Candidate: Rohit Rangarajan

Date: 24 March 2024

Approach:

Train a named entity linking model from scratch using spaCy and the data provided.

Why I chose this approach:

- The NEL problem is new to me but since I have used SpaCy library for other NLP tasks in the past, I decided to use it to solve this problem

Improvements that I would have done if I had more time

- clean up the companies file
  - add descriptions for companies wherever description was empty (if allowed, we can use services like CoreSignal API)
  - remove URLs from company names
    - use this regex to detect such cases in the name field

      .+ (.*\.(com|io|vc))

  - clean descriptions wherever junk info was there
- gather aliases/synonyms for company names (if allowed, we can use services like Seravia API)
  - add to knowledge base
  - this could improve model performance
- gather more annotated data for training
- try using some transformer model as base model within spacy instead of en_core_web_lg
- perform more detailed hyperparameter tuning when training the EL model
- stratify the train and test dataset based on entity id, so that the model can learn uniformly well across entity IDs

Observations

- some URLs in gold data not found in companies list
- different companies with same URL found in gold data
- same company name, but different URLs in companies file, e.g. Endeavor and a few others
- same company name, but different URLs in annotated file, e.g. "uber.com" and "uber.com/de/en" for company Uber
  - this was a challenge because, as part of my algorithm, to create the training data, I obtain the entity ID by matching the URL in companies list with the URL given in annotated data

    I am using URL and not name, because I found all URLs in the companies list to be unique, but found duplicates in company name
- "description" field has junk info e.g. "cloud-data_crunchbase_2011 worthy Appin tweetprocesor stanford group.pdf."
- for 165 companies, "description" field is empty. We need to collect those descriptions
- many cases where "name" field has URL, e.g. "name": "Andreessen Horowitz a16z.com"
- "name" field has special chars, e.g. "name": "Alb\u00e9a Group"

Steps to reproduce the solution:

1) Upload the solution.ipynb notebook to colab
2) Upload the data files

    a. news_articles-gold.jsonl
    b. news_articles-new.jsonl
    c. news_articles-linked.jsonl
    d. company_collection.json

3) Run all cells in the notebook