POSTDATA
Poetry Standardization and Linked Open Data

DH @ MADRID Summer School 2019

Creando un proyecto de humanidades digitales usando el modelado de datos y el procesamiento de textos

1 al 3 de julio de 2019

POSTDATA – Poetry Standardization and Linked Open Data

http://postdata.linhd.uned.es

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement Nº [679528])

POSTDATA Poetry Standardization and Linked Open Data · UNED · Fundación Uned · erc European Research Council

Los textos son objeto de estudio en muchos proyectos de humanidades digitales como fuente de información. Dependiendo del dominio de conocimiento en el que se trabaja y del objetivo final de estudio, se aplican distintas tecnologías para su procesamiento.
En toda actividad de investigación centrada en textos y en cualquier proyecto de humanidades se identifican necesidades similares, aunque con distintos enfoques. Estas necesidades son: la estructuración de la información y su modo de almacenamiento, el procesamiento de los objetos de estudio, y los textos necesarios para extraer información y características significativas.
El impacto de la Web Semántica y los datos abiertos enlazados (LOD) implica una estructuración y gestión de la información semánticamente enriquecida, reutilizable e interoperable.
Por otro lado, las nuevas tecnologías de procesamiento de lenguaje natural y de inteligencia artificial permiten acometer distintas formas de procesamiento de los textos con obtención de resultados cada vez más abundantes y relevantes.
El objetivo de este curso es el desarrollo de un proyecto de humanidades digitales centrado en el tratamiento de textos, extracción y almacenamiento de la información asociada al mismo desde distintos puntos de vista.
Este curso está dirigido a todos aquellos interesados en métodos digitales de investigación aplicados a las humanidades, y con especial énfasis en personas con formación humanística.

1. Descripción del curso

El curso "Creando un proyecto de humanidades digitales usando el modelado de datos y el procesamiento de textos" pretende dotar a los estudiantes de los conocimientos y capacidades necesarios para acometer el desarrollo de un proyecto basado en textos poéticos.
Se estructurará en dos bloques de sesiones. En el primero se incluyen sesiones divulgativas de conocimiento en el ámbito de las humanidades digitales con ponentes de reconocido prestigio que enriquecerán los conocimientos del curso desde una perspectiva más divulgativa. En el segundo bloque se incluyen cuatro sesiones correspondientes a las distintas fases del proyecto y que guiarán al estudiante en el proceso.
Este segundo bloque es en el que se impartirán los conocimientos necesarios para llevar adelante un proyecto. La primera sesión presentará los conocimientos y herramientas necesarios para el modelado conceptual de la información y la generación de ontologías en el ámbito de la Web Semántica. Las siguientes sesiones presentarán tecnologías de procesamiento de lenguaje natural e inteligencia artificial para la realización de distintas tareas relacionadas con el procesamiento de los textos, su análisis y posterior extracción de información. El resultado final serán metadatos, datos estructurales y datos de información prosódica que se almacenarán en forma de datos abiertos enlazados y en otros formatos para su visualización.

Los enlaces de acceso al curso requerirán una contraseña que se enviará con antelación a los estudiantes inscritos. Los vídeos de las sesiones podrán seguirse en los siguientes enlaces:

- En directo: https://canal.uned.es/live/event/5c9a13f0a3eeb03d678c936a
- En diferido: https://canal.uned.es/series/magic/n9eb2vhsok0cscs44ogcs8o440g48w

2. Programa

2.1. Bloque I

- **Conferencia 1: "Humanidades digitales, tecnologías del lenguaje y análisis de poesía".**
  Elena González-Blanco. Investigadora Principal del proyecto ERC-POSTDATA
  En esta conferencia se presentarán las distintas herramientas y tecnologías que permiten realizar el análisis de un texto. En el contexto del análisis de la poesía, se presentará la génesis del proyecto POSTDATA como proyecto ejemplar de las Humanidades Digitales y de las diferentes tecnologías involucradas.

- **Conferencia 2: "¿Qué es la estilometría? Usos y aplicaciones.**
  José Manuel Fradejas. Catedrático de Filología Románica. Universidad de Valladolid.
  El mundo actual es un mundo basado en textos. Los vídeos que se suben a YouTube diariamente tienen un breve texto describiéndolo. Las reseñas de TripAdvisor, Amazon, etc. son textos que encierran información valiosa por lo que los especialistas en el análisis de datos han prestado siempre una gran atención a los textos y para ello han creado numerosas herramientas. Estas herramientas pueden ser muy útiles a los filólogos, y de hecho lo son. Sin embargo, producen cierto miedo entre los "de letras". Esta presentación tratará de mostrar cuán útiles pueden ser a un filólogo y presentará algunos procedimientos de gran interés para el trabajo filológico, y tratará de espantar ese miedo a las máquinas. Estas son nuestras aliadas y pueden ser muy útiles en nuestro quehacer textual.

- **Conferencia 3: "Modelos computacionales de creatividad literaria: poesía y narrativa".**
  Pablo Gervás. Director del grupo de Investigación NIL. UCM.
  Desde la invención de la Inteligencia Artificial se ha dedicado esfuerzo regularmente a intentar construir programas capaces de generar historias o poesías. Este esfuerzo empezó centrándose en la aplicación de tecnologías existentes (planificación, razonamiento basado en casos, simulaciones basadas en agentes, ontologías, programación evolutiva...) y ha evolucionado progresivamente para tener en cuenta descubrimientos más recientes en ciencia cognitiva, psicología y neurología. La conferencia revisará la trayectoria de estos esfuerzos de investigación desde el punto de vista de su relación con las humanidades digitales. Se presentarán

ejemplos de sistemas de generación de poesía y narrativa desarrollados en la Universidad Complutense de Madrid.

- **Workshop DARIAH/DESIR: Digital Tools, Shared Data and Research Dissemination"**
  Deborah Thorpe. Training and Education Officer DARIAH Coordination Office Dublin, Trinity College Dublin, Trinity Long Room Hub Arts and Humanities Research Institute, The University of Dublin.
  En este workshop se presentará DARIAH-EU como una "Infraestructura de Investigación" en forma de red para mejorar y apoyar la investigación y la enseñanza digital en las Artes y las Humanidades. Se mostrará una visión general de los Grupos de Trabajo de DARIAH como apoyo a la formación y educación en humanidades. A continuación, se abordará la cuestión de la identificación y gestión de los datos de investigación en humanidades de manera específica, centrándose en los datos abiertos, y las buenas prácticas para la investigación abierta y la gestión de datos. El objetivo es que los participantes adquieran conocimientos sobre cómo DARIAH puede apoyar, de forma práctica e intelectual, su investigación en humanidades con un componente digital, y sobre cómo adquirir más habilidades y experiencia en datos de investigación en humanidades y en investigaciones abiertas. Esta sesión se impartirá en inglés.

2.2. Bloque II

Sesión 1: Introducción al modelado conceptual: Creación de una ontología.

Esta sesión se centrará en el modelado de datos y la creación de un conjunto de datos como datos abiertos enlazados. En primer lugar, se creará un modelado conceptual del dominio de conocimiento, en este caso de las obras poéticas. El modelo conceptual diseñado servirá de base para la definición de una ontología orientada a la web semántica y a los datos abiertos enlazados (LOD). Los conceptos que se tratarán se enumeran a continuación:

1. Definición de modelo conceptual: construcción y visualización
2. Definición de ontologías
   a. Identificación de entidades
   b. Construcción con *Protégé*
3. Análisis de documentos *OWL* y *RDF*
4. Generación de conjuntos de datos
5. Enlazado de datos

Sesión 2: Introducción al procesamiento de textos con Python

En esta sesión introducirá el entorno de exploración de datos y programación con *notebooks* de *Jupyter/ Colab*. Además, se dará una breve introducción al lenguaje de programación *Python*, sus primitivas básicas y algunos módulos para manipulación de datos. Los temas que se van a tratar son:

1. La consola y el intérprete de Python
2. Los Notebooks de Jupyter
3. Sintaxis y uso de módulos en Python
4. Lectura y escritura de archivos (texto y tabulares)
5. Gráfico de frecuencia de palabras

Sesión 3: Usando librerías de Python para procesamiento de texto. Una introducción práctica a la exploración, análisis y manipulación del texto

Esta sesión se dedicará a la exploración, análisis y manipulación de textos con la herramienta SpaCy para el procesamiento del lenguaje natural. Los conceptos que se tratarán se enumeran a continuación:

1. Introducción a SpaCy
2. Tokenización y separación de frases
3. Anotación gramatical
4. Reconocimiento de entidades nombradas
5. Lematización y tipos de palabras
6. Uso de bibliotecas de *Python*. La biblioteca NLTK

Sesión 4: Aproximaciones modernas al análisis de texto (word embeddings)

En esta sesión se hará uso de aplicaciones para realizar la escansión y la detección de encabalgamientos con el objeto de enriquecer los metadatos de un poema. Para terminar la sesión se realizará una introducción a los *word embeddings* y sus aplicaciones.

Inicialmente se abordará el problema de la silabificación de versos para extraer una aproximación de la longitud. Además, se mostrará cómo hacer un detector de la figura retórica del encabalgamiento de tipo *tmesis*. Finalmente, tanto la información de la escansión como del encabalgamiento será traducida a triples de acuerdo con la definición de la ontología inicial. Los temas que se van a tratar son:

1. Expresiones regulares
2. Encabalgamiento y tmesis
3. Escansión y separación silábica
4. Producción de un CSV
5. Conversión de CSV a triples

3. Objetivos

Los objetivos del curso son:

- Comprender los modelos conceptuales.
- Construir una ontología orientada a la poesía y en el ámbito de la web semántica.
- Conocer los formatos y la generación de datos abiertos enlazados (LOD).
- Aprender el uso de herramientas y librerías para el procesamiento de textos en *Python*.
- Analizar y extraer información de textos poéticos usando *Python*.
- Enriquecer los datos enlazados de un poema con información extraída usando técnicas de procesamiento del lenguaje natural y *Python*.

4. Materiales didácticos

Los estudiantes contarán con material teórico y bibliografía recomendada para la asimilación de conceptos.

Toda la documentación y códigos utilizados en el curso se pueden descargar de https://github.com/linhd-postdata/summer-school2019

El software que se va a utilizar durante el curso es:

- Navegador Web (*Firefox*, *Chrome*, *Safari*, *Edge*) **(Todas las sesiones)**
- *Protégé* **(Sesion 1)**
- *Open Refine* **(Sesion 1)**
- *Microsoft VS Code* **(Opcional)**
- Software hoja de cálculo para la visualización y manipulación de ficheros en formato CSV (*LibreOffice*, *Excel* o *Numbers*)
- *Python **3.7** y las librerías necesarias (incluyendo Jupyter)* **(Sesiones 2, 3 y 4. Opcional, véase el punto 1 en el siguiente apartado.)**

Existen dos opciones para configurar el entorno necesario para seguir el curso:

1. Desde los enlaces de descarga facilitados anteriormente, instalar los paquetes y librerías en el ordenador que se vaya a usar para seguir el curso.
   Existe también la posibilidad de que para las sesiones 2, 3 y 4 **se utilice la plataforma Google Colab**, a la que se puede acceder desde el navegador, y solo **es necesaria una cuenta de Google** para iniciar sesión. Esta opción no requiere de instalación adicional y es la opción recomendada.

2. Instalar la máquina virtual (Ubuntu 19) que se detalla a continuación y que lleva instalado todo el software necesario. Esta opción requiere de un ordenador **relativamente potente,** así como al menos **4GB de RAM libres** y **20GB de espacio disponible en el disco duro.**

   Si se opta por esta opción, los pasos que se deben seguir son los siguientes:
   1. Instalar VirtualBox, seleccionando el enlace correspondiente al sistema operativo de la máquina en la que se va a realizar la instalación en el apartado "*VirtualBox 6.0.8 platform packages*" de la página https://www.virtualbox.org/wiki/Downloads

2. Descargar la imagen de la máquina virtual en los siguientes enlaces:

   - **Google Drive** (enlace público, no es necesario cuenta de GMail): https://drive.google.com/open?id=1R0zpBYp6vzbZTrqzuX_0mibfuLAGdMQ8

   - **OneDrive** (enlace público): https://unedo365-my.sharepoint.com/:u:/g/personal/alvaro_perez_linhd_uned_es/EQER1xMO7LhArpeJKKYE9-EBs8N54PLXgQrJ3LvieUTYRw?e=DMP5ld

   - **FTP:** ftp://62.204.199.125/SS2019.ova
     - **Usuario: humanidades**
     - **Contraseña: Hum100***

En la máquina virtual vienen instaladas las siguientes aplicaciones (con accesos en la barre lateral de aplicaciones y el escritorio), útiles para el curso:

- Terminal
- Firefox
- *Protégé*
- *Open Refine*
- *Microsoft VS Code*
- *LibreOffice*
- Python 3.7 y las librerías necesarias (incluyendo Jupyter)

5. Cronograma

FECHAS	SESIONES
1 de julio 15:00-17:00 h	Presentación del curso (indicaciones generales y técnicas) Conferencia: Humanidades digitales, tecnologías del lenguaje y análisis de poesía • Salvador Ros Muñoz. Profesor Titular de Informática. UNED. • Elena González Blanco. Investigadora Principal del Proyecto POSTDATA ERC. Laboratorio de Innovación en Humanidades Digitales - UNED.
1 de julio 17:00-19:00 h	Sesión 1: Introducción al modelado conceptual: Creación de una ontología • Mª Luisa Díez Platas. Investigadora del proyecto POSTDATA. LINHD-ETSI Informática, UNED.
2 de julio 9:00-11:00 h	Sesión 2: Introducción al procesamiento de textos con Python • Álvaro Pérez Pozo. Investigador del proyecto POSTDATA. LINHD-ETSI Informática, UNED.
2 de julio 11:00-13:00 h	Sesión 3: Usando librerías de Python para procesamiento de texto. Una introducción práctica a la exploración, análisis y manipulación del texto • Javier de la Rosa. Investigador del proyecto POSTDATA. LINHD-ETSI Informática, UNED. • Álvaro Pérez Pozo. Investigador del proyecto POSTDATA. LINHD-ETSI Informática, UNED.
2 de julio 15:00-17:00 h	Sesión 4: Aproximaciones modernas al análisis de texto (*word embeddings*) • Álvaro Pérez Pozo. Investigador del proyecto POSTDATA. LINHD-ETSI Informática, UNED. • Javier de la Rosa. Investigador del proyecto POSTDATA. LINHD-ETSI Informática, UNED. • Mª Luisa Díez Platas. Investigadora del proyecto POSTDATA. LINHD-ETSI Informática, UNED.
2 de julio 17:00-19:00 h	Conferencia: "¿Qué es la estilometría? Usos y aplicaciones. • José Manuel Fradejas. Catedrático de Filología Románica. Universidad de Valladolid.

3 de julio 9:00-11:00 h	**Conferencia: "Modelos computacionales de creatividad literaria: poesía y narrativa".** • Pablo Gervás. Profesor titular en el Departamento de Ingeniería del Software e Inteligencia Artificial, Facultad de Informática, UCM. Director del grupo de investigación NIL y del Instituto de Tecnología del Conocimiento.
3 de julio 11:00-13:00 h	**Workshop DARIAH/DESIR: "Digital Tools, Shared Data and Research Dissemination".** • Deborah Thorpe. Training and Education Officer DARIAH Coordination Office Dublin, Trinity College Dublin, Trinity Long Room Hub Arts and Humanities Research Institute, The University of Dublin.

6. Evaluación

Los estudiantes del curso deberán realizar el proyecto final descrito en el documento correspondiente que se facilitará en el campus virtual.
La entrega del proyecto se realizará a través del campus virtual, mediante la tarea habilitada a tal fin. La fecha límite de entrega es el 15 de julio de 2019.

7. Bibliografía recomendada

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. *International Journal on SemanticWeb and Information Systems (IJSWIS)*,*5*(3), 1–22. Recuperado desde http://www.igi-global.com/article/linked-data-story-far/37496

Caro Castro, Carmen. (2012). Vocabularios estructurados, Web Semántica y LinkedData: oportunidades y retos para los profesionales de la documentación. Disponible en *Arquivologia, Biblioteconomiae Ciência de Informação : Identidades, Contrastes e Perspectivas de Interlocução*.Niterói.

Duran Cals, J., Conesai Caralt, J., Clarisó Viladrosa, R. Ontologías y Web Semánticas. UOC.
http://www.cartagena99.com/recursos/alumnos/temarios/Ontologias%20y%20web%20semantica.pdf

Gruber, T. R. (1995). Towards principles for the design of ontologies used for knowledge sharing. *International Journal of Human- Computer Studies*, 1995, 43(4-5): 907-928. Disponible en: http://tomgruber.org/wrinting/onto-design.htm.

HAUSENBLAS, M. 5 estrellas de Datos Abiertos. (s. f.). Recuperado 1 de octubre de 2018, de http://5stardata.info/

Manning, C.D. y Schütze, H. Foundations of statistical natural language processing. *MIT Press* (1999).

Manning, C.D., Raghavan, P. y Schütze, H. Introduction to Information Retrieval. *Cambridge University Press* (2008).

Russell, S. y Norvig, P. Artificial Intelligence (A Modern Approach) 3rd edition. *Prentice–Hall Hispanoamericana* (2010).

SCHORLEMMER, M. (s. f.). Diez años construyendo una web semántica. Recuperado 1 de octubre de 2018, de http://www.fgcsic.es/lychnos/es_es/articulos/construyendo_una_web_semantica

Suárez-Figueroa, MC., Gómez-Pérez, A., Motta, E(2012). Ontology Engineering in a Networked World| *Mari Carmen Suárez-Figueroa / Springer.* (2012). Retrieved from https://www.springer.com/gp/book/9783642247934

Toomey, Dan. Learning Jupyter. Pack Publshing (2016).

Vanderplas, Jake. Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media (2016).

W3C. Guía breve de la Web Semántica. Disponible en: http://www.w3c.es/Divulgacion/GuiasBreves/WebSemantica

8. Equipo docente

Salvador Ros Muñoz. Profesor Titular de Informática. UNED y Director Técnico del Proyecto Europeo POSTDATA

Elena González-Blanco. Investigadora Principal del Proyecto Europeo POSTDATA y *General Manager* of *Europe* en *CoverWallet*

Mª Luisa Díez Platas. Investigadora del proyecto POSTDATA. LINHD-ETSI Informática, UNED.

Javier de la Rosa. Investigador del proyecto POSTDATA. LINHD-ETSI Informática, UNED.

Álvaro Pérez Pozo. Investigador del proyecto POSTDATA. LINHD-ETSI Informática, UNED.

José Manuel Fradejas. Catedrático de Filología Románica. Universidad de Valladolid.
Pablo Gervás. Profesor titular en el Departamento de Ingeniería del Software e Inteligencia Artificial, Facultad de Informática, UCM. Director del grupo de investigación NIL y del Instituto de Tecnología del Conocimiento.
Deborah Thorpe. Training and Education Officer DARIAH Coordination Office Dublin, Trinity College Dublin, Trinity Long Room Hub Arts and Humanities Research Institute, The University of Dublin.