CD - 1055-17

FACULTAD DE CIENCIAS EXACTAS Y NATURALES FOLIO 4

Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales

Planilla a completar para presentación de Cursos de Posgrado

1.- DEPARTAMENTO de COMPUTACIÓN

2.- NOMBRE DEL CURSO: Seminarios de Ciencia de los Datos

3.- DOCENTES:

RESPONSABLE/S: Diego Fernandez Slezak, Alejo Salles
COLABORADORES: .....
AUXILIARES: .....

4.- CARRERA de DOCTORADO

5.- AÑO: 2016 CUATRIMESTRE/S: .....

6.- PUNTAJE PROPUESTO PARA CARRERA DE DOCTORADO: .....4.....

7.- DURACIÓN (anual, cuatrimestral, bimestral u otra): cuatrimestral

8.- CARGA HORARIA SEMANAL:

Teóricas: .....
Problemas: .....
Laboratorio: .....
Seminarios: .....3 presentaciones por doctorando.....
Teórico – Práctico: 5hs.
Salida a Campo: .....

9.- CARGA HORARIA TOTAL: ..... 80 .....

10.- FORMA DE EVALUACIÓN: ...Trabajo práctico, parcial, presentación de artículos y final.....

11.- PROGRAMA ANALÍTICO:

Ciencias de los Datos es una disciplina reciente que surje de la disponibilidad masiva de conjuntos de datos de todo tipo. Esta es una materia estrictamente interdisciplinaria que

reúne temas de ciencias de la computación, matemática, física, y ciencias naturales. Se trata de un abordaje desde una perspectiva algorítmica y de la teoría de la información aplicados al análisis de datos de grandes corpus. En particular, se propone el uso de técnicas modernas de procesamiento de información (como por ejemplo, Machine Learning), y las tecnologías de procesamiento masivo (como por ejrmplo MapReduce o Elastic Search).

Los objetivos de la materia son: 1) Proveer a los alumnos con un conjunto de herramientas matemáticas (de análisis estadístico de datos) que permitan el análisis estadístico clásico a conjuntos de datos. 2) Proveer a los alumnos herramientas computacionales para el modelado y adquisición de datos en grandes corpus. 3) Presentar las tecnologías recientes para procesamiento de cantidades masivas de dato. 4) Introducir al alumno en el análisis de datos utilizando Machine Learning.

Se llevaran a cabo experimentos (en el laboratorio y en grandes repositorios de datos tomados de la web) con particular foco en neurociencia, para inferir propiedades del computo humano a partir de datos observacionales. Se utilizara este problema especifico para abordar el problema de análisis de datos en muchas dimensiones, por ejemplo en el analisis de regularidades en grandes corpus de texto.

Programa:

1) Introducción a Python, NumPy, Pandas (nivelación para estudiantes de carreras distintas a Cs. De la Computación)
2) Introducción al Análisis Matemático Estadístico clásico
3) Accediendo a Grandes corpus de datos (Bases de datos no relacionales, MapReduce, Hive, Spark)
4) Algoritmos aproximados para procesamiento en grandes corpus de datos
5) Análisis de datos utilizando Machine Learning

12.- BIBLIOGRAFÍA:

- Ariely, D., & Jones, S. (2008). Predictably irrational: The hidden forces that shape our decisions. New York, NY: Harper.
- Bak, P. (1996). How Nature Works: The Science of Self-Organized Criticality. New York: Copernicus.
- Balci, F., Simen, P., Niyogi, R., Saxe, A., Hughes, J., Holmes, P., & Cohen, J. D. (2011). Acquisition of decision making criteria: Reward rate ultimately beats accuracy. Attention, Perception, & Psychophysics, 1–18.
- Broder, Andrei, et al. "Graph structure in the web." *Computer networks* 33.1 (2000): 309-320.
- Damasio, A. (2000). The feeling of what happens: Body and emotion in the making of consciousness. London, England: Vintage.
- Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." *Communications of the ACM* 51.1 (2008): 107-113.
- Dijksterhuis, A., Bos, M., Nordgren, L., and Van Baaren, R. (2006). On making the right choice: the deliberation-without-attention effect. Science 311, 1005.
- Gaber, Mohamed Medhat. *Scientific data mining and knowledge discovery*. Springer, 2009.

- Gobet, F., and Simon, H. (1996a). Templates in chess memory: a mechanism for recalling several boards. Cogn. Psychol. 31, 1–40.
- Gold, J., and Shadlen, M. (2002). Banburismus and the brain decoding the relationship between sensory stimuli, decisions, and reward. Neuron 36, 299–308.
- Hick, W. (1952). On the rate of gain of information. Q. J. Exp. Psychol. 4, 11–26.
- Isard, Michael, et al. "Dryad: distributed data-parallel programs from sequential building blocks." *ACM SIGOPS Operating Systems Review*. Vol. 41. No. 3. ACM, 2007.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). Judgment under uncertainty: Heuristics and biases. Cambridge, England: Cambridge University Press.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 263–291.
- Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2014.
- Littman, M. (1996). Algorithms for Sequential Decision Making. Ph.D. thesis, Brown University, Providence, RI.
- Luce, R. (1986). Response Times: Their Role in Inferring Elementary Mental Organization. New York: Oxford University Press.
- Manning, Raghavan & Schütze, An Introduction to Information Retrieval, Cambridge University Press, 2009.
- Page, Lawrence, et al. "The PageRank citation ranking: bringing order to the web." (1999).
- Shannon, C. (1950). Programming a computer for playing chess. Philos. Mag. 41, 256–275.
- Tan, Pang-Ning. *Introduction to data mining*. Pearson Education India, 2006.
- Von Ahn, L. (2006). Games with a purpose. Computer 39, 92–94.
- Von Ahn, L., and Dabbish, L. (2004). "Labeling images with a computer game," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (New York, NY: ACM), 319–326.
- Von Ahn, L., Liu, R., and Blum, M. (2006). "Peekaboom: a game for locating objects in images," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (New York, NY: ACM), 55–64.
- Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., and Blum, M. (2008). reCAPTCHA: human-based character recognition via web security measures. Science 321, 1465.
- Wagenmakers, E., and Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. Psychol. Rev. 114, 830.
- Zylberberg, A., Fernandez Slezak, D., Roelfsema, P. R., Dehaene, S., and Sigman, M.(2010). The brain's router: a cortical network model of serial processing in the primate brain. PLoS Comput. Biol. 6, e1000765.