Laboration exercise for Training School COST by Hercules Dalianis, Oct 7, 2015.

Build a machine learning based deidentification system for the DEID-corpus
Use the annotated American DEID-corpus, (The gold-standard corpus of de identified medical text) (Physionet, 2013) and build a deidentification system based using CRF Stanford NER (Stanford, 2015).
Compare your DEID results with the results of the rule based system by Neamatullah et al (2008). On Physionet you can download most parts of the information, however you must register on Physionet to get access to the DEID-corpus. When you have obtained access to the DEID-corpus you can contact me Hercules Dalianis at hercules@dsv.su.se, and get the pre-processed corpora for input to CRF Stanford NER.

The annotation classes in DEID are

Patient Name
Patient Name Initial
Relative/Proxy Name
Clinician Name
Date (not year)
Year
Location
Phone
Age over 89
Undefined
and
Overall (all classes collapsed)

In the report please write the evaluation with precision, recall and F-measure, exact match, partial match. Do n-fold cross evaluation or train on one set of the corpus and predict/evaluate on the remaining set of the corpus. Please explain and motivate your choices. How can you improve your results? Discuss also what these DEID-techniques can be used for in real life, are there any drawbacks?

A laboration exercise group must contain maximum two persons, and at least one person.
The lab report must contain at least two pages and contain the names of the group members.
Note that your laboration report should be in PDF format only and be submitted to Ilearn. The lab report will be graded with pass/fail.

Step 1
Contact Physionet and ask for access to the corpus. When you obtain permission and data, ask Hercules Dalianis for pre-processed data.

Step 2
Download Stanford NER.

Step 3
Follow instructions in http://nlp.stanford.edu/software/CRF-NER.shtml
(Prepare by running their example with *chapter 1 of Emma as training* and *chapter 2 as evaluation)*

Step 4
Train on one part of the DEID corpus and evaluate on a different part.

Hint 1: It seems to be very slow to train on the whole DEID corpus, try 2/5 to train and increase slowly to 50/50.

Hint 2: Increase memory,
java -Xmx14g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -prop sample.prop
14g in the switch *Xmx14g* means 14 Gigabyte as memory indication to the java compiler.

Step 5
Try Stanford NER interface *java -mx2000m -jar stanford-ner.jar* with your trained model to evaluate it qualitatively. Paste some parts of DEID corpus to analyze how well it de-identifies and where it misses.

Step 6
Finalise the laboration exercise and write report

References
Physionet. 2013, http://www.physionet.org/physiotools/deid/

Neamatullah I, Douglass M, Lehman LH, Reisner A, Villarroel M, Long WJ, Szolovits P, Moody GB, Mark RG, Clifford GD. Automated De-Identification of Free-Text Medical Records. *BMC Medical Informatics and Decision Making*, 2008, **8**:32. doi:10.1186/1472-6947-8-32

Stanford, 2015, Stanford Named Entity Recognizer (NER), http://nlp.stanford.edu/software/CRF-NER.shtml