Actividad 2: Análisis sintáctico

Objetivos

Con esta actividad se pretende implementar el algoritmo CKY probabilístico y aplicarlo para realizar el análisis sintáctico automático de una oración.

Pautas de elaboración

El algoritmo CKY probabilístico usa una gramática libre de contexto probabilístico. La gramática que se va a usar en esta actividad se encuentra en formato CNF (*Chomsky Normal Form*) y está compuesta de reglas que tienen asociada la correspondiente probabilidad de que se use la regla durante el análisis de una oración.

La gramática se define como sigue:

***G= (T, N, R,S)***

**Conjunto de terminales:** *T= {time, flies, arrow, an, like}*

**Conjunto de no terminales:** *N= {S, NP, VP, PP, Det, Nominal, Verb, Preposition, Noun}*

**Conjunto de reglas:**

*R={*

*S → NP VP 0.800*

*NP → time | flies | arrow 0.002*

*NP → Det Nominal 0.300*

*NP → Nominal Nominal 0.200*

*Nominal → time | flies | arrow 0.002*

*Nominal → Nominal Noun 0.100*

*Nominal → Nominal PP 0.200*

*VP → time 0.004*

*VP → flies| like 0.008*

*VP → Verb NP 0.300*

*VP → Verb PP 0.200*

*PP → Preposition NP 0.100*

*Verb → time 0.010*

*Verb → flies| like 0.020*

*Noun → time | flies | arrow 0.010*

*Det → an 0.050*

*Preposition → like 0.050*

*}*

**Axioma de la gramática:** *S*

Se debe realizar el análisis sintáctico de la frase:

*«Time flies like an arrow».*

Los pasos que se deben realizar en la actividad son:

Paso 1. Implementar en Python la estructura de la matriz de análisis sintáctico (CKY).

Paso 2. Implementar en Python la representación de la gramática proporcionada.

Paso 3. Implementar en Python el algoritmo CKY probabilístico.

Paso 4. Aplicar el algoritmo a la frase que se quiere analizar usando la matriz tabla, operando por columnas de abajo arriba y de izquierda a derecha. Para cada celda se deben determinar los constituyentes sintácticos posibles y calcular sus probabilidades.

Nota: tomar decisiones sobre cómo implementar el contenido de las celdas para poder aplicar el cálculo de la ruta de Viterbi más probable.

Paso 5. Implementar el algoritmo para el cálculo de la ruta de Viterbi más probable. Como resultado se debe obtener el árbol de análisis sintáctico más probable.

Proponer una forma de representación del árbol.

En el informe de la actividad se deben explicar las estructuras de datos y la implementación realizada, además de adjuntarse el código correspondiente.

Por último, se debe reflexionar sobre los resultados obtenidos, interpretarlos y analizar el rendimiento de aplicar el algoritmo CKY probabilístico para realizar el análisis sintáctico y sus limitaciones. Para ello se debe responder de forma razonada a las siguientes preguntas:

* ¿Es correcto el análisis sintáctico que se ha obtenido? Justifica la respuesta.
* ¿Cuáles son las limitaciones de aplicar el algoritmo CKY probabilístico para realizar el análisis sintáctico? Justifica la respuesta.
* ¿Qué posibles mejoras que se podrían aplicar para mejorar el rendimiento del análisis sintáctico? Justifica la respuesta.

**Extensión y formato**

Extensión máxima: 8 páginas, fuente Calibri 12 e interlineado 1,5.

**Rúbrica y criterios de evaluación**

|  |  |  |  |
| --- | --- | --- | --- |
| Análisis sintáctico  (valor real: 3 puntos) | Descripción | Puntuación máxima sobre 10  (puntos) | Peso  % |
| Implementación de las estructuras de datos para la matriz CKY y la gramática. | Se implementan las estructuras adecuadas que puedan ser usadas en la implementación de los algoritmos y la obtención de resultados. | 2 | 20 % |
| Implementación y aplicación del algoritmo de construcción de la matriz CKY | Se implementa el algoritmo y se calculan correctamente las probabilidades de las celdas de la matriz correspondiente a la frase que se analiza. | 3 | 30 % |
| Implementación del algoritmo para el cálculo de la ruta de Viterbi más probable | Se implementa correctamente el algoritmo. | 2 | 20 % |
| Árbol sintáctico | El árbol sintáctico resultante del análisis sintáctico de la oración es correcto. | 1 | 10 % |
| Reflexión sobre el análisis sintáctico | Las respuestas a las preguntas sobre el análisis de los resultados obtenidos al aplicar el algoritmo CKY probabilístico para realizar el análisis sintáctico, sus limitaciones y las posibles mejoras son correctas. | 2 | 20 % |
|  |  | **10** | 1. **%** |

Se debe entregar un informe con la descripción detallada de todos los pasos seguidos para realizar la actividad y los resultados intermedios que se van obteniendo en cada paso. Además, se debe detallar toda la información que se solicita en el enunciado.

El código implementado se deberá entregar como un anexo al informe. Si el código se entrega en un *notebook*, este se puede usar para incorporar toda la información que se requiere de cada paso y con un apartado final para las reflexiones. Además, se puede entregar el código con fuente de Python, debidamente comentado y adjuntarlo al informe que es obligatorio entregar.

Se valorará la explicación clara y argumentada.

Si solo se presenta el resultado final, independientemente de que este sea correcto o no, la actividad se calificará con cero puntos.

Se pueden usar bibliotecas de Python que den soporte a la implementación de las estructuras de datos. El resto de los recursos disponibles para el procesamiento del lenguaje natural como, por ejemplo, bibliotecas, API… no se deben utilizar. Si se utiliza alguno de estos recursos para resolver la actividad y no se realizan de forma manual todos los pasos descritos en el enunciado, la actividad se calificará con cero puntos.

El informe de la actividad debe tener el formato marcado en el apartado «Extensión máxima», que no puede ser sobrepasada. Por lo tanto, si el informe presentado tiene más páginas que el valor establecido, solo se corregirán y evaluarán las N primeras páginas. Por ejemplo, si se presenta un documento con 10 páginas y la extensión máxima es de 8 páginas, se evaluará solamente el contenido de las páginas 1 a la 8, por lo que todo el contenido presentado en las páginas 9 y 10 no contará para la nota de la actividad.

Se pueden añadir anexos al informe que contengan, por ejemplo, código, figuras, etc. Estos anexos no computan para la extensión máxima del trabajo, pero tampoco se valorará su contenido para la calificación final de la actividad. Por lo tanto, el informe debe ser entendible por sí solo y responder a todas las cuestiones planteadas en la actividad sin necesidad de recurrir al contenido presentado en los anexos.

Si se detecta plagio en el informe o actividades que copien parte de las respuestas de otro alumno, todos los alumnos involucrados obtendrán una calificación para la actividad de cero puntos.