Scalable Language Technologies Lab

Summer Term 2026

Exercise From Last Week

- Download the AOL Query Log from the web archive
- Load the AOL Query Log into a pandas DataFrame and experiment with it:
  - What queries did user 711391 issue?
  - How many queries are navigational?
  - Identify search sessions based on time (same user with at most 30 min between queries)
  - Identify query reformulations
  - Filter out search sessions that are purely navigational
  - Other stuff you can come up with :)
- Do not use neural models or embedding models yet
- Prepare a 5–10 min presentation with your findings

Agenda

- Your presentations
- Cluster onboarding [temir.org]

Don’t forget to hand in your presentations, code, and notebooks

Exercise

- Download these prompt logs and identify search sessions
- Create a dataset that contains search sessions from the AOL Query Log and the prompt logs
- Train a classifier on the joint dataset that predicts whether two queries are from the same search session

$$(\text{cute catd}, \text{cute cats}) \mapsto \text{Yes} \qquad (\text{cute cats}, \text{new york}) \mapsto \text{No}$$

- Encode the texts into vectors using DistilBERT (sentence embeddings)
- Implement a “fine-tuning” of DistilBERT: higher dot product means higher probability that two texts appear within the same search session
- Optional: You can use PyTorch Lightning for training
  PyTorch Lightning removes a lot of the boilerplate involved in Deep Learning. But you should write training code from scratch at least once to know how it works (its not hard; just tedious).
- Work on the assignment individually but exchange ideas and problems with each other
- Short 5–10 min presentations

Hints

- Use Hugging Face Transformers to load DistilBERT
- The contextualized embedding of the `[CLS]` token (the first token) is the sentence embedding (as is usual for this model)
- Use contrastive cross entropy
- You can find more information on what we have discussed in the last sessions if you search for “dual encoders” and “in-batch negatives”