COMP6714 Project-1 (Part 2)

Amritesh Singh
z5211987

Implementation Details of Q2

- The InvertedIndex class created in Q1 is used for constructing the TF-IDF index for the documents contained in the *men_docs* dictionary.
- <*mention, candidate_entity*> pairs are created for each mention word in the training data and each of its corresponding candidate entities.
- The following features are generated for each of these pairs:
    - **TF-IDF of candidate entities in the mention documents** (obtained from index created from InvertedIndex)

    - **IDF of candidate entities in the mention documents** (obtained from index created from InvertedIndex)

    - **TF of candidate entities in the mention document** (obtained from index created from InvertedIndex)

    - **TF of mention word's tokens in the candidate entity's parsed page**

    - **Okapi BM25**

      $$\mathrm{BM25}(d,q) = \sum_{i=1}^{M} \frac{\mathrm{IDF}(t_i) \cdot \mathrm{TF}(t_i,d) \cdot (k_1 + 1)}{\mathrm{TF}(t_i,d) + k_1 \cdot \left(1 - b + b \cdot \frac{\mathrm{LEN}(d)}{\mathrm{avdl}}\right)};$$

      Here, $k_1 = 1.5$, $b = 0.75$, *avdl* is the average document length in *men_docs*, and *LEN(d)* is the length of the mention word's document.

    - **Language model**

      $$p(t_i\,|\,d) = (1-\lambda)\frac{\mathrm{TF}(t_i,d)}{\mathrm{LEN}(d)} + \lambda p(t_i\,|\,C),$$

      Here, $\lambda = 0.5$, *LEN(d)* is the mention word's document length, and $p(t_i|C)$ is the background language model for t_i.

    - **Cosine similarity**

      $$\cos(\theta) = (e \bullet m) / \sqrt{(e \bullet e)}\sqrt{(m \bullet m)}$$

      Here, *e* is the TF-IDF for the candidate entity, and *m* is the TF-IDF for the mention word

- The training data (containing the features for each pair), the training labels (1 if candidate entity is the label for the mention, 0 otherwise), and the training groups (the size of candidate entities list for each mention) are then converted to XGBoost's DMatrix form.
- The XGBoost classifier is then trained using this data. The classifier has the following parameters: ***'max depth': 8, 'n_estimators': 5000, 'eta': 0.05, 'silent': 1, 'objective': 'rank:pairwise', 'min_child_weight': 0.02, 'lambda': 100***. The ***'num_boost_rounds'*** parameter is set to ***5000***.
- This classifier is then used to predict the values for the test data, and for each test group, the candidate entity with the maximum predicted value is set as the label for the corresponding mention word.
- The dictionary containing the mention words' IDs and their corresponding entity labels is finally returned by the function.