Project 3

顏昌毅 Q56124026
Code repository: https://github.com/changyiyen/iir-BIR
2023-10-31

- A set of 1000 XML documents related to systemic lupus erythematosus was downloaded from PubMed (placed in `corpus/`); another set of 5000 XML documents related to amyotrophic lateral sclerosis was downloaded from PubMed (placed in `als_corpus/`).
- Two attempts were made at the implementation and visualization of word2vec:
  - `word2vec_pytorch.py`: derived from Olga Chernytska's word2vec implementation (https://github.com/OlgaChernytska/word2vec-pytorch)
    - Dependencies: PyTorch (for building/training model itself), Numpy, BeautifulSoup (for XML parsing), NLTK (for tokenization)
    - Configuration uses `config.json`: contains model type and parameters, training data directory, etc.
      - "filetype" parameter: if set to "pubmedxml", will perform training/validation split and tokenization of input XML files (as listed in the "filelist" parameter), then save results to TSV files (`trainingcorpus.tsv` and `validationcorpus.tsv`); when set to something else, will begin model training and validation
    - Results will be saved to `weights/`, including `state_dict`
    - However, results are not easily visualized; started over
  - `word2vec_gensim.py`: uses Gensim package for word2vec fitting
    - Dependencies: Gensim (for building/training model itself), Numpy, BeautifulSoup (for XML parsing), NLTK (for tokenization), Scikit-learn, Pandas, and Matplotlib (for t-SNE and PCA visualization)
    - Configuration uses config_gensim.json: contains model parameters, training data paths, and paths to t-SNE and PCA graphs
    - Both CBOW and Skipgram models were trained
- Web interface mostly the same as in Project 2, but changed to include t-SNE and PCA graphs
  - Structured as a Flask application (search.py) with templates in `templates/` and image files in `static/`
  - Dependencies: Flask (for web interface), Gensim (for reading pre-trained model)
  - Shows cosine similarities and visualizations (t-SNE and PCA)