

Forprosjektrapport

Gruppen består av fire dataingeniør-studenter fra OsloMet: Atika Gondal, Pedram Rahdeirjoo, Tsegazab Tesfay og Erik Snilsberg

Oppgaven går ut på å forbedre søknadsprosesser ved hjelp av maskinlæring. Det vil si at vi skal bruke maskinlæring til å filtrere og sortere teksten i søknader.

Oppgaven ble tildelt av Innovasjon Norge. Kontaktpersonen vår i selskapet er Håvard Andersen.

Intern veileder på OsloMet er Jianhua Zhang.

Vi har holdt et møte med både eksternt og intern sensor tilstede, hvor vi har gått nøyere gjennom oppgaven og sørget for at alle har samme informasjon og vet hva som er planen videre.

Sammendrag

Innovasjon Norge har lyst til å bli en mer datadrevet organisasjon, og har også lyst til å ta i bruk maskinlæring. Oppgaven går ut på å forbedre prosesser ved hjelp av maskinlæring. For at dette skal bli en realitet må vi få oversikt over tilgjengelig data, som vi skal bruke til å trene modellene. Vi må ta i bruk ulike maskinlæringsmodeller ut ifra hvilke deler av bedriften de ønsker å ta i bruk dette verktøyet. Vi har, sammen med oppdragsgiver, bestemt oss for å bruke programmeringsspråket Python 3, fordi dette egner seg for maskinlæring. Vi vil bruke azure devops for planlegging og versjonskontroll. Det er foreløpig usikkert hvor mye data vi har tilgang til og vi må kanskje jobbe med mindre datasett. For gruppen så vil de neste ukene dreie seg om møter med IN sitt analyse-team og vi må også gjøre oss kjent med verktøyene vi skal bruke.

Dagens situasjon

Innovasjon Norge har som sagt lyst til å bli en mer datadrevet organisasjon og vil se på muligheter for bruk av maskinlæring i bedriften.

De beskriver seg selv som Norges største bank og tilbyr støtte i form av lån til gründere og oppstartsbedrifter med nyskapende ideer. For å motta lån og støtte må bedriftene sende søknader til IN og disse søknadene må dermed prosesseres og vurderes av IN sine rådgivere. Innovasjon Norge er interessert i å forbedre søknadshåndteringsprosessen ved å automatisere deler av tekstprosesseringen. Dette er en mulig anvendelse av maskinlæring innad bedriften. Målet er at et slikt verktøy skal kunne hjelpe rådgivere med å håndtere søknader på en mer effektiv måte, for eksempel ved at verktøyet gjør en tekstanalyse som rådgiverne får tilgang til.

For å kunne bruke maskinlæringsmetoder på søknader må vi ha tilgang på store mengder data i form av søknader. Innovasjon Norge har mye slik data, men på grunn av blant annet sensitiv informasjon er det ikke sikkert hvor mye av denne dataen som er tilgjengelig for oss. I løpet av de neste ukene vil oppdragsgiver sette opp møter med gruppen vår og analyse-teamet til IN for å fastslå hvilken data vi har tilgang til og hvordan vi kan bruke den. Foreløpig vil vi jobbe med mindre datasett.

Mål og rammebetingelser

Oppgaven ligner mer på en forskningsprosess enn utvikling av et system. Gruppen skal finne ut hvilke maskinlæringsmetoder som egner seg for tekstprosessering og fordeler og ulemper ved forskjellige metoder. Vi skal bruke programmeringsspråket Python 3 med PyCharm/Spyder som IDE eller Jupyter som web-basert IDE. Dette språket egner seg fordi det finnes mange biblioteker relatert til maskinlæring for språket. Det overordnede målet er å kartlegge mulighetene for bruk av maskinlæring til å forbedre søknadshåndteringsprosessen ved Innovasjon Norge. Vi skal også bruke azure devops, dette er et verktøy for planlegging og versjonskontroll som er mye brukt innad Innovasjon Norge.

Løsninger /alternativer

Så langt går løsningsmodellen ut på å implementere forskjellige maskinlæringsmodeller og trene modellene på et datasett av søknader gitt av oppdragsgiver. Treningen av modellene kan gjøres på en maskin hos Innovasjon Norge eller på private maskiner med tilstrekkelig kraftfulle skjermkort. Vi skal sammenligne de forskjellige modellene for å se hvilke metoder som egner seg best til tekstprosessering og analyse av søknader.

Analyse av virkninger

Testing av de ulike modellene kan ta tid, men vil være en nødvendig prosess for å finne ut hvilke modeller som passer best til vår løsning. Som sagt vil vi kanskje møte utfordringer med tanke på tilgjengelig datasett for trening av maskinlæringsmodellene. Mindre datasett vil gjøre modeller mindre nøyaktige, som igjen kan gjøre det vanskeligere å produsere nyttig informasjon for selskapet. Mindre datasett vil altså potensielt gjøre det vanskeligere å finne ut om maskinlæringsmetoder kan lønne seg i sektorene til selskapet som vi ser på.