## Data scrubbing

Twitter dataen som ble hentet ved å lytte på hashtaggen #Worlds2018 skapte store mengder med data. Denne dataen i form av twitter meldinger ble lagret i Json som er en semi-strukturert datastruktur(databasebok side 401). Tweet objektet som er lagret som et objekt med json er det man kaller «dirty data»( <https://en.wikipedia.org/wiki/Dirty_data>) Dirty data vil si ukomplett data, inkonsistent data og feilaktig data. Alt dette finner man innenfor twitter meldingene i datasettet som er samlet. Denne dataene ble først renset delvis med et python script som tok ut de delene av twitter objektet som var interessant for videre analyse. Disse attributtene ble deretter lagret i json format før det ble importert i excel for å bytte om filformatet til Excel sitt format. Dette ble gjort så det skulle være enklere å importere datasettet til rapidminer for større analyser.

Det ble også laget egne scripts for å hente ut kunne geo lokasjoner og koordinater til hvor de diverse brukerne var da twitter meldingen ble postet og lagret i et geo.json format så det kunne plasserer på dynamiske kart ved hjelp av et Python bibliotek som heter folium.