Assignments

All asignments will be uploaded here.

1. `clone` (or `pull` from) the **metu-cogs-520** repository to your machine.
2. Work under the directory named in your id.
3. `commit` and `push` the changes.

Conventions:

- all parts under their dedicated folders, each named **part_X** accordingly (e.g. **part_I** for the first part)
- a folder named **in** for your input files
- a folder named **out** for your output files

Try to make your code accessible, meaning taking good care of stucture, naming, comments etc.

Part I

- [**topic**] preprocessing texts
- [**due**] 4th of Nov.
- [**programming language**] any
- [**input language**] any (preferably the one you will work with in your project)
- [**tasks**] code a basic preprocesser with a similar scope examplified in the online tutorial, which utilizes third party processes when necessary
    1. review various corpora in your target language and find an appropriate one (preferably noisy)
    2. take a (representative) sample from that corpus. keep it small, no more than 5MB.
    3. preprocess the sample. first **(a)** try to correct spelling by utilizing a 3rd party library (e.g. NLTK for English in Python, İTÜ Turkish NLP pipeline for Turkish etc.). next **(b)** eliminate dates, numbers and punctuation (except those marking sentence boundaries). third **(c)** normalize cases except proper names via a (simple heuristic) process that you will engineer. you can also play with other preprocessing tasks, such as stemming.

Part II

TBA