Prosjekt 2 Big data

Milepel 1 Gruppe 14. Bom T. Tran Tobias G. Søyland.

Datasett: <https://www.kaggle.com/datasnaek/youtube-new?select=US_category_id.json>

Beskrivelse av datasett

Datasettet vi valgte oss er et datasett over trending videoer på nettsiden [youtube.com](http://youtube.com).

Youtube er en videostreaming nettside der brukere kan opprette kanaler(profiler) og legge ut videoer som andre brukere kan se , kommentere på og like / dislike.

Basert på en algoritme kun kjent av Youtube selv, plukkes videoer ut basert på seertallet , antall kommentarer og forholdet mellom likes og dislikes. Disse videoene blir lagt til trending listen.

Trending-listen er en liste av videoer som er «trending» og denne listen blir vist på hovedsiden til youtube samt videoer her får prioritering i forslags algoritmen til plattformen.

Det er flere hundre videoer per dag som er «trending» og videoer slutter å være trending etter algoritmen bestemmer at de ikke lenger er trending.

Vi valgte dette datasettet ettersom det av natur kreves å bli oppdatert ofte, og fordi youtube er den største video streaming plattformen og gir stor innsikt i hvilke temaer populasjonen er interessert i.

Datasettet består av 2 deler:

Den første delen er csv filer over alle videoer som har vært trending på youtube, siden youtube har forskjellige trending lister basert på hvilket land / kontinent man befinner seg i er det da flere csv filer for hvert land/kontinent.   
  
listen inkluderer: Canada , Tyskland, Frankrike , Storbrittania og USA.

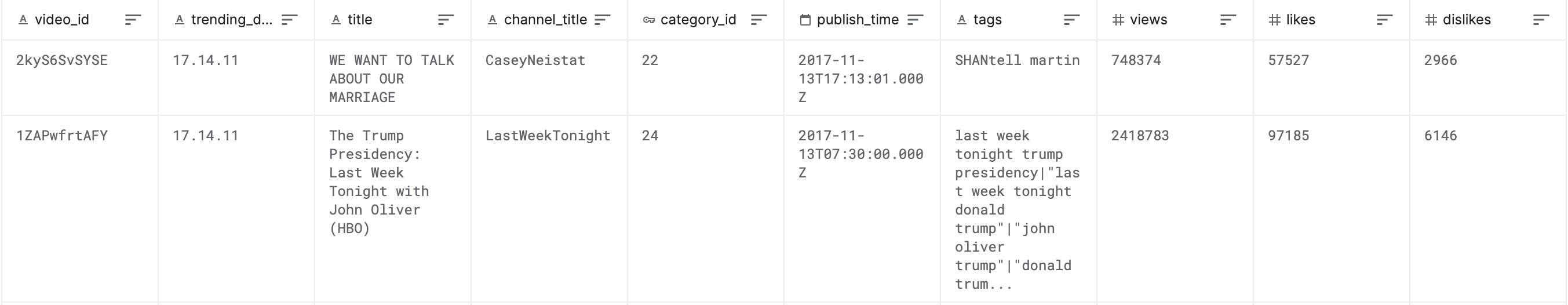
Den andre delen er en beskrivelse av video kategorier, ettersom hvert land/kontinent har sin egen definisjon av kategorier.

Det er en tillhørende json fil til hver csv fil.

Json filen har flere items hvor hvert item er en kategori med en id tillsvarende category\_ID I csv filen.

Oppsett:

Csv filen er delt opp i rader der hver rad er en video , og hver rad har flere felter som representerer info om videoen.



Video id er id for videoen, denne er unik for hver video og er det siste feltet i en youtube video adresse.

Video adresse struktur: <https://youtube.com/watch?v=> + video id feks «rYy0o-J0x20»

Dette blir en nettadresse til videoen.

Trending date: datoen videoen ble markert som trending.

Title: tittel på video.

Channel title: navnet på kanalen video ble publisert av.

Category Id : nummer på id som samsvarer til en kategori i Json fil.

Publish time: dato og klokkeslett videoen ble publisert. Noter at denne er forskjellig fra trending dato ettersom en video kan være 2 uker gammel før den blir trending.

Tags: tagger en video har , dette blir lagt til under publiseringen av video. Feks «sports» «gaming» etc.

Views: antall ganger videoen har blitt spilt av, ikke det samme som antall personer har sett den.

Likes: antall likes videoen har fått.

Dislikes: antall dislikes videoen har fått.

Beskrivelse av nettsiden:

Nettsiden vi har designet har to views der data blir vist.

Et view for å vise videoer som er trending den daværende uke , og et annet view som viser et kake-diagram over inndelingen av totale views fordelt på de forskjellige kategoriene youtube bruker til å kategorisere data.

I trending video view’et skal det også være mulig å klikke på videoen og bli henvist til youtube videoen på [youtube.com](http://youtube.com).

Dataen her trenger ikke å bli aktivt manipulert ofte , men ved noen tilfeller vil dette bli nødvendig, feks youtube legger til en ny kategori eller hvis man ønsker å vise data fra et annet kontinent.

I



d

Datamodellene våre:

trending\_this\_week:

{

"\_id":"trending\_this\_week":"2020-11-9"

"videos" :

[

{"referance" :"n1WpP7iowLc", "date":"2020-11-10"},

{"referance" :"0dBIkQ4Mz1M", "date":"2020-11-10"},

{"referance" :"5qpjK5DgCt4", "date":"2020-11-14"},

]

}

Categories\_views:

{

1:{"discription" : "Film and animation", "views" : "2 000 000 0" },

2:{"discription" : "Autos and vehicles", "views" : "3 000 00" },

3:{"discription" : "Music", "views" : "500 000 000" }

}

Begrunnelse av datamodellene våre: