## 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. Disse 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.

## Kart visualisering

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.

Dette kartet er tilgjengelig på [**insert link**] på nettsiden er kan du klikke på punktene på kartet for å se meldingen brukeren skrev samt lokasjonsnavn hvis dette var tilgjengelig.

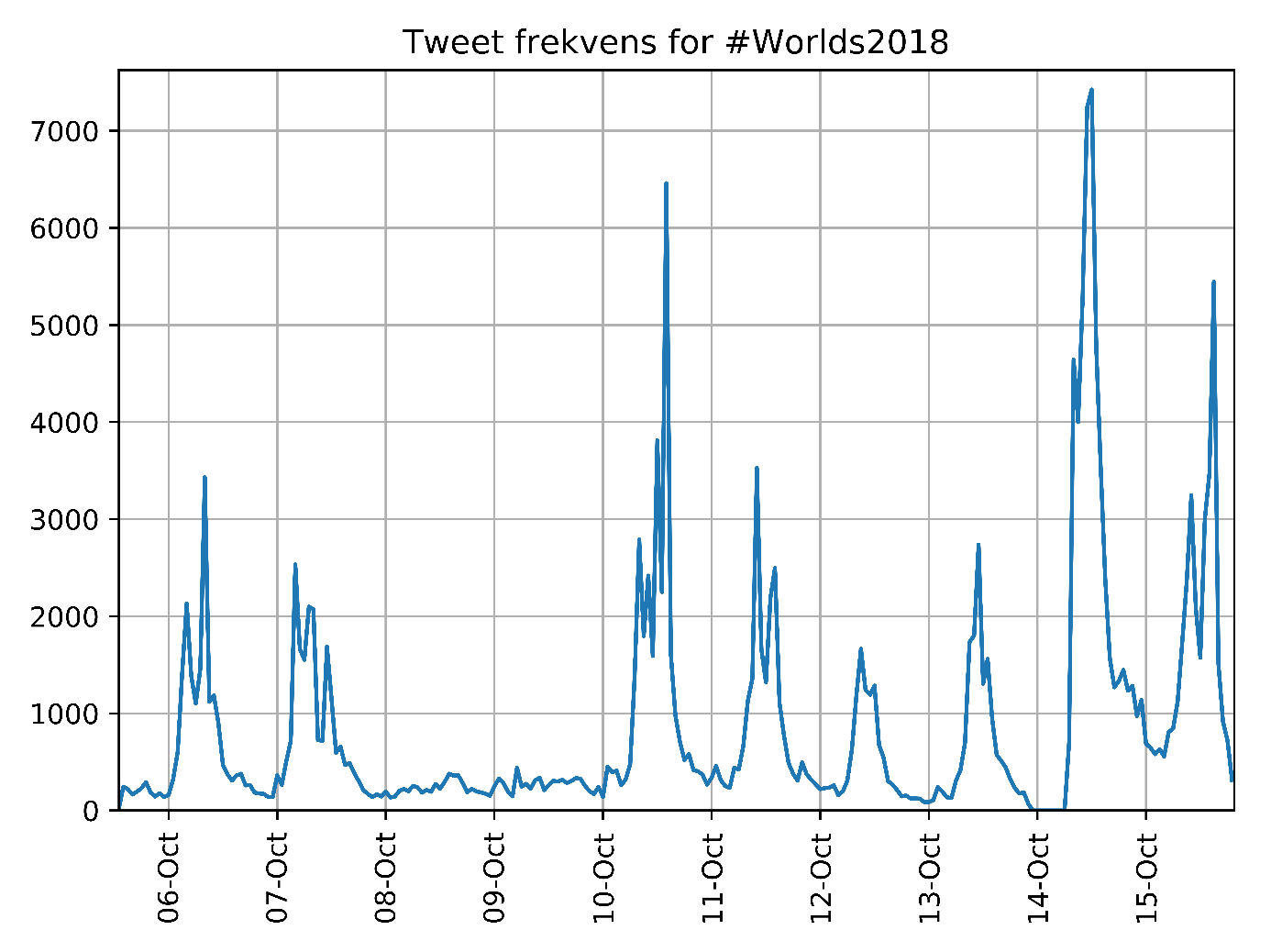
[**sett inn statisk bilde av kartet]**

Det ble laget to versjoner av kartet et med polygon markører og et med point. Polygon markørene er det som ble hentet ut av twitter objektet, problemet med å ha polygon som markører er at de kan strekke seg over en ekstremt stor del av kartet hvis en bruker har satt sin lokasjon til å være for eksempel USA. (se på kartet under eller følg linken for en dynamisk versjon [**link**]

Nevn antall som har delt lokasjondataen kontra hvor mange twitter meldinger vi samlet inn.

## Tidsserie visualisering

Får å måle hvor stor interessen har vært rundt denne hashtagen i perioden som dataene ble samlet inn har det blitt utført en tidsrekke analyse (engelsk: Time series analysis). Denne analysen er visualisert under, og viser hvor mange ganger det ble registrert en twitter melding med hashtagen «Worlds2018» per time i tidsforløpet.



Figur 1 X aksen angir datoen og Y aksen angir antall tweets per time som ble registrert i tidsrommet.

|  |  |  |  |
| --- | --- | --- | --- |
| Dato: | Antall kamper: | Antall forskjellige lag: | Type kamp: |
| Lørdag 6 oktober | 8 | 4 | Gruppespill best av 1 |
| Søndag 7 oktober | 8 | 4 | Gruppespill best av 1 |
| Onsdag 10 oktober | 6 | 6 | Gruppespill best av 1 |
| Torsdag 11 oktober | 6 | 6 | Gruppespill best av 1 |
| Fredag 12 oktober | 6 | 6 | Gruppespill best av 1 |
| Lørdag 13 oktober | 6 | 6 | Gruppespill best av 1 |
| Søndag 14 oktober | 6 | 6 | Gruppespill best av 1 |
| Mandag 15 oktober | 6 | 6 | Gruppespill best av 1 |
| Tirsdag 16 oktober | 6 | 6 | Gruppespill best av 1 |
| Onsdag 17 oktober | 6 | 6 | Gruppespill best av 1 |
| Lørdag 20 oktober | 10 | 4 | Kvartfinale best av 5 |
| Søndag 21 oktober | 7 | 4 | Kvartfinale best av 5 |
| Lørdag 27 oktober | 3 | 2 | Semifinale best av 5 |
| Søndag 28 oktober | 3 | 2 | Semifinale best av 5 |
| Lørdag 3 november | **FYLL MEG** | 2 | Finale best av 5 |

Tabell 1 Oversikt over hvilke dager det var kamper, og hva slags type kamper som ble spilt

Ved å studere tabell 1 og figur 1 er det lett å se hvilke dager det er mest populært og bruke hashtagen.

*Her burde det settes inn dataene som vetle fant, som viser seertall på twitch*