

# 캡스톤 2팀 회의록

(팀명: HIING(하잉) / 팀장: 조영훈 / 팀원: 전범수, 이헌재, 김동욱, 강신표)

작성일 : 2021.04.26

작성자 : 전범수

## 1. 개발 진행 공유

```
public ResponseEntity<Object> crawl() {
    WebDriver browser = null;
    AtomicBoolean failed = new AtomicBoolean();
    final String nationalitySuffix = "&persist_gl=1&gl=US";

    // remove duplicates
    final Set<String> urls = new HashSet<>();
    final List<Video> boxes = new ArrayList<>();
    try {
        ClassPathResource resource = new ClassPathResource("selenium/chromedriver");
        System.setProperty("webdriver.chrome.driver", resource.getURL().getPath());
        browser = new ChromeDriver();
        browser.get("https://www.youtube.com/feed/trending?bp=6gQJRkVleHBsb3Jl".concat(nationalitySuffix));
        List<WebElement> videos = browser.findElements(By.xpath("//*[@id=\"grid-container\"]/ytd-video-renderer"));
        for (WebElement element : videos) {
            try {
                WebElement link = element.findElement(By.xpath("//a[@id=\"thumbnail\"]"));
                final String href = link.getAttribute( name: "href");
                if (!urls.contains(href)) {
                    urls.add(link.getAttribute( name: "href"));
                    String id, title, captions;
                    id = href.substring(href.indexOf("v=") + 2);
                    title = element.findElement(By.xpath("//*[@id=\"video-title\"]/yt-formatted-string")).getText();
                    Response<Object> transcript = RetrofitClient.getApis().transcript(id).execute();
                    captions = transcript.body() != null ? transcript.body().toString() : null;

                    if (title != null && captions != null)
                        boxes.add(Video.builder().id(id).title(title).captions(captions)
                                    .ref(href).created(new Date()).build());
                }
            } catch (Exception e) {
                e.printStackTrace();
            }
        }
        System.out.printf("TOTAL VIDEOS DETECTED BY DRIVER: %d\n", videos.size());
        System.out.printf("URLS: %s\n\n", urls.toString());
    } catch (IOException e) {
        failed.set(true);
        e.printStackTrace();
    } finally {
        if (browser != null)
            browser.close();
    }
}
```

- selenium, java 이용 크롤링 서버 (spring boot) 제작 완료,
- 현재 영미권 유튜브의 Trend 영상을 하루 단위로 crawl 하도록 스케줄링을 설정해놨음.

```

@Getter
@NoArgsConstructor
@Document(indexName = "video")
public class Video {
    @Id
    private String id;
    private String title;

    @Field(type = FieldType.Text)
    private String captions;
    private String ref;

    @Field(type = FieldType.Date)
    private Date created;

    @Builder
    public Video(String id, String title, String captions, String ref, Date created) {
        this.id = id;
        this.title = title;
        this.captions = captions;
        this.ref = ref;
        this.created = created;
    }
}

```

- elasticsearch index, document 정의
- 후에 추가로 필요한 column이 생겨날 경우 수정 계획.

## 일정 및 향후 계획

- 멘토링 일정 예약  
: Collaborative filtering 관련 (추천 시스템) 멘토링 피드백 요청
- 크롤링 도메인 추가  
: 하루 약 30개 정도의 콘텐츠를 crawl 해볼 계획, 현재는 CC(subtitle) eng가 지원되지 않는 경우 필터링되는 방식을 사용하기 때문에, 다른 도메인에서도 크롤링해야 할당량을 채울 수 있음.
- AWS ElasticSearch 데이터베이스 구축  
: 크롤링 서버와 함께 실행한 뒤, 서버 스펙을 증설이 필요하면 증설하는 방향으로.
- 메인 페이지 튜토리얼 카드 제작  
: <https://kart.nexon.com/Kart/Guide/learn/Index.aspx> 참조해서 서비스 메인 내용에 대한 디자인 계획 수립
- 서비스 관련  
: 스케줄링 주기 (현재는 quartz를 이용, 1일에 한번씩 api가 호출됨) 선정 및 UX interaction 요소 고안(기획, UI/UX 팀)
- 디자인 / 기획  
: 검색 결과 페이지 디자인/기획, 문제 풀기 페이지의 스코어보드 위치 선정