Università degli Studi di Udine

Dipartimento di Scienze Matematiche, Informatiche e Fisiche

Corso di Laurea in Magistrale Comunicazione Multimediale e Tecnologie dell'Informazione

Recommender Systems – A.A. 2024/25

Docente: prof. Kevin Roitero

Selezione intelligente di subset di domande per il benchmarking di LLMs

Studente: Angela Rossi

Matricola: 174288

OBIETTIVI DEL PROGETTO

Il progetto si propone di esplorare strategie per ridurre la dimensione dei benchmark utilizzati nella valutazione dei Large Language Models (LLMs), mantenendo al contempo una buona approssimazione della metrica ottenuta sull'intero set di domande.
L'ipotesi centrale è che, selezionando in modo intelligente un subset rappresentativo di domande, sia possibile ridurre il costo computazionale della valutazione, senza compromettere significativamente l'affidabilità del risultato.

DATASET E SETUP SPERIMENTALE

Il dataset di partenza è un sottoinsieme del benchmark MMLU contenente 3 topic:

- high_school_macroeconomics
- professional_law
- professional_psychology

Ogni voce contiene:

- Il testo della domanda (question)
- Le risposte (choices)
- L'indice della risposta corretta (answer)
- L'embedding della domanda
- L'output del modello e se è corretto (correct)

Per ora il progetto si concentra su un solo modello LLM, valutato tramite la metrica correct.

BASELINE: SELEZIONE CASUALE

Nel primo esperimento ho calcolato la media di risposte corrette estraendo a caso gruppi sempre più ampi di domande (k=1, 2, . . . , 300). L'intervallo k è stato campionato da 5 a 300 con passo 5, per motivi di efficienza computazionale. L'operazione è stata effettuata in modo indipendente per ciascun argomento.

Risultati

I grafici mostrano che:

- La media campionaria converge rapidamente alla media globale
- Già con k tra 50 e 100, la stima diventa molto stabile
- Le curve hanno comportamenti diversi nei vari topic: più rumorose dove la varianza è maggiore

SELEZIONE RAPPRESENTATIVA CON KMEANS CLUSTERING

Come strategia alternativa alla selezione casuale, ho utilizzato KMeans clustering per selezionare subset di domande rappresentative basandomi sugli embedding vettoriali delle domande.

L'idea è che raggruppando le domande in k cluster, sia possibile scegliere la domanda più vicina al centroide di ciascun cluster, ottenendo così un subset che copre in modo ampio e bilanciato il contenuto del dataset.

Metodo
Per ognuno dei tre argomenti: professional_psychology, professional_law, e high_school_macroeconomics, ho applicato l'algoritmo KMeans testando da 1 a 300 cluster; per ciascun k ho ripetuto KMeans clustering 10 volte, individuando la domanda più prossima a ogni centroide in ciascun run, e calcolando la media delle accuracy risultanti. Questa ripetizione permette di mitigare l'effetto della randomizzazione iniziale nei centroidi. Sono quindi andata a calcolare la media della metrica correct sulle k domande così scelte e infine archiviato i risultati, confrontandoli con quelli ottenuti tramite selezione casuale.

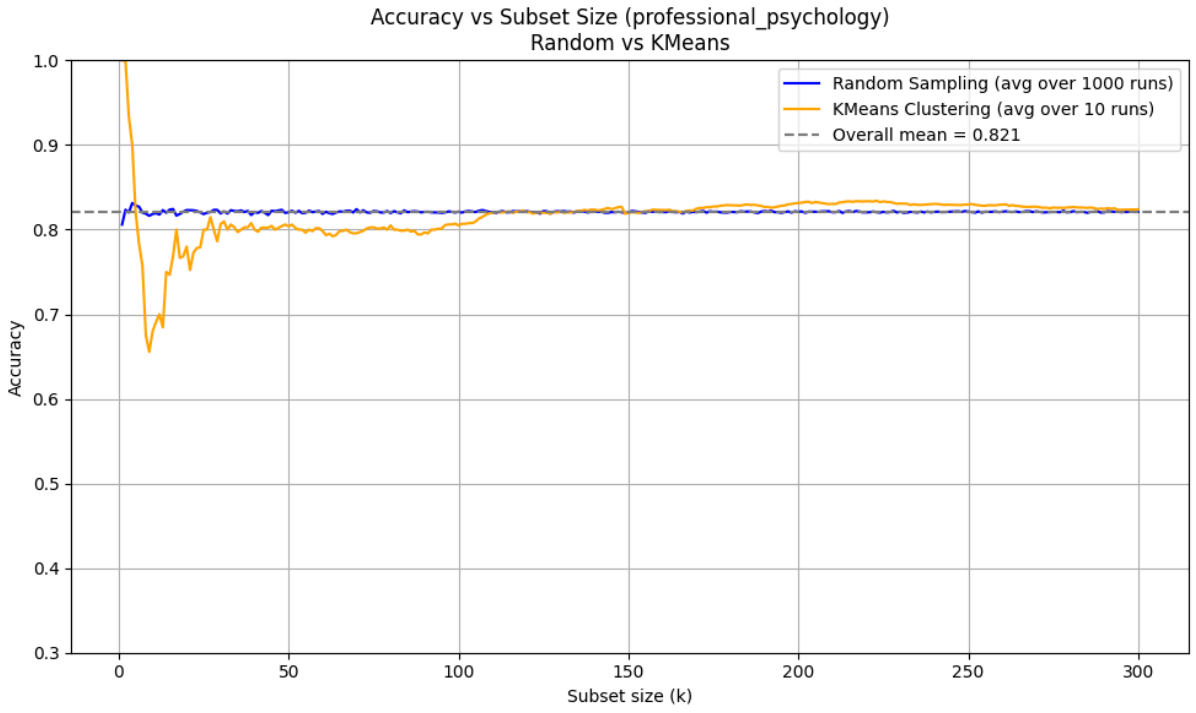
- In tutti i topic, KMeans tende a stabilizzarsi vicino alla media globale, pur con qualche fluttuazione.
- L'approccio fornisce una selezione più strutturata rispetto al caso casuale.
- In alcuni casi (es. professional_law), la convergenza è meno regolare, suggerendo che l'embedding non sempre cattura perfettamente la struttura del dominio.
- La selezione rappresentativa funziona meglio su topic con contenuti più omogenei (high_school_macroeconomics).

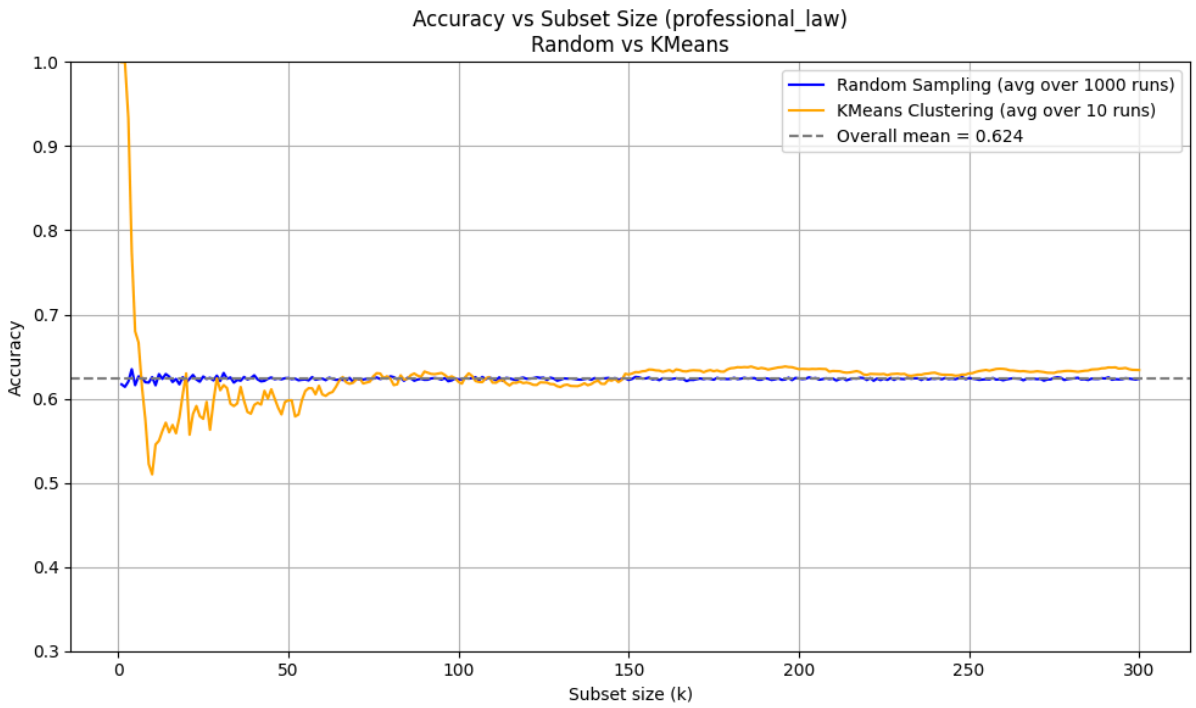
CONFRONTO TRA RANDOM E KMEANS

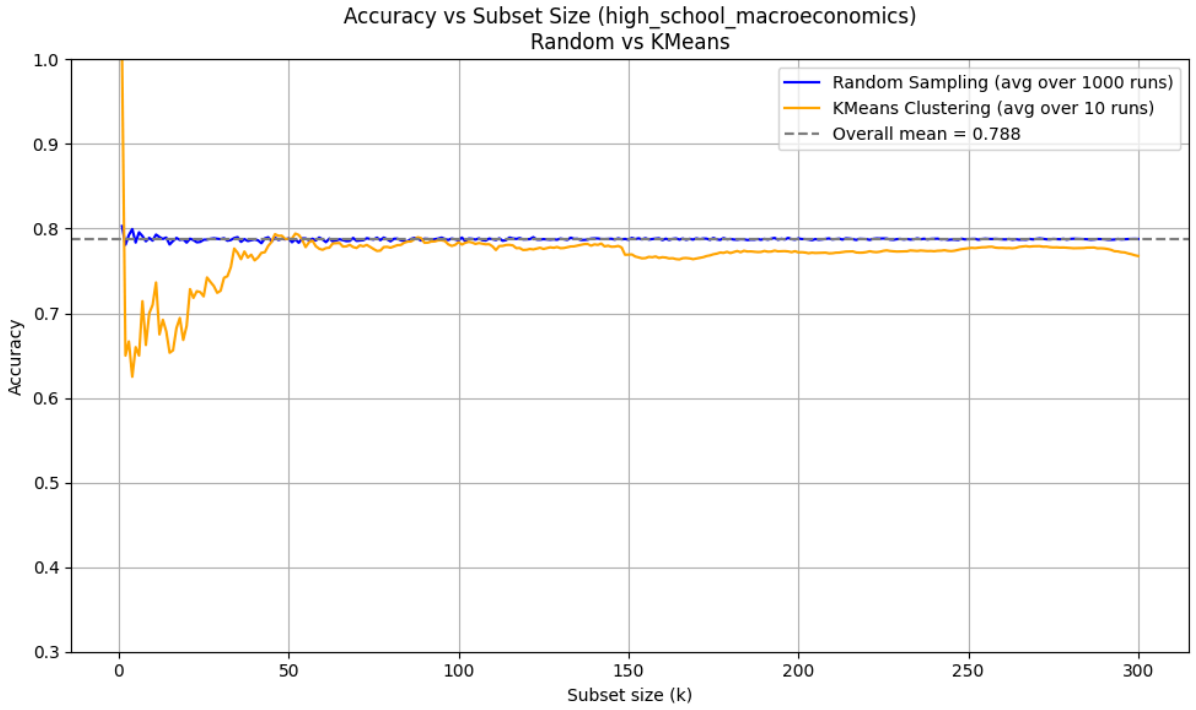
Per paragonare le due strategie, ho tracciato il numero medio di risposte corrette in base alla dimensione del subset, sia per la selezione casuale , in cui si fanno scorrere e si selenzionano randomicamente 1000 volte il set di risposte, sia per la selezione rappresentativa tramite KMeans, dove per ciascun k è stata effettuata una media su 10 esecuzioni con centroidi inizializzati casualmente.

I risultati mostrano che la selezione casuale è in grado di convergere alla media globale con sufficiente stabilità, ma con fluttuazioni più elevate, soprattutto in condizioni di basso k.

KMeans offre una curva più regolare, anche se in casi specifici, come professional_law nella Figura 6B, la convergenza non è lineare per alcune scelte di k; il che suggersi limiti nella capacità degli embeddings di garantire una rappresentazione media semantica. In generale, KMeans è migliore per coprire la varietà semantica presente nei e delle domande, soprattutto in ambiti meno diversificati.







CONCLUSIONI

- La selezione casuale è efficace, ma presenta variabilità elevata a bassi valori di k.

- KMeans offre una strategia più stabile e semanticamente informata.
- La rappresentatività è più facilmente ottenibile in domini più uniformi.
- Le strategie di riduzione del benchmark sono praticabili, e rappresentano un valido compromesso tra accuratezza e efficienza computazionale.
- KMeans si è dimostrato efficace nel fornire una selezione più stabile, ma la sua efficacia dipende dalla qualità degli embedding e dall'omogeneità semantica del dominio

SVILUPPI FUTURI

1. Valutare l’effetto della selezione di subset sull’ordinamento tra modelli diversi (in stile TREC).
2. Esplorare altre tecniche di selezione, come diversity-aware sampling o metodi apprendibili.