CS586: Data & Web Semantics
Assignment 1- Kshitij Gautam

Logic

1. Strategy & Algorithm (Link)
   a. Doing Reverse Engineering
      i. By saving the reference alignment `ref.saveRDF("./matcher.rdf");)` and compariing with the given RDF.
      ii. By checking what's matching and what's not matching
         1. Output is in (Correct_Wrong Mappings/all experiments)

   b. Experiments with Threshold (link)
      i. I experimented with threshold values and found that 0.71 was giving best results after manually reading (given rdf and my Matcher's generated rdf - **matcher_1.rdf**).
   c. **Experiments with Wordnet**
      i. Are written in Log file
      ii. I calculated similarity ration (number of common hypernyms/ total hypernyms) and returned Maximum of word, string and hypernym similarity
         1. `return (intersection.size()) / (n1Words.size() + n2Words.size());`
   d. Experiments with Existing Matchers
      i. I tried all the matchers(code is commented out). Few showed promises, some were very fast and some were very slow. You can read the experiments in the **Experimental_logs.txt**
2. Code and History of my commits is here
   a. https://github.com/gautamkshitij/semanticWeb.
   b. Given **MyMatcher.java**
      i. Please see Readme for more details.
3. Examples of Correct Mapping
   a. All correct mappings are stored in, An Example is shown in figure below.
      i. Correct_Wrong Mappings/all experiments/1. Correct_mapping.txt

4. Examples of Incorrect Mapping
   a. All wrong mappings are attached. Few examples are shown here.
   b. Correct_Wrong Mappings/all experiments/*wrongMappings.txt.*

```
3021    <map>
3022        <Cell>
3023            <entity1 rdf:resource="http://big.csr.unibo.it/
                sabine-eng.owl#European_Constitution"/>
3024            <entity2 rdf:resource="http://dbpedia.org/resource/European_Union"
                />
3025            <measure rdf:datatype="http://www.w3.org/2001/XMLSchema#float">
                0.7353</measure>
3026            <relation>=</relation>
3027        </Cell>
3028    </map>
3029
3030    <map>
3031        <Cell>
```

```
1970             <measure rdf:datatype="http://www.w3.org/2001/
                 XMLSchema#float">0.927776319526</measure>
1971        </Cell>
1972    </map>
1973    <map>
1974        <Cell>
1975            <entity1 rdf:resource="http://big.csr.unibo.it/
                sabine-eng.owl#EU"/>
1976            <entity2 rdf:resource="http://dbpedia.org/
                resource/European_Union"/>
1977            <relation>=</relation>
1978            <measure rdf:datatype="http://www.w3.org/2001/
                XMLSchema#float">0.766559621914</measure>
1979        </Cell>
1980    </map>
```

Results

- Final Result

Precision	Recall	F-measure	Found	Correct	Reference
90.6%	82.8%	86.6%	309	280	338

- Other Results and Logs
  - Are in **./Experiment_log.txt** and below.

Experiment Logs

- All experiments conducted using AML software are located in Readme file of the project. I have experimented with different combinations of matchers and different weights. The experiments were logged from the start of the project.
- **./Experiment_log.txt**

Reverse Engineering of the Reference Alignments

1. **Work Logging**
   a. **Reverse Engineering**
      i. I have modified the code and printed all the correct and wrong mappings now.

         example:
         Inflation is twice (correct and in wrong),
         Correct: Inflation—> Inflation
         Wrong: Inflation—> Inflation_cosmology (in the golden set Inflation is only once)
         so trying to get the Best Name instead of getting Set of names.

         ```
         names_Similarity = nameSimilarity(sLex.getBestName(i1), tLex.getBestName(i2), true);
         wu_pal = wn.wuPalmerScore(sLex.getBestName(i1), tLex.getBestName(i2));
         ```

Precision	Recall	F-measure	Found	Correct	Reference
93.6%	78.4%	85.3%	283	265	338

         → High Precision achieved, but 0.1% in F-1 Measure
      ii. I removed the wu_palmer similarity (significant F1 measure than Baseline)

Precision	Recall	F-measure	Found	Correct	Reference
91.1%	82.2%	86.5%	305	278	338

      iii. Experimenting with this setting and changing threshold (from 0.729 to different values
         1. Threshold: 0.9 (Precision: 93.5% and F: 81%)
         2. Threshold: 0.8 (P: 92, F: 85%)
         3. **Threshold: 0.71 (P 90.6%: F 86.6%)** _**better than 0.729**_
         4. Threshold: 0.6 (P: F:)

5. T: 0.65 (P: 89.5, F: 86. 4)
6. Threshold: 0.5 (P 84%, F 84.8%).

Precision	Recall	F-measure	Found	Correct	Reference
90.6%	82.8%	86.6%	309	280	338

    iv.

    v.   Analyzing Wrong mappings vs Correct Mappings
       1. Duplicate Mappings
          a. Genetically-modified_organism couldn't match

  vi.   **All previous logs are in Experimental_logs.txt**

2. Extra Code
   a. Printing Correct/Incorrect Mappings (instead of doing manually)

```java
public ArrayList<Mapping> printWrongMappings(Alignment ref, int whichOne) {

    if (whichOne > 2 || whichOne < 0) return null;
    //0 – wrong, 1– correct, 2 – unkown
    ArrayList<Mapping> askedMappings = new ArrayList<>();
    for (Mapping m : maps) {
        if (ref.contains(m)) {

            if (whichOne == 1) askedMappings.add(m);


            m.setStatus(MappingStatus.CORRECT);
        } else if (ref.contains(m.getSourceId(), m.getTargetId(), MappingRelation.UNKNOWN
            if (whichOne == 2) askedMappings.add(m);
            m.setStatus(MappingStatus.UNKNOWN);
        } else {
            if (whichOne == 0) askedMappings.add(m);
            m.setStatus(MappingStatus.INCORRECT);


        }

    }
    return askedMappings;

}
```

   b.
   c. Output
      i. Please see attached
    Correct_Wrong Mappings/Experiment_4_gettingBestName_threshold_0.71/.txt*