Python for NLP

August 26-30, 2019
LORIA, Nancy

https://synalp.loria.fr/python4nlp
Organisation and Funding

- LIFT (C. Gardent), CNRS GDR
- OLKi Impact Project (C. Cerisara), LUE IDEX
Audience

- Basic Python Required
- Humanity Students (linguists etc.) and researchers
- CS students and researchers
- Industrials
Objective

Learn to

- Retrieve and store textual data from web, api (e.g., Gutenberg books, web pages, social network data)
- Apply linguistic processing (POS tagging, Parsing, NER, etc)
- Compute basic statistics and their visualisation (Nb of sentences, of tokens etc.)
- Apply basic Machine Learning Techniques (Classification, Clustering, Regression)
- Use word embeddings
<table>
<thead>
<tr>
<th>Day</th>
<th>9h-10h</th>
<th>10h30-12h</th>
<th>14h-15h</th>
<th>15h30-17h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday, 26 August</td>
<td><strong>Introduction</strong>&lt;br&gt;Christophe Cerisara&lt;br&gt;Claire Gardent</td>
<td><strong>Collecting data</strong>&lt;br&gt;Yannick Parmentier</td>
<td>Lab Session</td>
<td>Lab Session</td>
</tr>
<tr>
<td>Tuesday, 27 August</td>
<td>Invited talk&lt;br&gt;Albert Gatt</td>
<td><strong>Linguistic Processing</strong>&lt;br&gt;Yannick Parmentier</td>
<td>Lab Session</td>
<td>Lab Session</td>
</tr>
<tr>
<td>Wednesday, 28 August</td>
<td>Invited talk&lt;br&gt;Malvina Nissim</td>
<td><strong>Analysing textual data</strong>&lt;br&gt;Claire Gardent</td>
<td>Lab Session</td>
<td>Lab Session</td>
</tr>
<tr>
<td>Thursday, 29 August</td>
<td>Invited talk&lt;br&gt;Malvina Nissim</td>
<td><strong>Classification and Clustering</strong>&lt;br&gt;Claire Gardent</td>
<td>Lab Session</td>
<td>Lab Session</td>
</tr>
<tr>
<td>Friday, 30 August</td>
<td><strong>Using word embeddings</strong>&lt;br&gt;Christophe Cerisara</td>
<td>Lab Session</td>
<td>Lab Session</td>
<td>Lab Session</td>
</tr>
</tbody>
</table>
Collecting Text

- Interaction Web Server / Browser
- What’s in a web page
- Processing web pages
- What’s an API
- Extracting text from Wikipedia and Social Networks
Processing Text

- Sentence segmentation and tokenization
- Morphological analysis, stemming
- POS tagging
- Named Entity Recognition
- Parsing
Analysing Text

- Descriptive statistics
- Univariate Analysis (distribution, dispersion)
- Bivariate Analysis (Contingency, covariance)
- Vizualisation (scatter plot, box plots, histograms, bar plots)
Classification and Clustering

- What is Machine Learning?
- Extracting Features
- Train/Dev/Test Data
- Supervised and unsupervised learning (Classification, regression, clustering)
Word Embeddings

- What are word embeddings?
- Downloading and Using word embeddings
## Registration

<table>
<thead>
<tr>
<th>Category</th>
<th>Fee</th>
</tr>
</thead>
<tbody>
<tr>
<td>UL Students</td>
<td>0 Euros</td>
</tr>
<tr>
<td>Students (&lt;500km)</td>
<td>300 Euros</td>
</tr>
<tr>
<td>Students (&gt;500km)</td>
<td>100 euros</td>
</tr>
<tr>
<td>Academics</td>
<td>400 euros</td>
</tr>
<tr>
<td>Private Sector</td>
<td>800 euros</td>
</tr>
</tbody>
</table>