FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Planificaciones

7506 - Organización de Datos

Docente responsable: ARGERICH LUIS RICARDO

OBJETIVOS

Introducir al alumno al mundo de la Ciencia de Datos (Data Science). Permitir que los alumnos utilicen herramientas modernas de análisis de datos para extraer información y realizar consultas. Brindar una detallada introducción a los temas mas importantes en la actualidad en el mundo del análisis de datos como Big Data, Machine Learning, Sistemas de Recomendaciones, Information Retrieval entre otros. Proporcionar herramientas para el análisis de datos masivos.

CONTENIDOS MÍNIMOS

-

PROGRAMA SINTÉTICO

Introducción a la Ciencia de Datos (Data Science)
Análisis Exploratorio de Datos.
Visualización de Datos.
Big Data, Procesamiento Distribuido
Complejidad y Compresión de Datos.
Hashing.
LSH.
Information Retrieval.
Procesamiento de Lenguaje Natural.
Visualización de Datos.
Reducción de Dimensiones.
PageRank y Derivados.
Algoritmos de Streaming.
Análisis de Redes Sociales.
Sistemas de Recomendaciones.
Análisis de Datos Topológico.
NoSQL.
Introducción a Machine Learning.
Clasificación.
Clustering.
Metadatos.
Blockchain.

PROGRAMA ANALÍTICO

Introducción a la Ciencia de Datos (Data Science)

Conceptos de estadística necesarios, conceptos de dimensionalidad, ejemplos de aplicacion de ciencia de datos en la práctica. Formatos de datos: CSV, JSON. Almacenamiento de matrices dispersas: CRS.

Análisis Exploratorio de Datos.

Herramientas para el análisis exploratorio de datos: R, Python. Pandas. DataFrames, concepto de índice. Consultas y modificaciones de Data Frames. Agrupación. Paradigma split-apply-combine. Tidy data. Formateo de los datos. Pivoteo de Dataframes, formatos tall y wide.

Visualización de Datos.

Conceptos de visualización de datos. Plots de líneas, barras, scatter plots, plot de burbujas, heatmaps, area plot, plots de correlación. Uso del color. Manejo de los ejes. Relevancia visual de los distintos atributos en un plot. Jerarquía visual. Conceptos de Tufte.

Big Data, Procesamiento Distribuido

Procesamiento distribuido y Map Reduce. Introducción a HDFS y Cloud-Storage.
Paradigma Map-Reduce usando Apache-Spark.
Primitivas básicas en PySpar: map, reduce, reducebykey, mapvalues, mappartitions, groupby, sort, filter,etc.
Acciones y transformaciones.
Procesamiento batch de datos masivos.
Explicación del funcionamiento interno de un sistema distribuido batch: fase de map, fase de reduce, fase de shuffle y sort.

Herramientas para procesamiento de datos masivos: Hive, Bigquery, Presto, Pig.

Hashing

Funciones de hashing. Funciones de hashing standard: FNV, Murmurhash, Cityhash, Pearson, Jenkins. Funciones de hashing criptográficas: SHA-2. Hashing Universal. Construcción de Carter-Wegman. Uso de múltiples funciones de hashing. Cuckoo Hashing. Estructuras de datos basadas en Cuckoo Hashing. Hashing Perfecto: Hash & Displace.

LSH

Locality sensitive hashing. Construcción de familias LSH para diferentes distancias. Hashmin para la distancia Jaccard. Amplificación de familias LSH, construcciones AND y OR. LSH para la distancia coseno, método de los hiperplanos. Cross-Polytope LSH. LSH para la distancia euclideana. Construcción de funciones LSH eficientes. LSH para semejanza máxima. De-duplicación de datos. LSH basada en los datos: Voronoi-LSH y K-Means LSH.

Complejidad y Compresión de Datos.

Complejidad de Kolmogorov, propiedades de la complejidad de Kolmogorov, incomputabilidad de la complejidad de Kolmogorov.
Aproximación de la complejidad mediante compresión de datos. Distancia normalizada de compresión.
Concepto de Modelo de mínima descripción.
Códigos de Huffman, Huffman estático y dinámico. Modelos de Orden Superior.
Compresión aritmética. PPMC, PPMD,PPMZ.
Block Sorting, MTF, Modelo Estructurado.
Familia de compresores LZ: LZ77, LZW, LZHuff. LZP.
DMC

Information Retrieval.

Indices invertidos, concepto, construcción de índices invertidos.
Almacenamiento de punteros: Código unario, gamma, delta.
Uso de índices invertidos para resolución de consultas: puntuales, booleanas, comodines, frases y proximidad.
Indices de n-gramas y léxico rotado.
TF-IDF.
Consultas ranqueadas, método del coseno. BM25.
Evaluación de consultas ranqueadas: Precision y Recall.
Indexación Semántica Latente.
Modelos probabilísticos.
Learning to Rank.

Procesamiento de Lenguaje Natural.

Modelos de n-gramas.
Smoothing: Corrección de Laplace, Good Turing Smoothing, Kneser-Ney.
Recuperación de bigramas y trigramas frecuentes.
Procesamiento de textos: Tokenización, stop-words, stemming.
Topic Models.
Probabilistic LSA y LDA.

Reducción de Dimensiones.

Métodos para reduccion de dimensiones lineales: SVD, PCA, MDS.
Métodos para reducción de dimensiones no-lineales: ISOMAP, Laplacian Eigenmaps, TSNE.
La hipótesis del manifold.
La maldición de la dimensionalidad.
Dimensionalidad intrínseca de un set de datos.

PageRank y Derivados.

Pagerank, formulación matemática. Modelos de Markov. Teletransportación. Solución de dead-ends. Spider-traps. Trust-Ran. Topic-Rank. Sim-Rank. Visual-Rank. Text-Rank. HITS.

Calculo de pagerank con datos masivos.

Análisis de Redes Sociales.

Características de las Redes Sociales como grafos. Distribución del grado y power-laws. Modelo de Erdos-Renyi. Modelo de Barabasi-Albert y preferential attachment. El fenómeno del mundo pequeño. Clausura triangular. Triángulos en Redes Sociales. Descomposición espectral de la matriz de adyacencias de una RS, eigenplots. Modelo de Watts-Strogratz. Detección de comunidades en Redes Sociales: Louvain, Infomap, Clustering Espectral, otros algoritmos. Centralidad: grado, betweenness, pagerank. Coeficiente de clustering en redes sociales. Algoritmos para procesar redes sociales masivas basados en Map Reduce. BFS distribuido.

Algoritmos de Streaming

Concepto de streams de datos y aplicaciones. Algoritmos de muestreo: Reservoir sampling. Momentos de un Stream. Flajolet-Martin. AMS. Calculo sobre ventanas: Decaying Windows, DGIM. Filtros de Bloom. Count-Min y aplicaciones. Misra-Gries.

Sistemas de Recomendaciones.

Sistemas de recomendaciones basados en contenido. TF-IDF. Collaborative Filtering: User-User e Item-Item. Métodos basados en factores latentes: SVD++, NMF: ALS, Factorization Machines.
Evaluación de Sistemas de Recomendaciones.

Análisis de Datos Topológico.

Algoritmos para análisis topológicos de datos. Homología Persistente. Complejo de Vietoris-Rips. Complejo de Cech. Teselación de Voronoi.
Algoritmo Mapper. Lentes Topológicos. Compresión topológica de un set de datos, representación, visualización e interpretación de los resultados.

Introducción a Machine Learning

Parámetros e hiper-parámetros. Cross-Validation. Grid-search y random-search. Underfitting y Overfitting. Bias y Variance. Teorema NFL (No Free Lunch). PAC Learning. KNN.

Clasificación y regresión.

Regresion lineal y logistic regression. Bayes y Naive Bayes. Perceptron y SVMs. SVM lineales, Kernel Trick. SVM online, algoritmo Pegaso. Arboles de decisión y Random Forests. XGBoost. Construcción de ensambles: Boosting, Bagging, Stacking.

Clustering.

Introducción al aprendizaje no-supervisado. K-Means, K-Means++, K-Means online. Variantes modernas de K-Means. Clustering Espectral. Clustering Jerárquico. DBSCAN. HDBSCAN. Evaluación de clustering.

NoSQL.

Formatos de Bases NoSQL. Bases para clave-valor. Bases para documentos, bases para datos estructurados, bases de tipo columna, bases para grafos. Ejemplos basados en una selección entre: Cassandra, MongoDb, Redis, Neo4J.

BIBLIOGRAFÍA

[Luis Argerich] Apunte del Curso.

[Jure Leskovec, Jeff Ullman, Anand Rajaraman] Mining Massive Datasets.
http://www.mmds.org/

[Christopher Manning] Introduction to Information Retrieval.
http://nlp.stanford.edu/IR-book/

[Matt Mahoney] Data Compression Explained.
http://www.mattmahoney.net/dc/dce.html

Todos los libros usados en la materia están disponibles online de forma gratuita y legal.

RÉGIMEN DE CURSADA

Metodología de enseñanza

Clases Teórico-Prácticas.
Exposición de los conceptos, explicación mediante ejemplos, comentario de tendencias actuales de investigación sobre cada uno de los temas. Ejercicios para practicar los algoritmos que se ven.

Modalidad de Evaluación Parcial

La evaluación parcial consta de 7 ejercicios que cubren los temas vistos en la materia hasta el examen. Los ejercicios son teóricos o prácticos y evaluan el conocimiento de los alumnos sobre los algoritmos vistos en clase y la capacidad de resolver situaciones nuevas usando estos algoritmos.

CALENDARIO DE CLASES

Semana	Temas de teoría	Resolución de problemas	Laboratorio	Otro tipo	Fecha entrega Informe TP	Bibliografía básica
<1> 09/03 al 14/03	Introducción a Data Science	Small Data: Análisis Exploratorio de Datos. Visualización. Pandas.				
<2> 16/03 al 21/03	Visualización de Datos.	Big Data: Apache Spark 1.				
<3> 23/03 al 28/03	Big Data: Apache Spark 2.	Big Data: PIG Latin.				
<4> 30/03 al 04/04	ITI 1.	ITI2.				
<5> 06/04 al 11/04	ITI3.	Hashing.				
<6> 13/04 al 18/04	LSH.	FERIADO.				
<7> 20/04 al 25/04	Dimensionalidad y Reducción de Dimensiones.	Information Retrieval 1.			Entrega de Informe de Diseño del TP.	
<8> 27/04 al 02/05	Information Retrieval 2.	Procesamiento de Lenguaje Natural.				
<9> 04/05 al 09/05	FERIADO.	Ejercicios.				
<10> 11/05 al 16/05	Examen PARCIAL.	Algoritmos de Streaming.				
<11> 18/05 al 23/05	Machine Learning 1.	Machine Learning 2.				
<12> 25/05 al 30/05	Examen RECUPERATORIO.	FERIADO.				
<13> 01/06 al 06/06	Clustering.	PageRank.				
<14> 08/06 al 13/06	Redes Sociales.	Sistemas de Recomendaciones.				
<15> 15/06 al 20/06	Análisis de Datos Topológico.	NoSQL.				
<16> 22/06 al 27/06	FERIADO.	Correción de TPs.			Entrega Final del TP	

CALENDARIO DE EVALUACIONES

Evaluación Parcial

Oportunidad	Semana	Fecha	Hora	Aula
1º	10	08/05	19:00	400
2º	12	22/05	19:00	400
3º	16	26/06	19:00	400
4º				

Observaciones sobre el Temario de la Evaluación Parcial
Se toma una evaluación parcial obligatoria que tiene dos recuperatorios de acuerdo al reglamento de cursada vigente.