네이버 뉴스 데이터,
수집 및 정제

질의 응답 AI를 위한 데이터 크롤링

3조 퓰리처
조장 : 조아빈
조원 : 김민준
조원 : 이명재
조원 : 김한솔

자연어 이해 인공지능 전문인력(AI학습 엔진 구축과 활용)

발표일자 : 2022.04.06

프로젝트 기간 : 22.03.07~22.04.07

발표 순서

01 추진 목표	02 과제 수행 범위	03 요구사항 정의
04 WBS	05 프로세스 설계	06 후기

“

질의응답 AI를 위한 데이터 크롤링

뉴스를 크롤링하여 질의응답을 위한 데이터 수집

데이터의 질적 향상 추구를 위한 중복 기사 제외

“

데이터 수집을 위한 뉴스 크롤링

뉴스 목록 확인

크롤링 주제에 맞는 뉴스 수집

상세 내용 수집 - 제목, 기사 내용, 분야, 날짜, 기자 명수집

—

WHY?

NAVER

HTML

수 많은 언론사를 하나의 공통의 HTML 접근가능

오류 최소화

공통된 HTML로 데이터 수집간 오류 최소화

다양한 연산자 검색

원하는 정보와 제외할 정보를 골라 데이터 검색용이

확실한 데이터접근

공통 HTML, 연산자 검색 등 상세한 검색 조건을 처리 가능

네이버 로고 출처 - https://ko.wikipedia.org/wiki/%ED%8C%8C%EC%9D%BC:Naver_Logotype.svg

01

2022 대선

02

우크라이나

올해 1분기 가장 이슈인 두 주제로 선정

대선 사진 출처- https://news.kbs.co.kr/special/election2022/president/index.html
우크라이나 사진 출처 - https://imnews.imbc.com/news/2022/world/article/6345153_35680.html

2022 대선

결과가 나온 키워드

22.03.10
~
22.03.15

정확한 데이터를 위한 검색 기간 설정

관련도 순

정확한 데이터를 위한 검색 방법

4,000 개

기간과 검색조건에 맞는 기사 수집 개수

우크라이나

진행 중인 키워드

현 시점부터 24시간 전 기사 수집

진행 중인 사건으로 투고된 날짜가
24시간이 지난 기사들 수집

최신 순

정확한 데이터를 위한 검색 방법

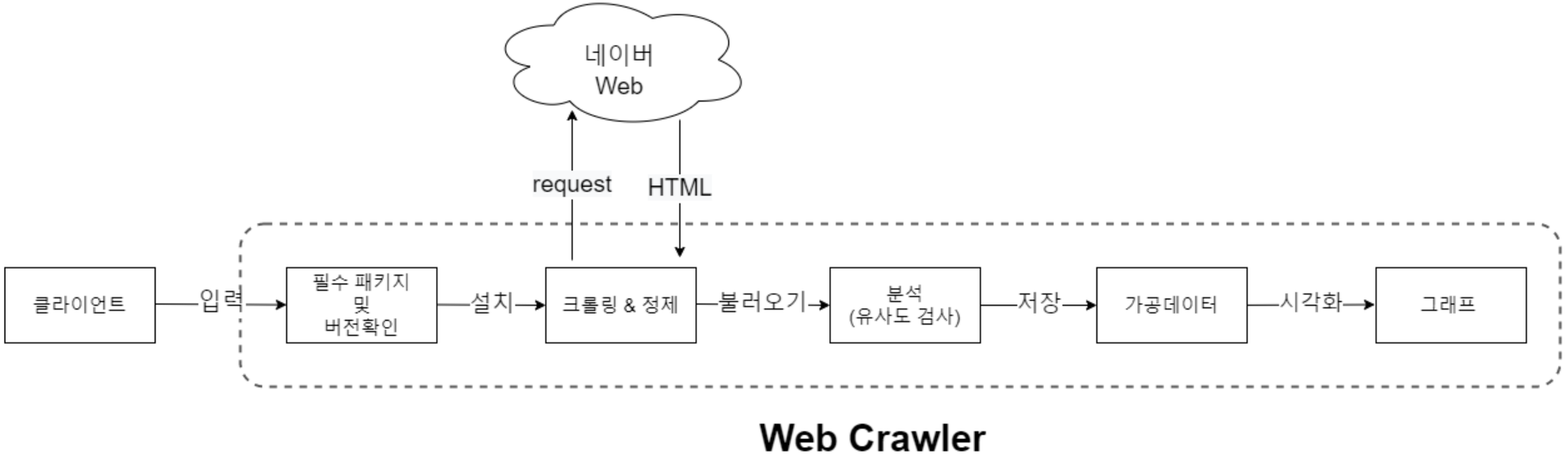
4,000 개

기간과 검색조건에 맞는 기사 수집 개수

● ● ○ 요구사항 정의

☑ **400자 이상**	▪ 4,000개 기사 중 본문이 400자 이상인 기사만 수집 ▪ 동영상만 있는 기사 및 속보 기사 제외
☑ **중복기사 제외**	▪ 유사도 검사를 통한 필터링
☑ **키워드의 시각화**	▪ 크롤링한 기사 중 가장 많이 언급된 단어로 랭킹 리스트 작성 ▪ 워드 클라우드를 통한 이미지 시각화

● ● ○ 프로세스 설계



01 필수 패키지 및 버전확인

- 크롤러 실행 전 필요한 모듈 및 프로그램 설치 후 설치된 버전확인

02 크롤링 & 정제

- 요구사항 정의서에 맞는 데이터 수집

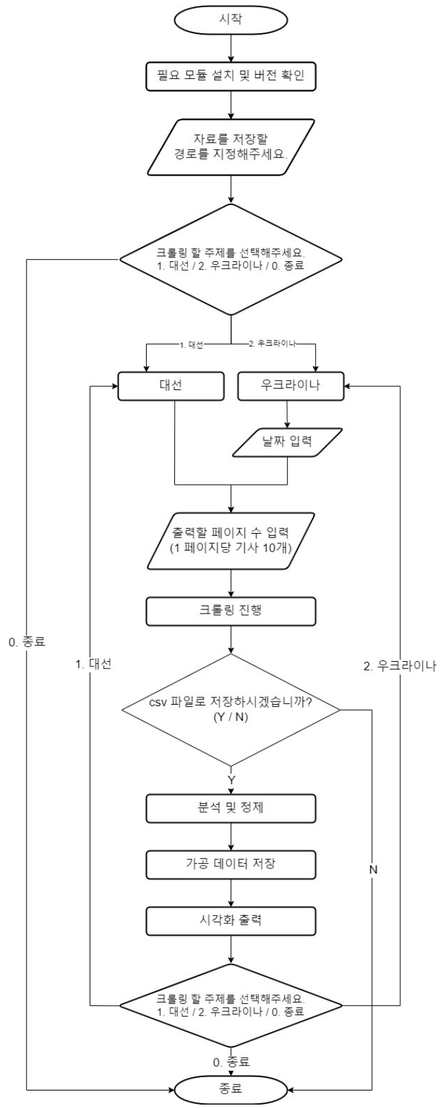
03 분석

- 유사도 검사를 통해서 중복기사 제외

04 시각화

- 가공 데이터로 워드 클라우드 및 다양한 그래프 구현

프로세스 설계 - Flow Chart(흐름도)



01 수집 페이지 수 지정

- 수집량은 4,000개로 지정
- 원하는 만큼 페이지 수집 가능하도록 기능 구현

02 날짜 Input

- ‘2022년 대선‘ 고정날짜 지정(22.03.10 ~ 22.03.15)
- ‘우크라이나’ 원하는 검색 날짜 검색 하도록 지정

● ● ○ 대선 - 테스트 결과 및 수집 결과

```
==================================검색 옵션 선택==================================

크롤링할 주제를 선택해주세요.

1. 대선 / 2. 우크라이나 / 0. 종료 : 1

대선 선택지입니다

출력할 페이지 수를 입력해주세요. (1 페이지당 기사 10개) : 400
```

```
==================================크롤링 시작==================================

1 페이지


400자 이상
1 번 기사
뉴시스
종합 뉴스 제목 :  20대 대선 재외투표 李 59.77% 尹 36.19% 沈 3.45%
이재우 기자


400자 이상
2 번 기사
연합뉴스
연합 뉴스 제목 :  '대선 패배' 민주당에 오히려 입당 러시…나흘만에 10만명 신청
정수연 기자


400자 이상
3 번 기사
주간조선
종합 뉴스 제목 :  대선 후보 수사 속도... 이재명 가족 소환 되나
이정현 기자
```

<화면 실행 결과>

4월 12일 22시 3분 news.csv 2022-04-12 오후 10:03 Microsoft Excel ... 1,688KB

4월 12일 22시 3분 yeonhab.csv 2022-04-12 오후 10:03 Microsoft Excel ... 220KB

	언론사	분야	날짜	제목	본문	기자	링크
0	뉴시스	정치	2022.03.15	20대 대선	기사내용	이재우 기	https://news.naver.com/main/read.naver?r
1	주간조선	정치	2022.03.15	대선 후보	지난 2월 2	이정현 기	https://news.naver.com/main/read.naver?r
2	경향신문	사회	2022.03.15	경찰, 김혜	[경향신문]	김태희 기	https://news.naver.com/main/read.naver?r
3	아시아경제	정치	2022.03.15	우상호 "서	우상호 총	이기민 기	https://news.naver.com/main/read.naver?r
4	뉴스1	사회	2022.03.15	세종선관위	20대 대선	장동열 기	https://news.naver.com/main/read.naver?r
5	노컷뉴스	정치	2022.03.15	대선 승리	핵심요약0	임진수 기	https://news.naver.com/main/read.naver?r
6	한국경제	정치	2022.03.15	대선을 찢(	카페 개설	홍민성 기	https://news.naver.com/main/read.naver?r

1680	한겨레21	사회	2022.03.15	윤석열 공	[표지이야	방준호 기	https://news.naver.com/main/re
1681	매일경제	정치	2022.03.14	[단독] 尹당	'마켓일기'	이희수 기	https://news.naver.com/main/re
1682	JTBC	사회	2022.03.14	저축은행부	동영상 뉴	박병현 기	https://news.naver.com/main/re
1683	뉴시스	정치	2022.03.14	[일문일답]	기사내용	이지율 기	https://news.naver.com/main/re
1684	세계일보	경제	2022.03.15	"尹 당선되	중개업소 "	김현주 기	https://news.naver.com/main/re

<파일 저장 및 CSV 저장 결과>

● ● ○ 우크라이나 - 테스트 결과 및 수집 결과

```
==================================검색 옵션 선택==================================

크롤링할 주제를 선택해주세요.

1. 대선 / 2. 우크라이나 / 0. 종료 : 2

우크라이나 선택지입니다

출력할 페이지 수를 입력해주세요. (1 페이지당 기사 10개) : 400
시작 날짜 입력 (2022.01.01) : 20220410
끝 날짜 입력 (2022.01.01) : 20220410
```

```
==================================크롤링 시작==================================

1 페이지


400자 이상
1 번 기사
중앙일보
종합 뉴스 제목 :  러시아 우크라 돈바스 공세 임박 관측…"동부서 13㎞ 길이 러군 차량 행렬"
이지영 기자


400자 이상
2 번 기사
연합뉴스
연합 뉴스 제목 :  [우크라 침공] 우크라 검찰총장 "전범 혐의 500명 파악"
최윤정 기자


400자 이상
3 번 기사
뉴시스
종합 뉴스 제목 :  美 "신임 러시아 우크라이나전 사령관, 민간인 공격만행 확대할 것" 경고
이재준 기자
```

<화면 실행 결과>

4월 12일 22시 3분 news.csv 2022-04-12 오후 10:03 Microsoft Excel ... 1,688KB

4월 12일 22시 3분 yeonhab.csv 2022-04-12 오후 10:03 Microsoft Excel ... 220KB

	언론사	분야	날짜	제목	본문	기자	링크
0	중앙일보	세계	2022.04.10	러시아 우	미국 민간	이지영 기	https://news.naver.com/main/re
1	뉴시스	세계	2022.04.10	美 "신임 러	[서울=뉴시	이재준 기	https://news.naver.com/main/re
2	데일리안	세계	2022.04.10	아이 4명 등	남편과 조	황기현 기	https://news.naver.com/main/re
3	세계일보	오피니언	2022.04.10	[사설] 권성	국민의힘 권성동 신임		https://news.naver.com/main/re
4	세계일보	오피니언	2022.04.10	[설왕설래]	1995년 1월	주춘렬 기	https://news.naver.com/main/re
5	세계일보	오피니언	2022.04.10	[주재우의	우크라이나 전쟁에 다		https://news.naver.com/main/re
6	세계일보	오피니언	2022.04.10	[WT논평]	By Jed Babbin (natio		https://news.naver.com/main/re
7	한겨레	세계	2022.04.10	러시아, 우	민간 위성업체 맥사 E		https://news.naver.com/main/re
8	뉴스1	세계	2022.04.10	젤렌스키 "	올라프 숄	원태성 기	https://news.naver.com/main/re
9	뉴시스	세계	2022.04.10	우크라의	러시아 미	김재영 기	https://news.naver.com/main/re
10	중앙일보	세계	2022.04.10	위성에 찍	지난 8일(현	정은혜 기	https://news.naver.com/main/re
11	MBC	세계	2022.04.10	돈바스 전	자료사진 제공 : 연합		https://news.naver.com/main/re
433	뉴시스	세계	2022.04.10	존슨 英종	기사내용	권성근 기	https://news.naver.com/main/r
434	KBS	세계	2022.04.10	EU 등 국저	유럽연합(E	이중근 기	https://news.naver.com/main/r
435	서울신문	세계	2022.04.10	英 존슨 총	"우크라이	김소라 기	https://news.naver.com/main/r
436	더팩트	경제	2022.04.10	[TF비즈토	삼성·LG 역	박경현 기	https://news.naver.com/main/r
437	더팩트	경제	2022.04.10	[TF비즈토	'매출 77조	박경현 기	https://news.naver.com/main/r
438	파이낸셜뉴	경제	2022.04.10	해인싸)변	[뉴욕=AP/	서혜진 기	https://news.naver.com/main/r

<파일 저장 및 CSV 저장 결과>

● ● ○ 시각화

TOP20 단어: [('우크라', 343), ('러시아', 267), ('가격', 178), ('달러', 132), ('전쟁', 127), ('정부', 116), ('기업', 113), ('이우', 106), ('대통령', 99), ('지원', 96), ('러시아군', 94), ('미국', 92), ('지역', 86), ('가능', 86), ('우크라이나', 80), ('시장', 80), ('세계', 80), ('동부', 79), ('배출', 79), ('상승', 78)]

TOP20 단어: [('후보', 6314), ('대통령', 6278), ('대선', 6267), ('선거', 5828), ('당선인', 4895), ('국민', 4519), ('민주당', 4126), ('서울', 3077), ('윤석열', 2917), ('의원', 2477), ('기자', 2299), ('정부', 2193), ('위원장', 2130), ('대표', 1773), ('정치', 1772), ('지방', 1615), ('이재명', 1556), ('수사', 1533), ('지역', 1493), ('이번', 1461)]

〈TOP20 단어 결과〉

〈워드 클라우드 결과〉

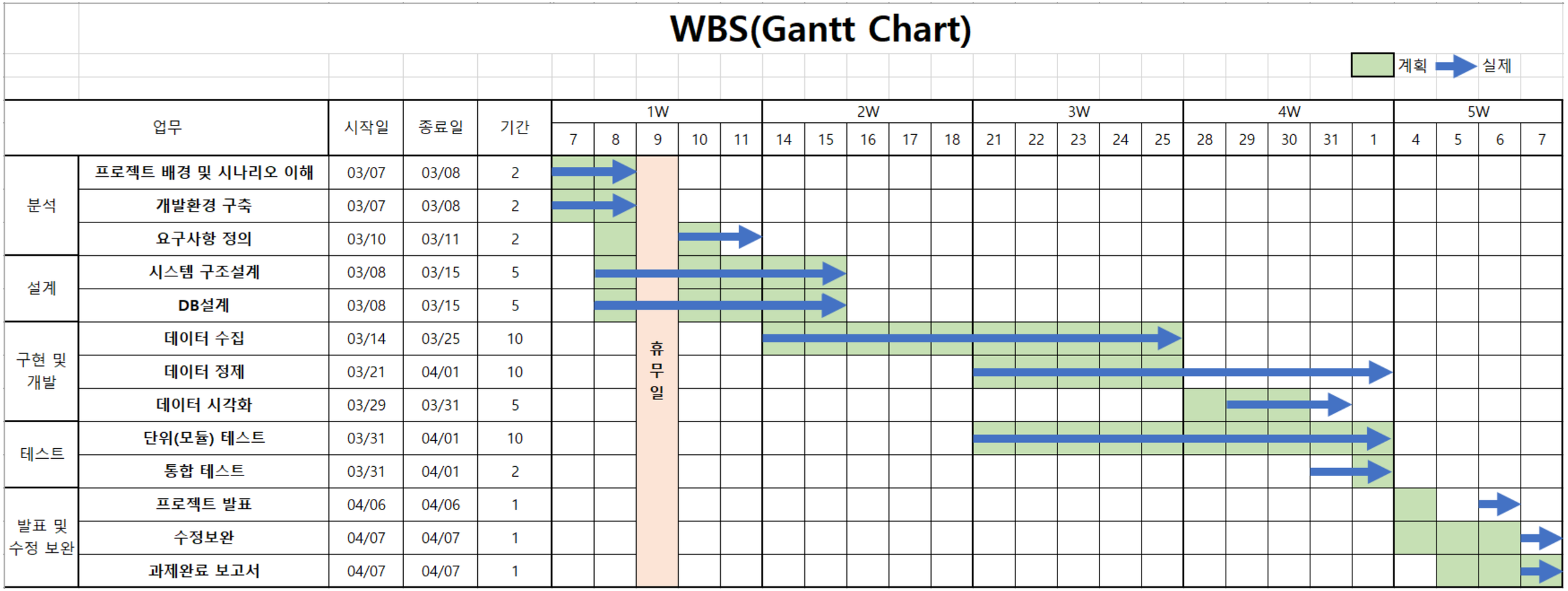
● ● ○ 테스트 시나리오

통합 테스트 시나리오					
프로젝트	네이버 뉴스 데이터 수집 및 정제	**단계**		구현	
시스템	데이터 수집, 정제	**버전**	v0.9.1	**수행자**	이명재 김한솔
작성자	3조 퓰리처 (조아빈, 김민준, 이명재, 김한솔)	**작성일자**	2022.03.31	**수행일자**	2022.04.01
시스템 / 서브 시스템 / 모듈	네이버 뉴스 데이터 수집 시스템				

시나리오	시나리오명	상세 설명(흐름도)	검증 포인트	Pass / Failure
TS-001	필요 모듈 설치 확인 및 실처, 버전 확인	필요 모듈 설치 여부 확인, 존재하지 않을시 설치하며 버전을 프린트 하여 확인	모듈 확인, 설치, 설치 버전 프린트	P
TS-002	자료 저장할 폴더경로 지정	이미 존재하는 폴더이면 그 폴더에 저장, 아닐 경우 입력한 폴더명으로 생성한 뒤 저장	일치하는 경로에 폴더생성/ 지정	P
TS-003	지정해둔 메뉴 선택 (1.대선 2.우크라이나 0.종료)	메뉴 번호(1, 2, 0)나 메뉴 이름(대선, 우크라이나, 종료)입력 시 해당 메뉴로 이동 잘못된 입력일 경우 올바르게 입력할 때까지 반복	메뉴 번호(0,1,2) 입력 시 메뉴에 대한 것을 실행	P
TS-004	출력할 페이지 수 입력 (1페이지당 기사 10개)	입력받은 수가 0이거나 음수가 아닌 올바른 입력일 때까지 반복	입력 받은 건수만큼 수집	P
TS-005	시작 날짜, 끝 날짜 입력	대선일 경우 날짜를 2022.03.10.~2022.03.15.로 고정 그 외엔 입력한 날짜로 진행 8자리 이상, 10자리 이하인 경우만 입력받음	입력된 날짜로 검색	P
TS-006	크롤링 진행	제목, 날짜, 언론사, 제목, 본문, 기자명, 기사링크, 분야를 수집 기자 카드 없을 시 공백처리 본문 내용 400자 미만일 미수집	지정된 내용을 추출	P
TS-007	원천 데이터 저장	크롤링한 기사들을 TXT파일과 CSV로 저장한다. TXT파일은 필수 저장이지만 CSV는 저장 여부를 물어본 뒤 저장, 저장된 시간의 월, 일, 시, 분이 파일 이름 앞에 붙어 저장된다.	TXT와 CSV 및 설정해둔 이름으로 저장	P
TS-008	시각화	CSV로 저장된 파일을 TOP20랭킹과 워드클라우드로 시각화	CSV파일을 읽고, 정상적으로 출력확인	P
TS-009	TS-003로 돌아가 메뉴 재선택 여부 (1.대선 2.우크라이나 0.종료)	메뉴 번호(1, 2, 0)나 메뉴 이름(대선, 우크라이나, 종료)입력 시 해당 메뉴로 이동 잘못된 입력일 경우 올바르게 입력할 때까지 반복	메뉴 번호(0,1,2) 입력 시 메뉴에 대한 것을 실행	P
TS-010	프로그램 종료	TS-003에서 올바르게 넘어온 프로세스인가 확인	0번 메뉴 선택 시 종료	P

● ● ○ WBS(Gannt Chart)

WBS(Gantt Chart)



업무		시작일	종료일	기간
분석	프로젝트 배경 및 시나리오 이해	03/07	03/08	2
	개발환경 구축	03/07	03/08	2
	요구사항 정의	03/10	03/11	2
설계	시스템 구조설계	03/08	03/15	5
	DB설계	03/08	03/15	5
구현 및 개발	데이터 수집	03/14	03/25	10
	데이터 정제	03/21	04/01	10
	데이터 시각화	03/29	03/31	5
테스트	단위(모듈) 테스트	03/31	04/01	10
	통합 테스트	03/31	04/01	2
발표 및 수정 보완	프로젝트 발표	04/06	04/06	1
	수정보완	04/07	04/07	1
	과제완료 보고서	04/07	04/07	1

1W					2W					3W					4W					5W			
7	8	9	10	11	14	15	16	17	18	21	22	23	24	25	28	29	30	31	1	4	5	6	7

보기만해도
흐 — 뭇한 프렌즈~

3조
풀리처

조원소개 및 후기 >

조아빈

“ 기본적인 설계부터 구현, 테스트까지 팀 단위 프로젝트의 전반적인 흐름을 학습할 수 있게 되었습니다. 크롤러를 코딩하는 과정부터 단순히 목표만을 위한 코딩이 아니라 효율을 증대 시키기 위해 시간단축을 하는 등 팀원 들과의 진행을 통해 배워 나갈 수 있었습니다. ”

김한솔

“관계 중 인간관계가 제일 힘들다고 여기는데 다들 어색함에도 불구하고 편한 분위기에서 큰 문제없이 1차를 끝내서 다행이다 ”

이명재

“설계로는 평범하게 구현할 수 있을 거라 생각 했는데 구현이 뜻대로 되지 않았다. 직접적인 코딩을 작성하기보단 프로그램 오류 발생시 원인분석과 수정 쪽으로 프로젝트에 참여했다.”

김민준

“프로젝트 기간동안 팀원들 간에 의사소통에서 서로 막힘이 없다는 것과, 직장생활에서 갈고 닦았던 소통, 문서 정리와 같은 것들이 생각보다 팀에 도움이 많이 되었다고 생각합니다. 그리고 서로가 잘하는 부분을 공유하고 문제를 공유하면서 풀어내면서, 결과물이 생기니 의욕이 올랐습니다. 개인적으로 좋았던 건 코딩의 기술적으로 많이 공유할 수 있어서 좋았습니다.”

지금까지

3조의 발표였습니다 !

발표에 대한 질의응답은 언제나 환영입니다 :)

발표 끝 !