News Crawler 개발 일지  
**1. 계기**  
 zdnet에서 it 기사를 찾아보고는 했는데, 매번 관심있는 기사를 찾는 것이 오래 걸려 사이트에 표시되어 있는 형태보다 조금 더 간결하게 보면서 귀찮지 않게 메일로 구독하는 방법은 없을까 생각했다. 그래서 다양한 모듈을 가진 파이썬을 이용해 이를 구현해 보기로 하였다.

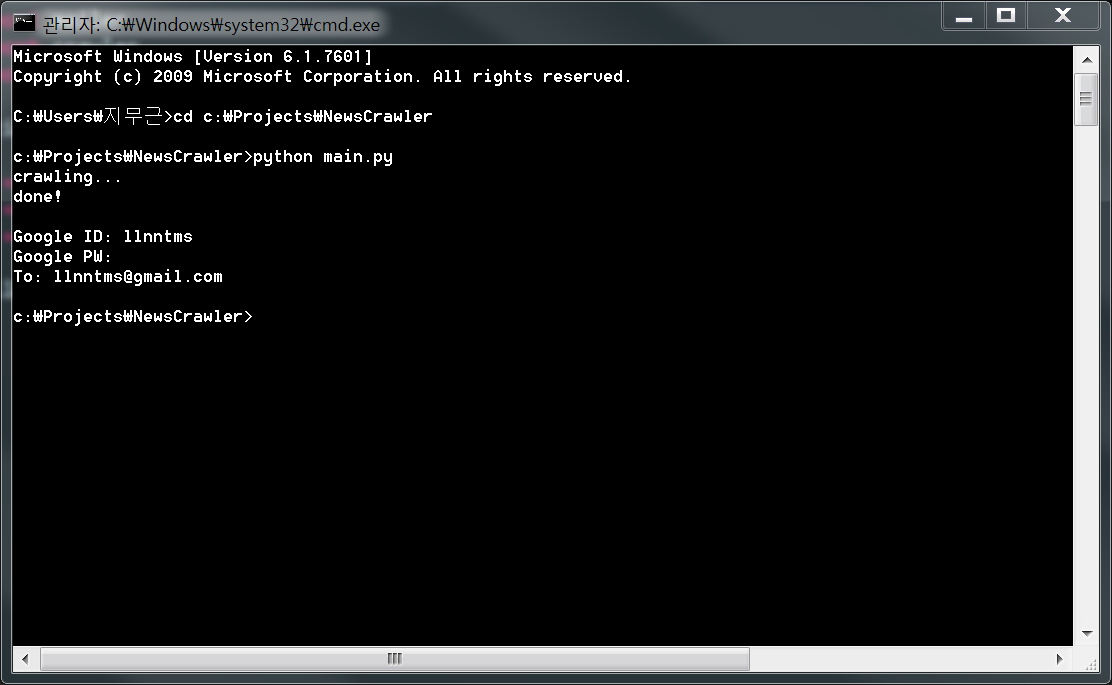
**2. 기능 구상 및 구현** 매일 아침 9시에 메일로 전날 모든 기사들의 제목 목록을 보내준다. 이때 제목에는 링크를 건다.

**3. 개발 일지**2017-02-07 기본 기능 구상  
 크게 세개의 파일로 나누어 개발한다. 사이트를 긁어 오는 모듈인 crawler.py(requests 모듈과 bs4의 BeautifulSoup모듈을 이용) 개인 메일로 보내주는 모듈인 mailer.py(smtp 모듈 이용)와 나머지 기능(긁어온 정보를 내가 원하는 형태로 바꿔주는 기능 등)을 가진 main.py이다.

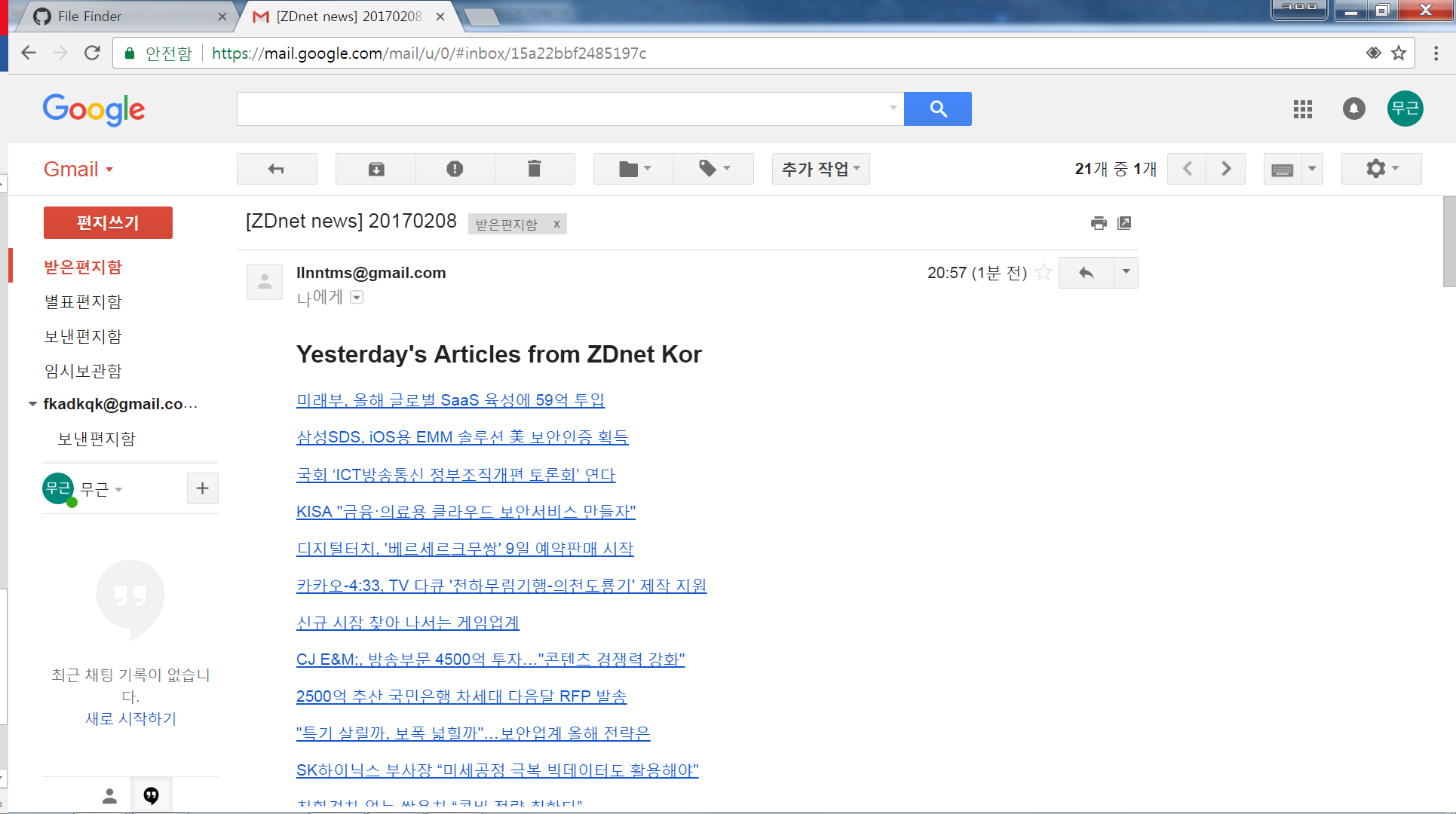
crawler는 어제자에 해당하는 기사만 가져와야 하기때문에 이를 구별하는 기능을 가져야한다.

추후 개발할 사항들:  
 1) 웹 크롤러가 특이 상황(사이트 주소나 구성이 바뀌는 등)에 유연하게 대처하는 것을 기대하기 어렵기 때문에 오류가 일어난 경우에 메일로 보고해주는 방법을 마련해야한다.  
 2) 매일 일정 시간에 main.py를 실행해야 하므로, 서버 등 항상 켜져있는 시스템에서 CRON을 이용한다  
 4) 크롤링한 것을 log로 남기고, 일정 기간이 지났다면 삭제한다.

2017-02-09 기본 기능 구현  
 crawler.py, mailer.py, main.py의 기본 기능을 구현했다.   
 crawler의 경우, 단순히 requests.get을 사용했을 때 정상적으로 페이지를 받아오지 못하는 문제가 있었지만, header의 user-agent를 직접 설정해 주니 문제가 해결되었고 정상적으로 html문서를 받아올 수 있었다. crawler.getAricleList는 기본 url을 zdnet 최신기사 1페이지로 설정하고, 받아온 각 기사의 발행일이 어제인 경우 articleList에 담는다. 이때 list에 담기는 정보는 기사 제목과 기사의 url 주소이다. 이는 추후에 더 업데이트할 예정이다. 페이지를 넘기며 기사를 수집하다가 발행일이 2일 전인 기사를 만나면 수집을 종료하고 수집된 기사 list를 return한다.   
 그 후에는 main.py에서 google smtp서버를 사용하기 위해 구글 ID와 PW를 입력 받고, 기사 목록을 수신할 email 주소를 입력 받는다. main.py는 이들을 mailer.py에 전달한다.  
 mailer의 경우, articleList에 담긴 정보를 이용해 링크가 걸린 기사 목록들로 html을 만든다. 이후 입력 받은 google ID와 PW를 사용해 SMTP\_SSL로 smtp.gmail.com:465로 접속한다(~~추후 서버를 이용하면 smtp 부분도 자체적으로 해결할 수 있을 것으로 생각된다.~~ 구글 smtp를 계속 이용하기로 함) 접속한 smtp 서버를 통해 html을 전송한다.



지금은 직접 실행하지만 서버로 옮기면 인자로 넘겨주어 지동 실행할 계획이다.



그리고 실제 메일로 받아본 기사 목록이다.

2017-02-13 centos 서버로 이식

현재 가지고 있던 centos 서버를 이용하여 python의 설치(최초 개발을 진행한 python의 버전 3.6이었지만, 자동으로 설치된 버전이 2.75이다. 호환성에 큰 문제가 발생한지 않는 한, 2.7을 기준으로 사용하겠다.)와 메일을 보내기 위한 sendmail package의 설치(메일 하나만 보내면 되기 때문에 패키지를 사용해 간단하게 진행한다.)를 완료하였고, 사용자 input으로 email 주소를 입력 받던 것을 명령 인자로 받을 수 있도록 main.py를 수정하였다. sendmail을 사용할 때, gmail에서 스팸으로 걸러진다. SPF 인증을 받아야 하는 것으로 나오지만, 여기서는 굳이 받지 않겠다. 일단은 gmail에서 주소록에 등록하는 것으로 해결한다. 또한, 매일 9시에 실행하도록 cron에 등록하였다. cron기본 시간을 수정하였다.