Asignatura: Procesado y Manejo de Datos Masivos
Código: 33083
Centro: Escuela Politécnica Superior
Titulación: Máster en Bioinformática y Biología computacional
Nivel: Máster
Tipo: Obligatoria
Nº de créditos: 6

GUÍA DOCENTE DE PROCESADO Y MANEJO DE DATOS MASIVOS

La presente guía docente corresponde a la asignatura Procesado y Manejo de Datos Masivos (PDM), aprobada para el curso lectivo 2017-2018 en Junta de Centro y publicada en su versión definitiva en la página web de la Escuela Politécnica Superior. La guía docente de IA aprobada y publicada antes del periodo de matrícula tiene el carácter de contrato con el estudiante.

ASIGNATURA

PROCESADO Y MANEJO DE DATOS MASIVOS (PDM)

1.1. Código

33083 del Máster en Bioinformática y Biología computacional

1.2. Materia

Ciencias de la Computación e Inteligencia Artificial

1.3. Tipo

Obligatoria

1.4. Nivel

Máster

1.5. Curso

1º

1.6. Semestre

1º

1.7. Número de créditos

6 ECTS

1.8. Requisitos previos

Se necesitan conocimientos básicos de programación.

1.9. Requisitos mínimos de asistencia a las sesiones presenciales

Dado el carácter práctico de la asignatura, se considera imprescindible para su superación la asistencia a un mínimo del 70% de las sesiones de clase.

1.10. Datos del equipo docente

Nota: se debe añadir @uam.es a todas las direcciones de correo electrónico.

Profesores de teoría:

Dr. Luis Lago (coordinador)
Departamento de Ingeniería Informática
Escuela Politécnica Superior
Despacho: B-307
Teléfono: +34 914972211
Correo electrónico: luis.lago
Horario de tutorías: Petición de cita previa en clase o por correo electrónico.

Dr. Kostadin Koroutchev
Departamento de Ingeniería Informática
Escuela Politécnica Superior
Despacho: B355
Teléfono: +34 914973210
Correo electrónico: k.koroutchev
Horario de tutorías: Petición de cita previa en clase o por correo electrónico.

Dr. Estrella Pulido Cañabate
Departamento de Ingeniería Informática
Escuela Politécnica Superior
Despacho: B-413
Teléfono: +34 914972289
Correo electrónico: estrella.pulido
Horario de tutorías: Petición de cita previa en clase o por correo electrónico.

Dr. Gonzalo Martínez Muñoz
Departamento de Ingeniería Informática
Escuela Politécnica Superior
Despacho: B-422
Teléfono: +34 914977528
Correo electrónico: gonzalo.martinez
Horario de tutorías: Petición de cita previa en clase o por correo electrónico.

1.11. Objetivos del curso

En esta asignatura los estudiantes adquirirán conocimientos y destrezas relacionados con el procesado y manejo de datos masivos.

Tras finalizar esta asignatura el estudiante será capaz de procesar y manipular grandes volúmenes de datos programáticamente y mediante línea de comandos. Así mismo, el estudiante aprenderá a manejar y parsear los formatos de datos más comunes en bioinformática (como FASTA, UNIPROT, etc.) mediante patrones de línea de comandos y técnicas de programación que minimizan las necesidades de memoria y de tiempo de ejecución y programación paralela. Finalmente, el estudiante será capaz de realizar operaciones complejas en bases de datos relacionales y no-relacionales, así como acceder a bases de datos biomédicas online programáticamente mediante sus APIs de acceso.

Las competencias básicas y generales que el estudiante adquiere en esta asignatura son:

- CG1 - Capacidad para comprender y aplicar métodos y técnicas de investigación en el ámbito de la Bioinformática.
- CG2 - Capacidad para proyectar, calcular y diseñar productos bioinformáticos.
- CG3 - Capacidad para trabajar en equipos multidisciplinares, comunicándose eficientemente y desarrollando su actividad de acuerdo con las buenas prácticas científicas.
- CG4 - Capacidad para la investigación, desarrollo e innovación, en empresas y centros tecnológicos, en el ámbito de la Bioinformática.
- CG5 - Capacidad para la aplicación de los conocimientos adquiridos y resolución de problemas en entornos nuevos o poco conocidos en el ámbito de la Bioinformática.
- CG6 - Capacidad de búsqueda, análisis y gestión de información; incluyendo la capacidad de interpretación y evaluación con un razonamiento crítico y autocrítico.
- CG7 - Capacidad de estudiar y resolver problemas biológicos y biomédicos con el soporte de herramientas computacionales.
- CB6 - Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación
- CB7 - Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio
- CB8 - Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios
- CB9 - Que los estudiantes sepan comunicar sus conclusiones y los conocimientos y razones últimas que las sustentan a públicos especializados y no especializados de un modo claro y sin ambigüedades

- CB10 - Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

Las cualificaciones ubicadas en el nivel de competencias transversales que el estudiante adquirirá en esta asignatura son:

- CT1 - Capacidad para trabajar en equipo de forma colaborativa y con responsabilidad compartida en el diseño y comunicación de estrategias experimentales.
- CT2 - Capacidad de identificar fuentes de información científica solventes para fundamentar el estado de la cuestión de un problema bioinformático y poder abordar su resolución.

La competencia de tecnología específica que el estudiante adquiere en esta asignatura es:

- CE1 - Capacidad de aplicar los conocimientos de biología, matemáticas, física y estadística a la bioinformática.
- CE8 - Capacidad de utilizar técnicas computacionales para procesado, almacenamiento y manejo de datos masivos.
- CE10 - Capacidad de diseñar, implementar y evaluar una solución informática para resolver necesidades en el procesamiento de datos.

Al final del semestre (objetivos generales), y de cada unidad (objetivos por tema) el estudiante deberá ser capaz de:

OBJETIVOS GENERALES	
G1	Utilizar sistemas Linux para la gestión de grandes volúmenes de datos
G2	Utilizar técnicas de programación avanzada para el procesado de datos en bioinformática
G3	Realizar operaciones complejas en bases de datos relacionales y no-relacionales, así como acceder a bases de datos biomédicas online programáticamente mediante sus APIs de acceso.
G4	Conocer los métodos de acceso on-line, off-line y programático via REST de los principales repositorios de datos biomédicos.

OBJETIVOS ESPECIFICOS POR TEMA	
TEMA 1.- 1 Uso práctico de la línea de comandos para la extracción y correlación de información	
1.1.	Utilizar comandos avanzados de linux relacionados con la gestión de ficheros y procesos.
1.2.	Manejar con destreza, pipes, filtros y expresiones regulares.

TEMA 2.- Estrategias programáticas de parseo y extracción de datos	
2.1.	Identificar las familias de formatos más utilizados en bioinformática y las distintas partes de los formatos
2.2.	Utilizar biopython para parseo de datos.
2.3	Utilizar técnicas de programación paralela para procesado de datos
TEMA 3.- Bases de datos	
3.1.	Implementar consultas complejas en SQL
3.2	Identificar las limitaciones de las bases de datos relacionales que hacen necesarias las bases de datos NoSQL
3.3	Conocer los repositorios de datos biomédicos de NIH y EBML. Conocer los métodos programáticos de acceso a estos repositorios utilizando python y REST. Conocer los principios de las normativas de NIH/FDA sobre investigación y experimentos biomédicos y campos relacionados.
3.4	Implementar búsquedas estructuradas con flujo de datos simple utilizando el REST sobre BD de secuencias, secuencias anotadas y bibliografía biomédica en uno de los repositorios.

1.12. Contenidos del programa

Programa

1 Uso práctico de la línea de comandos para la extracción y correlación de información
- 1.1 Repaso de conceptos: pipes y filtros en la Shell
- 1.2 Expresiones regulares.
- 1.3 Comandos avanzados de Linux

2 Estrategias programáticas de parseo y extracción de datos
- 2.1 Familias de formatos de datos más comunes en bioinformática. Estrategias de parseo de los mismos.
- 2.2 Estructura de datos avanzadas relacionadas con el parseo
- 2.3 Estudio de librerías especializadas: BioPython
- 2.4 Programación paralela

3 Bases de datos
- 3.1 Bases de datos relacionales y SQL avanzado.
- 3.2 Introducción a las bases de datos no-relacionales.
- 3.3 Bases de datos públicas de uso común en bioinformática y acceso programático mediante APIs

1.13. Referencias de consulta

Bibliografía:

1. The Linux Cookbook: Tips and Techniques for Everyday Use. Michael Stutz. http://dsl.org/cookbook/cookbook_toc.html
2. The Linux Command Line. William Shotts. http://linuxcommand.org/
3. Numpy and Scipy Documentation: https://docs.scipy.org/doc/
4. pandas: powerful Python data analysis toolkit. http://pandas.pydata.org/pandas-docs/stable/
5. Biopython Documentation: http://biopython.org/wiki/Documentation
6. Spark Cookbook. Rishi Yadav. Packt Publishing. 2015.
7. Spark in action. Peter Zecevic, Marko Bonaci. Manning Publications. 2017.
8. Beginning Neo4j. Chris Kemper. Apress. 2016.
9. NoSQL for mere mortals. Dan Sullivan. Addison-Wesley. 2015.
10. Bioinformatics Data Skills: Reproducible and Robust Research with Open Source Tools, Vince Buffalo, ISBN: 978-1449367374
11. https://www.ebi.ac.uk/Tools/webservices/ (2017-07-04).

2. Métodos docentes

Los métodos docentes usados en la asignatura serán fundamentalmente prácticos y podrán incluir cualquiera de los siguientes:

- Clases Teóricas apoyadas con material multimedia
- Resolución de problemas o casos prácticos en el aula
- Metodologías e-learning
- Aprendizaje basado en problemas
- Trabajo autónomo de laboratorio
- Prácticas asistidas por ordenador
- Tutorías individuales o en grupos reducidos

3. Tiempo de trabajo del estudiante

		Nº de horas	Porcentaje
Presencial	Clases teóricas	28 h	50 h (33%)
	Clases prácticas	14 h	
	Tutorias	2 h	
	Realización de exámenes	6 h	
No presencial	Estudio semanal	20h	100h (67%)
	Realización de actividades prácticas	60h	
	Preparación del examen (convocatoria ordinaria)	10h	
	Preparación del examen (convocatoria extraordinaria)	10h	
Carga total de horas de trabajo: 25 horas x 6 ECTS		**150 h**	

4. Métodos de evaluación y porcentaje en la calificación final

Dos itinerarios de evaluación:

Evaluación continua: Necesario 70% de asistencia a clase y entrega puntual de las prácticas propuestas.

NOTA = 0.25*EXAMEN + 0.75*PRÁCTICAS

Evaluación no continua: Estudiantes que no cumplen los requisitos mínimos de asistencia o que entregan con retraso alguna de las prácticas.

NOTA = 0.7*EXAMEN + 0.3*PRÁCTICAS

5. Cronograma

El siguiente cronograma es aproximado. La distribución inicial de las clases puede sufrir variaciones en función de las necesidades docentes.

Semana	Tema	Contenido
1	1	1.1. Repaso de conceptos: pipes y filtros en la Shell 1.2. Expresiones regulares
2	1	1.3. Comandos avanzados de Linux
3	2	2.1. Familias de formatos de datos más comunes en bioinformática. Estrategias de parseo de los mismos. 2.2. Estructuras de datos avanzadas relacionadas con el parseo
4	2	2.3. Estudio de librerías especializadas: BioPython 2.4. Programación paralela
5	3	3.1. Bases de datos relacionales y SQL avanzado
6	3	3.2. Introducción a las bases de datos no-relacionales
7	3	3.3. Bases de datos públicas de uso común en bioinformática y acceso programático mediante APIs
8	3	3.3. Bases de datos públicas de uso común en bioinformática y acceso programático mediante APIs