2023-12-02 회의

만들어진 네이버 크롤러를 기반으로 우리에게 주어진 지하철 역 데이터를 전처리.

전체 15개년 데이터의 컬럼명이 불일치 하기 때문에 일치 시키는 작업이 필요

크롤러를 통해 우선 필요 데이터(위도, 경도, 뉴스)를 csv파일로 모두 만들고 hdfs로 이동 시킨 뒤
해당 날짜 해당 역에 대한 뉴스기사 제목과 내용 그리고 위도 경도를 hdfs에서 ADD COLUMN 해준다.

그 후 데이터 분석의 용이성을 위해 승차 하차로 나누어진 데이터를 유동인구라 칭하며 각각 승차 하차에 나누어진 인원을 더해서 하나의 row로 만들어 준다.