Proyecto Cadenas de Markov

Objetivo: Aplicar los conceptos básicos de las Cadenas de Markov.

El proyecto es en parejas.

Apliciación

Consideraremos que un texto se puede modelar palabra a palabra. De forma que la aparición de una palabra en una posición del texto solo depende de la palabra inmediatamente anterior. De esta forma sería una Cadena de Markov de primer orden.

Se debe generar un vocabulario a partir de una serie de textos, el cual nos daría el *espacio de estados* de la Cadena. También las *probabilidades de transición* se obtendrían a partir de estos textos con las frecuencias observadas de las parejas de palabras que aparezcan en ellos.

A partir de esta Cadena podemos hacer generación de texto partiendo de un *estado inicial* y simulando la Cadena. Analizaremos la Cadena con los conceptos vistos.

Elección de Corpus

Este se deja a discreción de las parejas. Debe incluir entre 20 y 100 textos de al menos 200 palabras cada uno. Por ejemplo: canciones, cuentos, artículos de wikipedia, descripciones de juegos, etc.

Implementación

Usando Python (u otro lenguaje de su preferencia) deben programar (por lo menos) lo siguiente:

- Procesamiento de texto
    - Lectura de los archivos de entrenamiento
    - Separación y estructuración por palabras (tokenización)
    - Estandarización (remover mayúsculas, tildes, etc.)
    - Manejo de puntuación
    - Otros procesos necesarios
- Construcción del modelo
    - Creación de la distribución de probabilidad inicial
    - Crear la representación de los estados (mapeo de palabras a índices)
    - Cálculo de frecuencias de transiciones entre parejas de palabras consecutivas
    - Construcción de la matriz de transiciones y obtención de probabilidades
- Análisis de la Cadena de Markov
    - Programar funciones que permitan:

        - Algunas clases de comunicación
        - Periodicidad de algunos estados relevantes estados
    - Una función que calcule la probabilidad de transición a *n* pasos dado un valor de *n*. ¿Qué significan estos valores en este contexto?
        - Usar las ecuaciones de Chapman-Kolmogorov como parte de este apartado
- Implementar una simulación de una secuencia de un largo dado, iniciando por una palabra dada. Tener la posibilidad de sugerir una lista aleatoria de 10 palabras de vocabulario

Informe escrito

Reporte final lo más corto y conciso posible que incluya lo siguiente:

- Corpus de textos usado y lógica de la elección
- Histograma u otra visualización de la distribución de probabilidad inicial
- Preprocesamiento realizado
- ¿Cómo construyeron el espacio de estados?
- ¿Cómo construyeron la matriz de transición?
- Clasificaciones interesantes de algunas:
    - Clases de comunicación
    - Periodicidad de estados
- Análisis de las implicaciones de las propiedades anteriores de la Cadena
- Ejemplos de algunas probabilidades a *n* pasos y su análisis
- Uso de las ecuaciones de Chapman-Kolmogorov para esta aplicación
- Mostrar fragmentos de secuencias de texto generadas que ayuden a ilustrar los análisis teóricos anteriores
- Conclusiones que incluyan los casos o hallazgos más interesantes.

No es necesario incluir matrices o secuencias muy grandes (como el vocabulario), más sí algunos segmentos que las ilustran y mostrar su estructura.

La entrega de este documento será en pdf por correo.

Sesión de entrega

La presentación del proyecto se realizará en clase a forma de sustentación con una duración aproximada de 15 minutos por pareja. Además se entregará el documento acompañante.

Rúbrica de evaluación

Criterio	1	2	3	5
Formulación del Modelo Markoviano	Modelo ausente o erróneo.	Modelo parcial o con fallas.	Modelo básico correcto y funcional.	Modelo riguroso, completo y bien justificado.
Análisis Teórico y Práctico	Análisis incorrecto/irrelevante.	Análisis con errores conceptuales.	Análisis básico correcto. Falta profundidad o detalle.	Análisis completo y profundo.
Aplicación Ecuaciones Chapman-Kolmogorov	Aplicación incorrecta.	Errores serios.	Cálculo básico correcto. Falta verificación y/o interpretación detallada.	Aplicación, verificación e interpretación completas y rigurosas.
Calidad de la Sustentación Oral	Incomprensible/Irrelevante.	Parcial, poco clara, dificultad.	Cubre puntos principales. Respuestas aceptables.	Excepcional: Clara, completa y demuestra dominio.
Calidad del Informe Escrito	Incompleto/erróneo.	Parcial, con fallas notables.	Cubre puntos principales. Formato/redacción aceptables.	Excepcional: Riguroso, conciso, bien presentado.