Comments and answers - Assignment 3

Project Text Analysis

Leon Wetzel, Teun Buijse and Roman Terpstra

Exercise 1

Part 3

```
Path Similarity

('car-automobile', 1.0)
('coast-shore', 0.5)
('monk-slave', 0.2)
('moon-string', 0.1111111111111111)
('food-fruit', 0.09090909090909091)
('journey-car', 0.05)

Leacock-Chodorow Similarity

('car-automobile', 3.6375861597263857)
('coast-shore', 2.9444389791664407)
('monk-slave', 2.0281482472922856)
('moon-string', 1.4403615823901665)
('food-fruit', 1.2396908869280152)
('journey-car', 0.6418538861723948)

Process finished with exit code 0
```

As you can see the ranking is quite different, mostly when it comes to medium-level similarity pairs. The similarity measures in the NLTK corpus reader do not consider these to be very similar, while the subjects did, in our eyes, succesfully recognize that these are more similar than monk-slave for example. This could be explained by the fact that car-automobile, for example, can be used interchangeably. While food-fruit have very different meaning but belong to the same category. ## Exercise 2

Part 1

To start the server, we use the following bash command:

```
shell script $ java -mx4g -cp "*" edu.stanford.nlp.pipeline.StanfordCoreNLPServer  -port 9000 -
timeout 15000 -serverProperties server.properties
```

We count 2 ORGANIZATION's, 1 LOCATION and 20 PERSON's. Not all the shown named entities are correct. For example, *Augusta Ada King* and *Ada* are named as ORGANIZATION, which is not correct in this particular context.

Part 2

There appear to be 6 other models for named entities. We take a closer look at the 4 and 7 classes type of models.

- `english.conll.4class.distsim.prop`
- `regexner.patterns`
- `english.muc.7class.distsim.crf.ser.gz`

- `english.conll.4class.distsim.crf.ser.gz`
- `english.all.3class.distsim.prop`
- `english.muc.7class.distsim.prop`

Let's start off with the 4 classes. We alter `server.properties` by replacing the old value of `ner.model` with `edu/stanford/nlp/models/ner/english.conll.4class.distsim.crf.ser.gz`. We then feed the text file to the application, which returns more ORGANIZATION labels, although one can conclude that more often than not a ORGANIZATION label is not properly used in this context. In addition, the MISC label is introduced, being a label often present at indications of nationality.

Let's use the 7 classes model. The `ner.model` in `server.properties` will be changed to `edu/stanford/nlp/models/ner/english.muc.7class.distsim.crf.ser.gz`. We can see that next to our familiar named entities, we also see a new named entity DATE. The ORGANIZATION entity is also more present than the original model.

Part 3

For this part of the exercise, we use the `edu/stanford/nlp/models/ner/english.conll.4class.distsim.crf.ser.gz` model.

```
C:\Users\leonw\Documents\PTA\venv\Scripts\python.exe
         C:/Users/leonw/Documents/PTA/week3/exercise2.py
countess O
independence O
Lord O
Countess ORGANIZATION
Ada ORGANIZATION
Byron ORGANIZATION
question O
attempt O
disease O
England LOCATION
augusta O
Ada PERSON
Byron PERSON
bent O
byron O
club O
instrument O
Lord LOCATION
others O
December O
earth O
baron O
king O
working O
Isabella PERSON
programmer O
lord O
godhead O
Engine ORGANIZATION
engineer O
algorithm O
history O
mother O
War ORGANIZATION
friendship O
overlord O
Greek ORGANIZATION
employment O
relationship O
end O
mathematics O
```

approach O
Luigi PERSON
concern O
charles O
year O
class O
November O
oeuvre O
greek O
poet O
adenosine O
forefather O
endowment O
engine O
science O
King ORGANIZATION
skill O
work O
Babbage PERSON
england O
world O
anne O
pastime O
talent O
interest O
relate O
Independence ORGANIZATION
cock O
article O
sight O
access O
person O
feat O
logic O
study O
request O
set O
note O
church O
machine O
society O
marriage O
kinship O
imagination O
individual O
insanity O
november O
overture O
charlemagne O
father O
Anne PERSON
Between ORGANIZATION
engineering O
analyst O
_ O
wedlock O
computer O
child O
Lovelace ORGANIZATION
Biography O
month O
war O
doubt O
vision O
deaminase O

```
wife O
stage O
creature O
Lovelace PERSON
programâ O
lovelace O
company O
eminence O
locomotive O
car O
december O
technology O
calculator O
mathematician O
workplace O
Analyst O
motion O
writer O
sake O
mentality O
mind-set O
calendar O
bill O
chiefly O
tool O
Metaphysician MISC
death O
Charles PERSON
capability O
Menabrea PERSON
universe O
path O
isabella O
don O
Ada LOCATION
Byron LOCATION
Analytical ORGANIZATION
adult O
founder O
Notes O
€ O
populace O
'' O
hardening O
== O
campaign O
worldly O
effort O
Augusta ORGANIZATION

Process finished with exit code 0
```