Bienvenidos!
Semana 4

Datos No Estructurados y Semiestructurados

Especialización en Economía, opción Ciencia de Datos

FCS - Udelar

Guillermo Lezama

La clase de hoy

- Comentarios de organización del curso
- Transformar PDFs
- Preprocesamiento (Stemming y Lematizacion)
- Del texto a matrices
- Similitudes entre textos

- Cuestionarios
- Tareas
- Emails
- EVA

Proyecto Final

- Similar a actividad grupal.
- La diferencia: si van a trabajar con los datos.
- Objetivo: Transformar un conjunto de datos desestructurados en algo estructurado + contar algo de esos datos + articularlo con una historia a contar.
- 9 de Junio: Contar qué piensan hacer
- Entrega: 31 de Julio
- 3 opciones: Texto, imagen, Sonido.
- Grupos de max 4.
- Agregar si hay algún tipo de texto, imágenes o sonidos que quieran trabajar

El lunes

- 1 o 2 slides
- Pregunta que quisieran responder => Necesitan un tema!
- ¿Qué tipo de datos estructurados precisarían para responder eso?
- ¿Qué fuente de datos no estructurados van a usar
- Es un cuarto de la nota del trabajo final.
- ¿Cómo evalúo el trabajo final? Complejidad de la tarea, entender por qué importa lo que hacen…

Entregable

- 1 excel o csv con filas y columnas
- 1 pagina contando, en sus propias palabras, qué hicieron y por qué lo hicieron (el por qué, es lo del lunes)
- 1 notebook (puede ser en R tambien)

- Obtener
- **Procesar**
- Analizar

- ¿Preguntas?

De dónde sacar los textos?

- Muchas veces están en PDF
- Imágenes
- Columnas
- pdfplumber
  - Varias páginas

Stemming y Lematización

- Palabras parecidas complican análisis: gato, gatos, gatitos.
- Reducir dimensionalidad es bueno!
- Reducir palabras a una raíz común
  - 'jugando', 'jugó', 'jugar’ son palabras que refieren a cosas parecidas
  - Stemming: ‘jug’
  - Lematización: ‘jugar’

- Lematización es precisa
  - Necesita de mucho conocimiento del idioma
  - Más costoso

- Stemming es rápido y simple, pero a veces carece de sentido

Matriz Documento-término

- En general comparamos muchos documentos.
- Filas: Documentos
- Columnas: Palabras
- En cada celda, cuantas veces aparece esa palabra.
- Pablito clavó un clavito, qué clavito clavó pablito?

pablito	clavó	un	clavito	qué
2	2	1	2	1

TF-IDF

- **TF-IDF** es una técnica utilizada para evaluar la importancia de una palabra en un documento dentro de un conjunto de documentos.
- **TF (Term Frequency):** Mide la frecuencia de una palabra en un documento específico.
- **IDF (Inverse Document Frequency):** Mide cuán común o raro es un término en todo el conjunto de documentos.
- Penaliza las palabras comunes y da más peso a las palabras que son raras en el conjunto de documentos, ayudando a identificar términos clave en un documento.
- país, gobierno, políticas…

Cosine Similarity

- **Cosine Similarity** es una métrica que mide la similitud entre dos vectores no nulos mediante el coseno del ángulo entre ellos.

$$AxB/||A||\ ||B||$$

- Entre -1 (opuestos) y 1 (idénticos).
- 0 (no hay relación)

Para próxima clase

- Esta semana: un cuestionario cada clase
- Tarea del notebook
- Revisar el mail