Pruebas Científico de Datos Sanitas

Elija una de las pruebas mostradas a continuación y desarróllela mostrando su proceso de pensamiento y el cómo creó las soluciones necesarias en cada aspecto de la prueba.

Procesamiento de Lenguaje Natural (Natural Language Processing)

1. **Resúmenes de los premios de investigación de la NSF**

En el siguiente conjunto de datos comprimido encontrará resúmenes de artículos de investigación que fueron extraídos por la NSF (Fundación Nacional de Ciencia).

Tú tarea es desarrollar un modelo no supervisado que clasifique los resúmenes dentro de un tema específico, es decir, agrupar los resúmenes basado en su similitud semántica (el conjunto de datos comprimido se encuentra adjunto en el correo de esta prueba).

Diseñe un notebook en *Jupyter* (si usas Python) o un reporte *Rmarkdown* (en caso de que utilices R) y hazlo disponible en la red, ejemplo *github*.

Pista o tip: Muestre claramente los procesos de ciencia de datos que quiere ejecutar sobre los datos y resalte los aspectos que considere más relevantes y que valga la pena discutir. Tenga en cuenta que no todo el resumen del artículo es útil para llevar a cabo el desarrollo, sea selectivo.

Muestra de un resumen de artículo:

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.

==========

2. Análisis de críticas de contenidos

Para las grandes empresas de contenidos es relevante tener en cuenta la percepción de los usuarios sobre sus contenidos, es por ello que, consideran todas las reseñas que dejan los críticos/aficionados sobre sus contenidos, y en base a ello se generan acciones y decisiones sobre cómo orientar los nuevos contenidos.

Para el conjunto de datos adjunto, se ha identificado todas las reseñas de los críticos/aficionados de ciertos contenidos y una etiqueta de polaridad de su percepción sobre el contenido, la que se marca como positiva o negativa.

El objetivo es identificar cuándo una reseña puede ser positiva o negativa.

Diseñe un notebook en *Jupyter* (si usas Python) o un reporte *Rmarkdown* (en caso de que utilices R) y hazlo disponible en la red, ejemplo *github*.

Pista o tip: Muestre claramente los procesos de ciencia de datos que quiere ejecutar sobre los datos y resalte los aspectos que considere más relevantes y que valga la pena discutir. Tenga en cuenta la importancia de analizar los grupos de palabras y la raíz de las palabras.

Recuerde calcular métricas de los modelos que desarrolle.