PRUEBA TÉCNICA PARA ROL "CIENTÍFICO DE DATOS"

A continuación se especifican 2 casos de uso, de los cuales usted **solo debe seleccionar 1** para desarrollar. En el desarrollo del caso usted debe demostrar sus habilidades de analítica en la resolución de problemas e identificación de patrones. Se espera que para el desarrollo del caso de uso haga uso oportuno de las herramientas de analítica necesarias para la resolución del mismo siguiendo alguna metodología de desarrollo de problemas analíticos.

Al final del caso se espera recibir el HTML del código empleado para el desarrollo del mismo, así como el código y sus componentes en un estado reproducible clonando el repositorio. Para la entrevista, si lo desea puede complementar el código con ayudas visuales tipo presentación aunque esto no es estrictamente necesario.

CASO DE USO 1:

Este conjunto de datos comprende varios resúmenes de papers, uno por archivo, proporcionados por la NSF (Fundación Nacional de Ciencia). Al final se muestra un abstract de ejemplo.

Actividad

- Su tarea es desarrollar un modelo no supervisado que clasifique los resúmenes en un tema. De hecho, su objetivo es agrupar resúmenes en función de su similitud semántica.

Puede obtener una muestra de resúmenes aquí. Sea creativo y establezca claramente su enfoque. Aunque no esperamos resultados precisos sino una buena canalización de su trabajo. La idea es crear un cuaderno, como Jupyter en Python o en el lenguaje analítico de su preferencia y ponerlo a nuestra disposición, preferiblemente, en github.

Sugerencia para el éxito en su búsqueda: Desarrolle y especifique el proceso de ciencia de datos que va a desarrollar y resaltar aspectos importantes que podría considerar para ser discutidos.

Nota: no se necesitan todos los campos en cada archivo de resumen.

Buena suerte.

Ejemplo de abstract:

==========

Title : CAREER: Markov Chain Monte Carlo Methods
Type: Award
NSF Org : CCR
Latest Amendment Date : May 5, 2003
File : a0237834
Award Number: 0237834
Award Instr.: Continuing grant

Prgm Manager: Ding-Zhu Du
CCR DIV OF COMPUTER-COMMUNICATIONS RESEARCH
CSE DIRECT FOR COMPUTER & INFO SCIE & ENGINR
Start Date : August 1, 2003
Expires : May 31, 2008 (Estimated)
Expected Total Amt. : $400000 (Estimated)
Investigator: Eric Vigoda vigoda@cs.uchicago.edu (Principal Investigator current)
Sponsor : University of Chicago
5801 South Ellis Avenue
Chicago, IL 606371404 773/702-8602
NSF Program : 2860 THEORY OF COMPUTING
Fld Applictn: Program Ref : 1045,1187,9216,HPCC,
Abstract : Markov chain Monte Carlo (MCMC) methods are an important algorithmic device in a variety of fields. This project studies techniques for rigorous analysis of the convergence properties of Markov chains. The emphasis is on refining probabilistic, analytic and combinatorial tools (such as coupling, log-Sobolev, and canonical paths) to improve existing algorithms and develop efficient algorithms for important open problems. Problems arising in computer science, discrete mathematics, and physics are of particular interest, e.g., generating random colorings and independent sets of bounded-degree graphs, approximating the permanent, estimating the volume of a convex body, and sampling contingency tables. The project also studies inherent connections between phase transitions in statistical physics models and convergence properties of associated Markov chains. The investigator is developing a new graduate course on MCMC methods.
==========

CASO DE USO 2:

Diabetes

En este conjunto de datos tiene 3 salidas diferentes:

1. No readmisión;
2. Un reingreso en menos de 30 días (esta situación no es buena, porque tal vez su tratamiento no fue apropiado);
3. Un reingreso en más de 30 días (este no es bueno como el último, sin embargo, la razón podría ser el estado del paciente).

Actividad

- Su tarea es **clasificar** un resultado paciente-hospital o **clusterizar** con el objetivo de encontrar patrones que den una visión distinta.

- La idea es crear un cuaderno, como Jupyter en Python o en el lenguaje analítico de su preferencia y ponerlo a nuestra disposición, preferiblemente, en github.

Sugerencia para el éxito en su búsqueda: Desarrolle y especifique el proceso de ciencia de datos que va a desarrollar y resaltar aspectos importantes que podría considerar para ser discutidos.

Buena suerte

Habilidades a evaluar en la prueba y en la entrevista:

- Conocimiento teórico de desarrollo de modelos analíticos
- Metodología de desarrollo analítico
- Programación en lenguajes de código abierto
- Conocimientos de cloud computing
- Entendimiento de operacionalización de productos analíticos
- Entendimiento de métodos de trabajo colaborativo