Task 1: Natural Language Processing

The following text was extracted from the Wikipedia page about Edward Snowden[1].

> In 2013, Snowden was hired by an NSA contractor, Booz Allen Hamilton, after previous employment with Dell and the CIA. On May 20, 2013, Snowden flew to Hong Kong after leaving his job at an NSA facility in Hawaii, and in early June he revealed thousands of classified NSA documents to journalists Glenn Greenwald, Laura Poitras, and Ewen MacAskill. Snowden came to international attention after stories based on the material appeared in The Guardian and The Washington Post.

1.1 Insert the text from Task 1 into the Stanford Core NLP demo: http://corenlp.run/

1.2 What types of output are returned?

- *Part-of-Speech*
- *Named Entity Recognition*
- *Basic Dependencies*
- *Enhanced++ Dependencies*
- *Open IE*

1.3 What is the difference between Named Entity Recognition (NER) and Entity Linking (EL)?

In named entity recognition, named entities are recognized and typed, but not linked to a knowledge resource. Example:

- *NER: Snowden (PER)*
- *EL: https://en.wikipedia.org/wiki/Edward_Snowden*

Task 2: Relation Extraction

2.1 Use Ollie to annotate the text from Task 1: https://github.com/knowitall/ollie (Section "Local Machine").

Output

```
- 0,903: (Snowden; was hired by; an NSA contractor)
- 0,771: (Snowden; was hired in; 2013)

- 0,909: (Snowden; flew to; Hong Kong)
- 0,807: (he; revealed; thousands of classified NSA documents)
- 0,789: (Snowden; revealed; thousands of classified NSA documents)
- 0,771: (his job; be leaving at; an NSA facility)
- 0,728: (Snowden; flew on; May 20 , 2013)
- 0,58: (he; flew to; Hong Kong)
- 0,55: (thousands of classified NSA documents; be revealed in; early June)
```

[1] https://en.wikipedia.org/w/index.php?title=Edward_Snowden&oldid=836189168

- 0,432: (he; flew on; May 20 , 2013)
- 0,425: (Snowden; revealed thousands of classified NSA documents to journalists Glenn Greenwald , Laura Poitras in; early June)
- 0,366: (Laura Poitras; be thousands of classified NSA documents to; journalists Glenn Greenwald)
- 0,288: (he; revealed thousands of classified NSA documents to journalists Glenn Greenwald , Laura Poitras in; early June)

- 0,925: (Snowden; came after; stories based on the material appeared in The Guardian and The Washington Post)
- 0,92: (Snowden; came to; international attention)
- 0,655: (stories; be based on; the material)

2.2 Describe the output and list the most important advantages and disadvantages of relation extraction.

Important relations in the text are extracted as triples.
Advantages:
- *The relations are not limited to any pre-defined constraints.*
- *Context information is extracted "(he; flew to; Hong Kong)" and "(he; flew on; May 20 , 2013)".*
- *Each triple comes with a confidence score.*

Disadvantages:
- *Named entities are not linked and the relation between them is just presented as terms.*
- *All relations connect relations with exactly two concepts involved.*
- *The relations are not put together, e.g. to connect "(he; flew to; Hong Kong)" and "(he; flew on; May 20 , 2013)".*

Task 3: DBpedia Languages

3.1 What will the following query return when executed on the English DBpedia?

```
SELECT LANG(?label) WHERE {
      ?sub rdfs:label ?label .
} GROUP BY LANG(?label)
```

Query result: ja, el, en, sr, pt, nl, de, …
The query returns all languages for which there is at least one entity in the English DBpedia that has a label in that language.

3.2 Rewrite the query to add the frequency of labels per language.

```
SELECT LANG(?label), COUNT(*) WHERE {
      ?sub rdfs:label ?label .
} GROUP BY LANG(?label)
```

3.3 Execute your query from Task 3.2 on https://dbpedia.org/sparql (English DBpedia) and https://fr.dbpedia.org/sparql (French DBpedia). List the three most frequent languages per DBpedia. What do the results tell us about the use of languages in different DBpedia language versions?

Top three languages in the English DBpedia:
en 4862847
de 1020378
fr 709837
Top three languages in the French DBpedia:
fr 3267206
en 989974
de 516788
These results demonstrate that the different DBpedia language editions are extracted from different Wikipedias and that there is not integration between them, e.g. to collect all labels.

Task 4: Language Identification

Given are the following sentence *s* and two language profiles:

- sentence *s*: “in may they may go in their therme”
- English: [the, may, in_, by_, and, one, ing, his, ted]
- German: [sch, her, das, die, bei, ein, in_, the, go_]

4.1 Create the document profile of *s* as follows:

4.1a First, write down all character-based 3-grams of *s*. Ignore 3-grams starting with a blank and those at the start or end of the sentence.

in_, n_m, may, ay_, y_t, the, hey, ey_, y_m, may, ay_, y_g, go_, o_i, in_, the, hei, eir, ir_, r_t, the, her, erm, rme

4.1b Now, sort the set of 3-grams by descending frequency and assign ranks to them. The most frequent 3-gram gets rank 1. If more 3-grams appear equally often, they get the same rank.

Rank 1: the
*Rank 2: in_, may, ay_*
Rank 5: n_m, y_t, hey, ey_, y_m, y_g, go_, o_i, hei, eir, ir_, r_t, her, erm, rme

4.2 Determine the language of *s* based on its document profile and the language profiles as follows:

4.2a Compute the Jaccard similarities between the document profile and both language profiles. Which language is identified for *s* according to this similarity?

English:

$A = \{the, may, in\_, by\_, and, one, ing, his, ted\}, |A| = 9$

$B = \{the, in\_, may, ay\_, n\_m, y\_t, hey, ey\_, y\_m, y\_g, go\_, o\_i, hei, eir, ir\_, r\_t, her, erm, rme\}, |B| = 19$

$A \cap B = \{the, may, in\_\}, |A \cap B| = 3$

$|A \cup B| = |A| + |B| - |A \cap B| = 25$

$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{3}{25} = 0.12$

German:

$A = \{sch, her, das, die, bei, ein, in\_, the, go\_\}, |A| = 9$

$B = \{the, in\_, may, ay\_, n\_m, y\_t, hey, ey\_, y\_m, y\_g, go\_, o\_i, hei, eir, ir\_, r\_t, her, erm, rme\}, |B| = 19$

$A \cap B = \{her, in\_, the, go\_\}, |A \cap B| = 4$

$|A \cup B| = |A| + |B| - |A \cap B| = 24$

$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{4}{24} \approx 0.17$

German language is identified for s when using Jaccard similarity.

4.2b Compute the "out of place" measures between the document profile and both language profiles (with default penalty $K = 10$). Which language is identified for *s* according to this measure?

English:

the: $|1 - 1| = 0$, *may:* $|2 - 2| = 0$, *in_:* $|2 - 3| = 1$

$0 + 0 + 1 + K \cdot 16 = 161$

German:

her: $|5 - 2| = 3$, in_: $|2 - 7| = 5$, the: $|1 - 8| = 7$, go_: $|5 - 9| = 4$

$3 + 5 + 7 + 4 + K \cdot 15 = 169$

English language is identified for s when using the "out of place" measure.

Task 5: Text Similarity

Given are the following three sentences[2] [3]:

s_{E1} (English): *Theresa May has announced plans to call a snap general election for the UK on 8 June.*

[2] http://www.bbc.com/news/uk-politics-39629603

[3] http://www.spiegel.de/politik/ausland/theresa-may-grossbritannien-soll-am-8-juni-vorzeitig-waehlen-a-1143687.html

s_{E2} (English): *Donald Tusk said the 27 other EU states would forge ahead with Brexit, saying the UK election of the parliament would not change their plans.*
s_{G1} (German): Regierungschefin *Theresa May hat angekündigt, dass die Neuwahl des Parlaments in Großbritannien am 8. Juni stattfinden sollen.*

Compute the sentence similarities based on the following three measures:

- **Longest common subsequence:** length of the longest subsequence of characters that appears in both sentences
- **Biword similarity:** Split the text into biwords (word-based 2-grams, ignore commas and dots) and compute their overlap using the **Dice coefficient**.
- **Knowledge-based similarity:** Manually annotate named entities and time expressions. Then compute the knowledge-based similarity as the mean of the Jaccard similarity computed with named entities and the **Jaccard similarity** computed with time expressions.

English Sent.	German Sent.	Longest common subsequence	Biword similarity	Knowledge-based similarity
s_{E1}	s_{G1}	14 “Theresa May ha”	$\frac{2 \cdot 1}{16+16} \approx 0.063$ Common biwords: “Theresa May” Biwords in s_{E1}: *16* Biwords in s_{G1}: *16*	$\frac{1}{2} \cdot \frac{2}{2} + \frac{1}{2} \cdot \frac{1}{1} = 1$ Entity Overlap: $\frac{\lvert\{Theresa\_May, UK\}\rvert}{\lvert\{Theresa\_May, UK\}\rvert}$ Time Overlap: $\frac{\lvert\{8\ June\}\rvert}{\lvert\{8\ June\}\rvert}$
s_{E2}	s_{G1}	5 “ament”	$\frac{2 \cdot 0}{24+16} = 0$ Common biwords: “Theresa May” Biwords in s_{E2}: *24* Biwords in s_{G1}: *16*	$\frac{1}{2} \cdot \frac{1}{5} + \frac{1}{2} \cdot \frac{0}{1} = 0.1$ Entity Overlap: $\frac{\lvert\{UK\}\rvert}{\lvert\{T\_M, UK, EU, Brexit, D\_T\}\rvert}$ (T_M: Theresa May, D_T: Donald Tusk) Time Overlap: $\frac{\lvert\{\}\rvert}{\lvert\{8\ June\}\rvert}$