Reproducibility Appendix
Analysis of Questions, Winter 2023/4

Kacper Grzymkowski
MSc student
WUT
kacper.grzymkowski.stud@pw.edu.pl

Jakub Fołtyn
MSc student
WUT
01151388@pw.edu.pl

Marceli Korbin
MSc student
WUT
01142124@pw.edu.pl

Mikołaj Malec
MSc student
WUT
01142129@pw.edu.pl

Anna Wróblewska
supervisor
lecturer at WUT
anna.wroblewska1@pw.edu.pl

Reproducibility checklist

Overall results:

- MODEL DESCRIPTION – we consider the second, complexity clustering. It was a clustering based on the embedding representation of the questions along with the DSI metric and the question words present in the question (words such as “what”, “where”, “if”, etc. Those three aspects were combined, meaning that the clustering was done along those 3 axes. This way, we combined both the topic similarity between questions with the indicators of their complexity.
- LINK TO CODE – Github source code: (link), Python requirements list:
  - numpy – version 1.23.5
  - pandas – version 1.5.3
  - tqdm – version 4.64.0
  - matplotlib – version 3.7.1
  - sentence_transformers – version 2.2.2
  - awq – version 0.1.8
  - transformers – version 4.24.0
  - sklearn – version 1.2.1
  - nltk – version 3.7
  - torch – version 1.12.1
  - kmodes – version 0.12.2
  - seaborn – version 0.11.2
  - gensim – version 4.1.2
  - lda – version 3.0.0

  R requirements list:

  - quanteda – version 3.0.0
- INFRASTRUCTURE – scripts can be run using Python 3.9.1. We ran them on a workstation with 11th generation Intel i5 processor and a Nvidia GeForce RTX 3060Ti graphics card.
- RUNTIME PARAMETERS – Split across components
  - SBERT embeddings: 1 minute
  - DSI calculation: 30 minutes (on a sample)
  - Question word extraction: 1 minute
  - Clustering: 1 minute
  - T-SNE: 10 minutes
  - LDA – 3 minutes
  - guidedLDA – 5 minutes
- PARAMETERS – 11 850
- VALIDATION PERFORMANCE – There is consistency in the trends of clustering validation metrics across the majority of 10 random subsets, it suggests robust and stable performance, validating the reliability of the selected number of clusters
- METRICS –
  - Silhouette score: measures cluster cohesion and separation, ranging from -1 (poor) to 1 (good clustering),
  - Inertia: sum of squared distances to closest cluster center.

Multiple Experiments:

- NO TRAINING EVAL RUNS – 19 evaluations of cluster number
- HYPER BOUND – Bounds for each hyperparameter:
  - Number of clusters: range from 10 up to 190 included
- HYPER BEST CONFIG – Hyperparameter configurations for best-performing models: based on Silhouette Method and Elbow Method best cutoff is on 30 clusters.

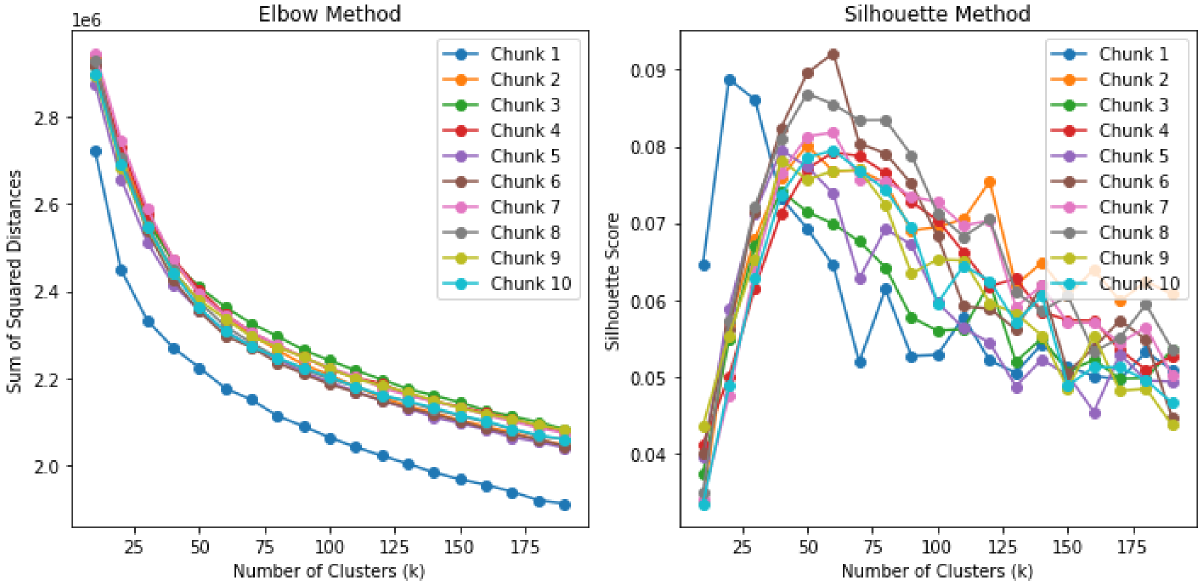

Figure 1: Expected performance

- HYPER SEARCH – Number of hyperparameter search trials: 1.
- EXPECTED PERF – Shown in 1

Datasets – utilized in the experiments and/or the created ones:

- DATA STATS – There are 100,000 questions in the dataset.
- DATA SPLIT – As the problem in the project involved clustering and no labels were present in the dataset, we did not perform a train/test data split. We did, however, took only a random 10% of the data for the project computations, which means that we worked only on 10,000 questions in total.
- DATA PROCESSING – No preprocessing was made on the data.
- DATA DOWNLOAD – Data can be downloaded from this link. It is located in folder *redistribute/QG/train/train.txt.target.txt*
- DATA LANGUAGES –The data is in the English language.