Task 1 *(required)*: Model Evaluation and Selection

Codebase link: https://github.com/iamcalledayush/Task-1

Code Files Description:

- *evaluation_metrics.csv and evaluation_summary.txt contain the results*
- *evaluate_models.py contains the script for evaluating models.*

To run the code:
just run the evaluate_models.py file (it has all the comments to ensure a good understanding of the code). It will generate the results (csv and txt files) after evaluating all the models.

Evaluation Methodology and Chosen Metrics:

Since it's a similarity search-based task, there's no explicit ground-truth labels for each query, I employed a set of 5 metrics to assess for retrieval quality and efficiency (I randomly sampled 1,000 queries, every model is evaluated on the same subset of 1,000 queries, ensuring that comparisons are fair).

1. **Average Top-1 Similarity Score**: This is an obvious metric. For each query, the search engine returns a ranked list of items based on cosine similarity. The top-1 similarity score, i.e., the similarity of the best match is recorded for every query. This metric will serve as an indicator of the quality of the best available match.
2. **Weighted Average Top-5 Similarity Score**: In addition to the top match, I assess the overall quality of the retrieved results by computing a weighted average of the similarity scores for the top five results.
    a. ***Weighting Method***: I applied a decay weighting scheme (initially defined as [0.5, 0.2, 0.15, 0.1, 0.05]) which emphasizes the highest-ranked result while still considering subsequent results.
    b. ***Re-normalization***: In cases where fewer than five results are returned, the available weights are re-normalized so that they sum to 1 again. This adjustment is done because we want the weighted average to remain comparable across queries regardless of the number of returned results.

3. **Query Coverage**: This metric reflects the percentage of queries for which the top result's similarity exceeds a predefined threshold (set at 0.2). High query coverage indicates that the model reliably produces at least one meaningful match for the vast majority of queries.
4. **Average Number of Results Returned**: This is the average number of results returned per query (with a maximum cap of 5). This metric helps understand the model's recall and ensures that both the quality and the quantity of retrieved items are considered.
5. **Average Response Time**: Fast response times are really important to ensure a smooth experience for donor, especially when scaling to large volumes of queries.

Performance Results for Each Model and Model Selection

Eight models were evaluated using a consistent set of 1,000 donor queries. Results were as follows:

1. **Model: infgrad/stella-base-en-v2**
    - Top-1 Similarity: 0.8100
    - Weighted Top-5 Similarity: 0.7873
    - Coverage: 100%
    - Avg. Results: 5.00
    - Response Time: 0.0213 sec
2. **Model: Alibaba-NLP/gte-large-en-v1.5**
    - Top-1 Similarity: 0.7844
    - Weighted Top-5 Similarity: 0.7580
    - Coverage: 100%
    - Avg. Results: 5.00
    - Response Time: 0.0892 sec
3. **Model: jinaai/jina-embeddings-v3**
    - Top-1 Similarity: 0.7650
    - Weighted Top-5 Similarity: 0.7334
    - Coverage: 100%
    - Avg. Results: 5.00
    - Response Time: 0.0827 sec
4. **Model: BAAI/bge-large-en-v1.5**

- Top-1 Similarity: 0.7662
- Weighted Top-5 Similarity: 0.7414
- Coverage: 100%
- Avg. Results: 5.00
- Response Time: 0.0619 sec

5. **Model: nomic-ai/nomic-embed-text-v1.5**
    - Top-1 Similarity: 0.7192
    - Weighted Top-5 Similarity: 0.6852
    - Coverage: 100%
    - Avg. Results: 5.00
    - Response Time: 0.0393 sec
6. **Model: Alibaba-NLP/gte-base-en-v1.5**
    - Top-1 Similarity: 0.7122
    - Weighted Top-5 Similarity: 0.6815
    - Coverage: 100%
    - Avg. Results: 5.00
    - Response Time: 0.0430 s
7. **Model: all-MiniLM-L12-v2**
    - Top-1 Similarity: 0.6198
    - Weighted Top-5 Similarity: 0.5751
    - Coverage: 99.60%
    - Avg. Results: 4.97
    - Response Time: 0.0068 sec
8. **Model: jxm/cde-small-v2**
    - Top-1 Similarity: 0.6081
    - Weighted Top-5 Similarity: 0.5919
    - Coverage: 100%
    - Avg. Results: 5.00
    - Response Time: 0.0362 sec

*The best model according to the results is **stella-base-en-v2.** It has the best Top-1 similarity score and weighted top-5 similarity score, along with a 100% coverage and a decent response time as well.*

Improvement on Baseline

Initially I selected the 4 metrics other than the Weighted Top-5 Similarity and noticed that average results were around 5 for all the models, hence, resulting in 5 most relevant retrievals above the threshold similarity score. Hence, I decided to weigh the top 5 retrievals for each model to gain a better insight on similarity score. The result of this observation was including an extra metric: **Weighted Top-5 Similarity.**

This metric provides a more comprehensive view of a model's performance. It captures not only the best match but also the consistency of the results across the top 5 items. With this additional insight, we can differentiate models that might have a high Top-1 similarity but inconsistent performance on subsequent retrieval results.