Machine Learning Engineer | Assessment task

Task: Named Entity Linking (NEL)

Data description

As part of this task, you can find the following three datasets:

1. **news_articles-gold.jsonl**: A set of 40 news articles with one article per line formatted as JSON. In each article you can find the following fields:
   a. **title**: the title of the news article.
   b. **text**: the full-text article as extracted from the news website.
   c. **source**: link to the original article.
   d. **annotations**: a dictionary of companies that appear in the article. Given as a dictionary with the company name as it is mentioned as keys, and their identifying URLs as values.
2. **news_articles-new.jsonl**: A set of 60 news articles in the same format as described above but without the **companies** field.
3. **company_collection.json**: A collection of companies in JSON format with the following fields:
   a. **url**: URL of the company website, which is a unique identifier.
   b. **name**: the name of the company.
   c. **description**: a short text describing the company.
   d. **founded**: the year the company was founded.
   e. **headquarters**: the location of the company (country, state, city).
   f. **industry_label**: one or multiple industry labels assigned to this company (separated by |).

Task description

Your task is to process the news articles from the **news_articles-new** dataset, identify and link potential companies that appear in these articles. The result should be a file called **news_articles-linked.jsonl** in the same format as the **news_articles-gold** dataset. The company identifiers should be taken from the **company_collection** but if you find companies that are not part of the collection you can add them with an empty string as value, e.g. companies: {'Apple':'apple.com', 'MalWart': ''}. The solution should be implemented in Python and you are free to use any library or framework of your choice, as long as we can easily reproduce your results. **Note**: Named Entity Recognition (NER) is a part of this task, however the main task is still Named Entity Linking (NEL).

Deliverables

This problem can be solved in a lot of different ways and solutions can become quite complex. We do not expect a solution with near state-of-the-art results but are more interested in how you approach the problem with regards to your knowledge and creativity.

Alongside the result file (“news_articles-linked.jsonl”), please also include the source code and requirements so that we can reproduce the environment and results. This can be in any format, but feel free to return it under one that makes it the most suitable to reproduce your results.

In addition, please also include a document with a short explanation of why you have chosen this approach, what challenges you found and how the solution could potentially be improved in the future if you had more time and resources.

Task duration

The duration of this task should not extend **3 to 4 hours**. Feel free to spend more time on it but as mentioned above we do not expect overly complex implementations.