지를 파악하지 못한 한계에서 나오는 부작용이라고 생각합니다. 본 서비스가 상용화되어 이용한 사람들을 대상으로 연령대, 지역 등의 간소화된 프로필을 제공받음으로써, 어느 지역에 어느 정도의 수요가 발생하는지 예측하는 것에 있어서 기반 데이터를 추출할 수 있을 것으로 생각됩니다. 이렇게 되면 종이신문의 수요량을 고려한 발행을 통해 불필요한 공급비용의 감축이 가능합니다.

4. 본 서비스의 주된 목적이자 개발의의라고 할 수 있습니다. '신문'이라는 매체에 대한 군 장병들의 거부감을 개선하고 접근성을 높여 올바른 인식과 군 소식에 더 관심을 가지는 장병들의 양성에 기여하는 교육적인 목표를 달성합니다.

6. 기타

해당 서비스를 구축하면서 아쉬웠던 점은, 최신 신문 데이터의 부재였습니다. 국방 공공데이터, 국가 보훈처, 국방홍보원 등에서 신문들을 직접 수집하다보니 신문들이 존재함에도 불구하고 누락되었거나 국방 공공데이터에는 최신 신문들이 갱신되지 않은 경우가 있어 2021년에 발간된 신문에 대해 월별로 한 부로 한정해 진행했습니다.

따라서 만약 신문 데이터를 공공데이터 API를 통해 제공받을 수 있다면, 실시간으로 새로 발간된 신문을 서비스에 업로드해 개발되어있는 파이프 라인과 연결을 통한 분석 및 결과를 실시간으로 제공하는 서비스로 확장이 가능합니다. 이를 통해 해당 서비스의 확장 및 개선을 도모하고 싶습니다. 동시에 현재 서비스에 등록되지 않은 다른종류의 국방관련 신문 데이터가 추가된다면 다양한 신문들의 추가도 가능할 것으로 보입니다.