**Temario Examen Curso Big Data**

**Fecha: sábado 3 de mayo**

**Unidad I: Preprocesamiento y Análisis Exploratorio de Datos**

* Tipos de datos (estructurados, no estructurados)
* Tratamiento de valores faltantes (NaN, imputación, eliminación)
* Detección y tratamiento de outliers
* Normalización y estandarización
* Exploración de datos: estadísticas descriptivas, correlación, visualización
* Importancia del preprocesamiento antes de aplicar algoritmos

**Unidad II: Procesamiento de Lenguaje Natural (NLP)**

* Concepto y aplicaciones del NLP
* Tipos de datos en NLP: reviews, noticias, opiniones, etc.
* Pipeline de procesamiento de texto:
  + Limpieza (minúsculas, símbolos, espacios)
  + Tokenización
  + Palabras de parada (stopwords)
  + Derivación (stemming) vs lematización
  + Etiquetado gramatical (POS tagging)
  + Reconocimiento de entidades nombradas (NER)
  + Extracción de relaciones
* Herramientas de NLP (NLTK, spaCy, etc.)

**Unidad III: PySpark y Procesamiento Distribuido**

* ¿Qué es Apache Spark ?
* ¿Qué es PySpark y qué lo hace útil para Big Data?
* Machine Learning con PySpark (overview): VectorAssembler, Pipelines