IHLT Laboratory

Jordi Turmo
TALP Research Center
turmo@cs.upc.edu

Session 1

Goal of IHLT lab sessions

- Learn to use basic NLP functions for managing text content
- Solve simple programming exercises

Programming platform: Jupyter (python)

[works saved as notebooks - *.ipynb -]

NLP package for Python: `nltk`

Similar open-source NLP suites out of this framework:
Stanford CoreNLP, Freeling, Apache OpenNLP, IXA Pipes

Evaluation of IHLT lab

- groups of 2 people, although individual works will be accepted
- One project presented in two parts = "Two projects"
- A set of exercises solved in lab sessions
- Grade=0.45*Part1+0.45*Part2+0.1*Exercises
  (this represents the 40% of the final IHLT grade)

- Exercise of a session: at the end of the session send the Jupyter notebook to `turmo@cs.upc.edu`

Topic of the project

How similar are two sentences between them? compare different approaches

Relevance of the topic:
IR, QA, summarization, automatic translation, plagiarism detection, ...

First part of the project

Deadline: 10/11/2017

A pair of texts is a paraphrase when both texts describe the same meaning with different words

- Implement at least three approaches to detect paraphase using sentence similarity metrics. Explore one lexical dimension per metric. Evaluate each approach using the provided evaluation framework (Raco). Compare and comment the results achieved by the approaches.

- Jupyter notebook: p1-[Student1]-[Student2].ipynb
- Comparisons file name: p1-[Student1]-[Student2].pdf
- send email to `turmo@cs.upc.edu` with title *'IHLT project1'*

Second part of the project

Deadline: 21/12/2017 (oral presentation)

- Implement other approaches to detect paraphrase using sentence similarity metrics: one based on the syntactic dimension alone and the rest based on combining it with those metrics you explored in the first part of the project. Evaluate each approach using the provided evaluation framework (Raco). Compare and comment the results achieved by these approaches among them and also among those explored in the first part of the project.

- Jupyter notebook: p2-[Student1]-[Student2].ipynb
- slides: p2-[Student1]-[Student2].pdf
- send email to `turmo@cs.upc.edu` with title *'IHLT project2'* before the oral presentation

Framework installation and execution

Windows:

- Install Anaconda (https://www.continuum.io/downloads). This installs python 3.6 and Jupyter.
- Install `nltk` and `scipy` python libraries (https://www.python.org)
- Execute Icon `anaconda-navigator` (select `jupyter-notebook`, select New/Python3)

Linux/OS:

> `sudo apt-get install python3`

> `pip3 install -U pip`

> `pip3 install jupyter`

> `sudo pip3 install -U nltk`

> `sudo pip3 install -U scipy`

> `jupyter notebook` (select New/Python3)

Stop server with Ctrl-C

Exercise

Validate the installation process

- Open a new python3 jupyter notebook
- Change the name of the session to S1-[Student1]-[Student2]
- Import without errors `nltk` library
- Save the session and exit jupyter server.
- Send S1-[Student1]-[Student2].ipynb by email to `turmo@cs.upc.edu` with title *session1*