PubDictionariesVectorExtension

Participants

- Vincent Emonet
- Jin-Dong Kim

Summary

To extend PubDictionaries for it to benefit from vector representation and computation

PubDictionaries

- A repository of dictionaries.
- Each entry of a dictionary is a mapping between a label and an identifier.
- A dictionary can be uploaded or downloaded as a CSV file
- Once a dictionary is created in PubDictionaries, it will immediately become actionable for dictionary lookup and text annotation.
- Currently, dictionary lookup and text annotation is based on string similarity computation.

Goal

- To add an option to use vector distance computation in the place of string similarity computation.
- Use vector distance computation for
    - Automatic synonym expansion
    - Dictionary lookup
    - Text annotation

Method

- For all existing PubDictionaries:
    - ✅ Generate embeddings (using the FlagEmbedding best small model according to current MTEB benchmarks)
    - ⏳ Load the embeddings in a vector database: using pgVector as PubDictionaries uses the postgres database already
- Visualise first results: https://qdrant.blah.137.120.31.102.nip.io/dashboard#/collections/pubdictionaries-flag/visualize
- Develop scripts for dictionary lookup and automatic synonym expansion
    - In ruby? To better fit PubDictionaries stack