Python programming
— exercises for text and web mining

Finn Årup Nielsen

DTU Compute
Technical University of Denmark

September 22, 2014

Word and sentence segmentation

Segment the following short text into sentences and words:

```
>>> s = u"""DTU course 02819 is taught by Mr. Finn Årup Nielsen,
Ph.D. Some of aspects of the course are: machine learning and web
2.0. The telephone to Finn is (+45) 4525 3921, and his email is
faan@dtu.dk. A book published by O'Reilly called 'Programming
Collective Intelligence' might be useful. It costs $39.99 or 285.00
kroner in Polyteknisk Boghandle. Is 'Text Processing in Python'
appropriate for the course? Perhaps! The constructor function in
Python is called "__init__()". fMRI will not be a topic of the
course."""
```

Try both with the `re` module as well as with a function from `nltk`.

Email mining

Change the feature set to less words or other words.

Code available here: https://gist.github.com/1226214

Web extraction

Extract information from the course website of DTU 02819, e.g., "Qualified Prerequisites" from the Course Base 2014/2015 page, and/or grades distribution from Karakterfordeling and/or information from the Course evaluation.