Assignment 1

Due: March, 02 2018, before 5 PM

Submit via Canvas

Task 1 (10 points)

Install the Stanford CoreNLP pipeline: https://stanfordnlp.github.io/CoreNLP/

If you intend to work on English only, you might want to install both of the English language models: English and English (KBP)

The installation of CoreNLP presupposes an installed Java SE. I use successfully Java 8:
http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html

You can use the startup script in our Canvas code section to launch CoreNLP in server mode on your local computer. Access the interface using a browser and the localhost:9000 URL.

Activate:

- Parts-of-speech
- Named entities
- Dependency parse
- Lemmas
- Constituent parse
- Coreference

Process a sentence pair like:
John met Susan in the mall. She told him that she is traveling to Europe next week.

Identify a task or multiple tasks that might benefit from NLP features.

In the context of the specific task(s), describe your strategies for vectorization or use of:

- Part-of-speech properties (identify the tag-set and discuss the mapping of encoded features to a vector representation)
- Named entities
- Lemmas
- Coreference
- Dependency parse tree
- Constituent parse tree

Be concrete and justify your vectorization strategy.

Think of the detailed linguistic features encoded in each level of NLP-analysis. What might be a relevant feature for the task or application?

Task 2 (10 points)

Install spaCy and the relevant language models: https://spacy.io/usage/

spaCy does not provide the modules Coreference and Constituent parse, but all the other modules. Make sure that you can process the sentence pair mentioned in task 1.

Study the output of the spaCy NLP pipeline and compare it with the Stanford CoreNLP pipeline.

Develop the same vectorization strategy.

Task 3 (10 points)

(I will assume that you will use Python in this task. You can use any other modern and reasonable language, but maybe make sure that I would understand the code.)

Use the Stanford CoreNLP pipeline and study the documentation:
https://stanfordnlp.github.io/CoreNLP/corenlp-server.html

Identify how you can receive for example a JSON object from the server and how you can map it to Python data structures and vectorize those data structures into appropriate Numpy vectors (or arrays).

Use spaCy to generate NLP-objects and vectorize the relevant features as discussed in the two previous tasks.

Write Python (or other programming language code) for the vectorization of the features discussed in the previous tasks.

If you need help or assistance with the installation of Python, modules, or the NLP components, the conversion of output formats to Python data structures, etc., please contact me.

Task 4 (Bonus task, make-up points: 10)

Familiarize yourself with the two modules:

- spaCy vectors:
- fastText

Use both modules for text classification based on vectors and embeddings. Use some collection of texts that you create yourself, restricting yourself to a few classes, two would be fine.

Describe your approach in words and provide code examples with a functional test environment.