DIGITAL TOOLS FOR HUMANISTS SUMMER SCHOOL 2022

Program Week 1 – June 7 to June 10 2022

Tuesday June 7 - Morning

Introduction to Digital Humanities and a refresher on computers and networking
Vittore Casarosa (ISTI-CNR and University of Pisa)
One (simple) way to think of Digital Humanities is to think that it is just the use of "digital tools" in the study and research activities carried on by scholars in the Humanities. After a brief introduction to the main application fields where digital tools can be used (practically all of them), for the benefit of all those who were exposed to Computer Science a long time ago, or have been only marginally touched by it, we will briefly review the basics of computer architecture and the representation of information within a computer.
We will also see how the evolution of computer technology and of communication networks has led, in the early '90, to the explosive growth of the Internet and the Web, and how the actual Web is (slowly) evolving towards the Semantic Web.

Tuesday June 7 - Afternoon

Designing a project in Digital Public History
Enrica Salvatori (University of Pisa)
The main characteristics of a hypothetical DH project involving private and public realities of the territory will be illustrated, with the description of the main phases of its organization, implementation, maintenance and conservation. In the practical part we will try to create a work team on a concrete project and to design a possible work plan.

Wednesday June 8

Digitization of written sources
Federico Boschetti (ILC-CNR)
The theoretical part of the course illustrates methods and open source instruments for the preprocessing, acquisition by OCR/HTR, and postprocessing of text extracted from images of two-dimensional text-bearing objects (manuscripts, printed editions, maps, etc.). Techniques of preprocessing to optimize the images are described. The necessary steps to the text acquisition, such as training of OCR/HTR models, recognition, and accuracy evaluation are addressed. Finally, alignment techniques of multiple outputs and the application of linguistic models to improve the accuracy of the OCR/HTR results are discussed.
The practical part of the course provides the students with the skills necessary to manage the workflow for the acquisition of textual samples from different kinds of documents.

Thursday June 9

Machine Learning for automatic text analysis: tasks, methods, and tools

Alejandro Moreo (ISTI-CNR), Fabrizio Sebastiani (ISTI-CNR)

Many text analysis tasks are either tedious, or expensive, or time-consuming, or difficult to carry out; examples are (a) assigning subject codes (from a predefined taxonomy) to scientific papers, (b) determining, among a set of candidates, the most likely author of a text of unknown or disputed paternity, (c) marking a textual comment (on a product, on a political candidate, etc.) as conveying a positive or a negative opinion about its subject. Can these tasks be automated to some degree? Can we build tools that support the work of humans who carry out these tasks? These are the goals of machine learning as applied to automatic text analysis. In this course we will present machine learning methods for automating some of these tasks; a practical hands-on session will also introduce open-source tools that implement these methods.

Friday June 10

Natural Language Processing methods

Rachele Sprugnoli (University of Parma)

Natural Language Processing (NLP) is an interdisciplinary field whose goal is to create machines that understand natural languages. NLP applied to Humanities disciplines helps in dealing with large amount of data, extracting information and finding relationships and patterns between words.
The lesson will feature: (i) an introduction to the main areas of research within the field; (ii) hands-on activities on some NLP tasks, such as lemmatization, part-of-speech tagging, named entity recognition, topic modelling and keyword extraction.

DIGITAL TOOLS FOR HUMANISTS SUMMER SCHOOL 2022

Week 2 – June 13 to June 16 2022

Monday June 13 - Morning

Digital collections and digital libraries

CLARIN-IT group (ILC-CNR), Vittore Casarosa (ISTI-CNR and University of Pisa)

The session will start with an overview of the services provided by CLARIN (the European Research Infrastructure for Language Resources and Technology): discovery services, text processing and exploration services, deposit services. It will continue with a presentation and a hands on session on the ParlaMint corpora for multilingual parallel parliamentary data. Finally, it will provide a brief overview and some examples of use of the European Social Sciences & Humanities Open Marketplace, a portal to find and access resources (tools, services, training materials, workflows and datasets) for research in Social Sciences and Humanities.

The morning session will conclude with a brief overview of two of the most used systems for creating and managing digital collections, namely WordPress and Omeka.

Monday June 13 - Afternoon

Designing a project in Digital Public History

Enrica Salvatori (University of Pisa)

On the web we now frequently find digital libraries and archives, i.e. collections of digitized and annotated objects (items with metadata) that can be researched. This basic condition is not usually enough to enhance the digitized cultural heritage and make it really useful: in order to do this (Digital) Public History practices must be activated.

Tuesday June 14

Methods and tools for digital philology

Roberto Rosselli Del Turco (University of Torino)

Digital philology is a fairly recent discipline aiming at applying ICT methods and tools to the textual criticism area. Quite a number of new digital editions have been published during the last twenty years or so. Many of these editions, however, are achieved by programming and configuring complex frameworks, within the reach of medium-large research groups only. Encoding the edition texts in the TEI XML format allows the individual scholar to prepare a digital edition, but the on-line publication and navigation of such a site still remain a complicated and potentially expensive operation.

EVT (Edition Visualization Technology) is an open source tool whose purpose is to allow the scholar to publish TEI-based editions in an easy way, making available to the end user an user-friendly user interface and several research tools. This course will introduce the subject of digital philology, of text encoding using the TEI standard and a “hands one” final part when students will be able to experiment with EVT.

Wednesday June 15

Deep Learning tools for image classification and retrieval

Fabio Carrara (ISTI-CNR), Fabrizio Falchi (ISTI-CNR), Nicola Messina (ISTI-CNR)

The ongoing artificial intelligence renaissance is driven by deep learning – a sub field of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks. Deep learning methods are particularly good at learning representations of the data that make it easier to extract useful information when building classifiers or other predictors — automatically extracting information from humanities data and images in particular.

The theoretical part of the course will illustrate the basic aspects of deep learning and some applications to computer vision and multimedia retrieval.

The practical part of the course will provide the students with skills for learning and using tools for image representations suitable for classification and retrieval.

Thursday June 16

Meaning-making and storytelling in the age of databases, websites, and social media

Seamus Ross (University of Toronto)

It would be hard to ignore that the ways in which society represents, disseminates, and uses information has undergone dramatic changes in recent decades. The emergence of new types of documents in conjunction with the shifts in social nature of the construction of the documentary heritage poses challenges and opportunities to the process of humanities scholarship. This session examines the nature of the document and contemporary narratives as seen through the ways databases and websites represent content, make it accessible, and how they are used. In understanding these documents we will examine them with the lenses of the archive, their syntax, semantics and context and process (i.e., pragmatics), provenance, genre, diplomatics, and remix. The session aims to illuminate the impact of these two kinds of documents on humanities research and meaning-making in the coming decade.

During the applied afternoon session participants will experiment with interpreting the narrative inherent in a database and engage with a web archive to gain an appreciation of the role of these document classes as literature and historical sources.