UNIVERSIDAD NACIONAL DE ASUNCIÓN

FACULTAD POLITÉCNICA – INGENIERÍA INFORMÁTICA

DISEÑO DE COMPILADORES

TRABAJO PRÁCTICO – GRUPO 2

TOKENIZADORES EN EL PROCESAMIENTO DE LENGUAJES NATURALES (NPL)

MOTIVACIÓN. Las herramientas de procesamiento de lenguajes se aplican sobre los lenguajes naturales y necesitan contemplar diferentes situaciones que se presentan como consecuencia propias de estos. Se pueden mencionar a las ambiguedades, expresiones literarias o modismos propios de zonas geográficas o contextos específicos como los principales. El analizador léxico tiene que la capacidad de manejar las situaciones anteriores, para lo que se apoya en en herramientas específicas, tales como preprocesadores léxicos que facilitan la tarea de identificación de palabras.

Los tokenizadores son una herramienta de procesamiento que pueden trabajar en identifcar y organizar palabras de tal manera a simplificar el proceso posterior de analisis sintáctico.

Objetivo del Trabajo Práctico

Construir un tokenizador mínimo, MNLPTK por sus siglas en inglés (a Minimal Natural Language Processing Tokenizer), que actúe como una solución de speech analytics. Para ellos podrá identificar palabras (lexemas) en un texto de entrada en idioma español, que se entiende es el resultado de una interacción telefónica. Procesará los lexemas identificados de manera a generar una ponderación sobre la llamada en general considerando una evaluación en términos de:

A. El desempeño del funcionario que atiende la llamada. Se busca poder obtener una métrica que califique si el funcionario se desempeñó de acuerdo a criterios establecidos.
B. La evaluación de la llamada en general. Para este caso se pretende evaluar la experiencia del cliente, es decir si se pudo resolver o satisfacer la necesidad del cliente. Esta evaluación no siempre depende del desempeño del funcionario, ya que el nivel de servicio o calidad del producto adquirido por el cliente juega un rol principal.

Estos dos indicadores puede permitir tomar acciones que busquen fidelizar al cliente, para lo cual:

- Un bajo desempeño del funcionario, puede fortalecerse con capacitación adicional.
- Una experiencia negativa de la llamada, tiene que ver con una reacción de seguimiento por lo general a cargo del departamento comercial, con el objeto de mejorar la experiencia del cliente (descuentos, promociones, bonificaciones, resolución de problemas puntuales, otros).

El tokenizador retorna una nota de evaluación para el desempeño del funcionario y una serie de TOKENS que dan una idea de la justificación y una nota de evaluación de la experiencia del cliente y una serie de TOKENS que dan una idea de los ocurrido durante la llamada.

Descripción del alcance del Trabajo Práctico

Los tokens definidos para el lenguaje de entrada en el texto deben permitir obtener una evaluación de la atención y de la experiencia del cliente. Por lo tanto se podrían definir TOKENS tales como EXP_MALA, EXP_NEUTRA, EXP_BUENA y ATC_MALA, ATC_NEUTRA, ATC_BUENA.

Para un procesamiento más eficaz, la llamada se transcribe en dos archivos, que serán las entradas al tokenizador. En el primer archivo se encuentra la interacción únicamente del personal de atención al cliente. En el segundo, por el contrario, se transcribe solamente la interacción del cliente. Esto permite poder dividir al tokenizador en dos procesos bien especializados, para realizar la evaluación.

El funcionamiento entonces se podría encarar de la siguiente manera: El programa comienza un ciclo de lectura, primeramente del archivo del funcionario de atención al cliente, y de acuerdo a los patrones definidos puede obtenerse una evaluación de la gestión. En forma posterior, se realiza la lectura y evaluación de la interacción del cliente.

Algunos criterios que se deben tener en cuenta son:

a. Experiencia de atención: Saludo de bienvenida (“buen dia”, “buenas tardes”, “buenas noches”) y despedida (“Hasta luego” _consideramos una sola por simplicidad). Identificar al cliente: Utilizar la palabrá “documento” o “cédula” es indicativo que se está pidiendo identificar al cliente. Utilizar palabras tales como “gracias”, “por favor”.
b. Experiencia del cliente: Si utiliza palabras negativas tales como mal, desastre, harto, cansado, cancelar y otras debería sumar calificaciones negativas. De forma similar (antónimos) se sumaran calificaciones positivas. El contraste de ambos indicadores podría generar un valor en una escala del 1 al 5 de la experiencia del cliente con el servicio en general.

   Para el caso de las palabras neutras, se puede considerar aquellas que no son ni buenas o malas. Al terminar la lectura de ambos archivos se podría preguntar si existe alguna palabra neutra que se desee incorporar como mala o como buena. Es decir, la posibilidad que manualmente se pueda actualizar estas listas.

Al finalizar el proceso de análisis léxico de los archivos de entrada, se tiene:

a. Una ponderación para experiencia de atención y una para la del cliente.
b. El detalle de como se llegó a esa ponderación:
   - Experiencia de atención: que si cumplió y que no.
   - Experiencia del cliente: listado de palabras malas, buenas.
c. La posibilidad de recorrer las palabras neutras y asignarlas ya sea como malas o buenas. Se actualizarán los patrones de los tokens conforme hayan asignaciones válidas.

MEJORAS: Es posible tener en cuenta:

a. NÚMERO Y GENERO. El proceso de comparación con el patrón puede arrojar una raíz común y diferenciarse en número o gémero. Por ejemplo para el caso de palabras buenas o malas considerar el número y genero: mal, agregar malo, mala, malos, malas, malísimo, etc.
b. Interface gráfica amigable.

METODOLOGIA. Utilizar los conocimiento dados en clases para construir el tokenizador, teniendo en cuenta que es posible:

1. Utilizar el algoritmo de simulación de un AFD para obtener los lexemas, para patrones definidos por una expresión regular por comprensión.
2. Utilizar herramientas IA para el soporte y ayuda tal como CHATGPT u otras herramientas relacionadas. La codificación, ejecución y procesamiento debe ser responsabilidad del alumno. Observación: *Es probable que las herramientas de IA generen conocimientos inexactos, incompletos o incorrectos que deberán ser identificados y corregidos apropiadamente*.

El lenguaje de programación a utilizar para la implementación será elegido libremente el alumno.

ENTREGABLES Y DEFENSA DEL TRABAJO PRÁCTICO. Son entregables del trabajo práctico:

a. Un documento en PDF donde:
    1. se describa el trabajo práctico, las decisiones adoptadas, mejoras y todo lo que contribuya a defnir e alcance y estrategias aplicadas.
    2. El código fuente.
    3. El resultado para un caso de ejemplo.
    4. Cualquier observación sobre el funcionamiento o situaciones no resueltas o no contempladas.
b. La defensa del trabajo práctico es presencial y no puede recuperarse. Es requisito para rendir el examen final. En caso de ausencia no se aceptará la entrega en forma electrónica sin defensa.

Al finalizar el trabajo, se comprenderá el impacto que se tiene en utilizar el analizador léxico como herramienta de procesamiento de lenguajes. En forma simétrica y comparando, se podrán generar corolarios en la convenicia de utilización del analizador léxico o sintáctico como protagonista de una estrategia de procesamiento de lenguaje y el rol de un tokenizador para el análisis de la entrada de datos.

Evaluación

El proceso de evaluación tendrá en cuenta:

a. una parte práctica, que se basa en la defensa del trabajo práctico, donde se presentará sus funcionalidades y ejemplos con cadenas para procesar. Se evaluará su eficiencia y su correcto funcionamiento y validaciones. La interface es opcional.
b. Una parte teórica, donde se evalua el contenido del documento PDF base del trabajo práctico y el algoritmo obtenido.

Observaciones

a. El trabajo práctico no es recuperable.
b. Fecha de entrega: el último lunes de clases antes de la interrupción para el periodo del segundo parcial. Cada alumno dispondrá de hasta un máximo de de 15 minutos para la defensa del TP.
c. El trabajo práctico se debe realizar en forma individual.