Jobbfinner DAT158

Vegard Rose, Pelle Nielsen

1: BESKRIV PROBLEMET

OMFANG / *SCOPE*

Målet med prosjektet er å lage en jobbanbefalingsplattform. Den skal hjelpe brukeren å finne relevante jobber basert på ferdigheter, erfaring og interesser. Plattformen bruker ML for å analysere store mengder stillingsannonser og finner de stillingene som passer best til brukerens profil.

I dag må jobbsøkere søke manuelt fra nettsider som Finn.no og LinkedIn. Dette kan være tidskrevende og gir ofte irrelevante treff. Ved å bruke tekstbasert ML (TF-IDF + cosine similarity) kan vi automatisk finne likhet mellom brukerens ferdigheter og stillingsbeskrivelser. Den kan returnere personlige anbefalinger på noen få sekunder.

Prosjektet er ment for studenter og arbeidssøkere som ønsker en effektiv og personlig jobb veileder. Modellen kan senere brukes som en modul i et større system.

For å måle ytelse og relevans brukes:

- Cosine Similarity: For å måle likhet mellom brukerprofil og jobbteksten.
- Responstid: Hvor raskt anbefalingene returneres i webappen.
- Brukertilfredshet: Målt gjennom hvor relevante forslagene oppleves.

Minimum suksesskriterium: Systemet skal levere gode og relevante forslag innen 3 sekunder. Den skal klare å foreslå jobber med matchende ferdigheter.

2: DATA

Datasettet som er brukt heter “Job Dataset” (syntetisk, 2024), og er hentet fra Kaggle. Datasettet består av ca. 1,6 millioner jobber fra ulike bransjer og land, generert for forsknings- og utdanningsformål.

Den består av 8 rader med ulik informasjon om stillingen. Vi renset og filtrerte datasettet for å fjerne rader uten nøkkeltekst, og en mindre versjon på 10 000 rader ble brukt i webapplikasjonen for ytelsen sin skyld.

Teksten i datasettet ble behandlet med str.lower(), fjerning av tomme verdier og kombinasjon av relevante kolonner til en tekststreng.

Ettersom datasettet er syntetisk og generert uten personopplysninger, så vil ingen personvernproblemer oppstå. Dette sikrer etiske hensyn.

3: MODELLERING

For å finne relevante jobber ble det brukt en TF-IDF(Term Frequency-Inverse Document Frequency) modell for å representere jobbbeskrivelser og ferdigheter som numeriske vektorer. For å måle hvor nært brukerens input matcher hver jobb i datasettet brukes cosine similarity.

Trinn i modelleringen:

1. Kombinere tekst fra ‘Job Title’, ‘Role’, ‘Skills’, ‘Job Description’.
2. Fjerning av Stoppord og Tokenisering.
3. TF-IDF
4. Beregne cosine similarity mellom brukerinput og jobb.
5. Sorter, gi retur av de 5 mest relevante jobber.

Baseline løsningen uten maskinlæring ville vært et enkelt søk på nøkkelord, men her ville ikke løsningen forstått kontekst eller relevans mellom ord som “developer” og “software engineer”.

4: DEPLOYMENT

Det brukes Streamlit for å implementere modellen og webapplikasjonen, med dette kan brukere skrive ferdigheter og interesser direkte i et tekstfelt og få resultater i sanntid.

Ved deploy:

- Streamlit laster og vektoriserer data ved oppstart.
- Brukerinput transformeres og matches mot TF-IDF matrisen.
- Resultatene vises for brukere med tittel, selskap, lønn og beskrivelse.

Dette systemet er enkelt å vedlikeholde, dersom man skal utvide eller legge til nye jobber kan train_job_model.py kjøres på nytt. Modellen kan også enkelt hostes gratis på Streamlit Cloud.

5: REFERANSER

- Kaggle: Job dataset (Synthetic, 2024) - https://www.kaggle.com/datasets/ravindrasinghrana/job-description-dataset
- Scikit-learn: TF-IDF Vectorizer & Cosine Similarity.
- Streamlit, https://docs.streamlit.io
- OpeinAI. (2025). ChatGPT(GPT-5) [LLM]. https://chat.openai.com