INSTITUTO POLITÉCNICO NACIONAL
Escuela Superior de Cómputo

Tecnologías de Lenguaje Natural
Profesor: Ituriel Enrique Flores Estrada

Septiembre 2024

PRÁCTICA 1

Generar el reporte de dos programas en Python (Jupyter Notebook) para hacer uso de expresiones regulares y para la normalización de documentos. Cada programa deberá corresponder, respectivamente, a la Parte 1 y a la Parte 2 descritas a continuación:

PARTE 1. EXPRESIONES REGULARES

Para las cadenas de texto incluidas en el Anexo “A”, y únicamente para los numerales siguientes 1 a 5, identificar las líneas que cumplan con los siguientes:

1. Contengan una “r” seguida por una “g”. La “r” y la “g” no necesariamente tienen que estar en posiciones consecutivas. **(2 puntos).**

2. Describan comidas que cuesten al menos 100.00. **(2 puntos)**

3. Contengan una “a”, seguida por una “b”, seguida por una “c” (puede haber otros caracteres entre la “a” y la ”b” y entre la “b” y la ”c”. En caso de existir caracteres entre las letras indicadas, dichos caracteres no pueden ser a ni b, ni b y c, respectivamente. Ejemplos de cadenas invalidas: “A apple, a banana”; “Bad but beatiful car” **(2 puntos)**

4. Contengan en la descripción de gastos una “a” minúscula y un digito entre 0 y 9 en cualquier orden. Es decir, el carácter “a” puede aparecer antes o después del digito. **(2 puntos)**

5. Contengan el carácter “d”, **posiblemente** seguido de otros caracteres, seguido de una “i”. Coincidencias incluirían palabras tales como: diver, doily, drip, diplomat, etc. **(2 puntos)**

A continuación, se describen ejercicios adicionales de expresiones regulares. Cabe señalar que cada uno de los siguientes ejercicios contienen el texto a analizar.

6. Identificar títulos de películas producidas **antes** de 2002. El año de producción es el encerrado en paréntesis. **(2 puntos)**
   a. The Shawshank Redemption (1994)
   b. The Godfather (1972)
   c. The Godfather: Part II (1974)
   d. 2001: A Space Odyssey (1968)

e. The Good, the Bad and the Ugly (1966)
f. Angry Men (1957)
g. Schindler's List (1993)
h. The Lord of the Rings: The Return of the King (2003)
i. Fight Club (1999)
j. 2010: The Year We Make Contact (1984)
k. 101 Dalmatians (1996)

7. Identificar recetas que contengan la palabra 'chocolate' y cualquier variación de ella en términos de combinación entre minúsculas y mayúsculas, y también repetición de caracteres. Por ejemplo, una cadena válida puede ser: “ChOcoolATe. **(2 puntos)**
   a. Cake 1: sugar, flour, cocoa powder, baking powder, baking soda, salt, eggs, milk, vegetable oil, vanilla extract, chocolATE chip.
   b. Cake 2: cream cheese, sugar, vanilla extract, crescent rolls, cinnamon, butter, honey.
   c. Cake 3: dark chocolate cake mix, instant CHOCOLATE pudding mix, sour cream, eggs, vegetable oil, coffee liqueur.
   d. Cake 4: flour, baking powder, salt, cinnamon, butter, sugar, egg, vanilla extract, milk, chopped walnuts.
   e. Cake 5: gingersnap cookies, chopped pecans, butter, cream cheese, sugar, vanilla extract, eggs, canned pumpkin, cinnamon, CHOColate.
   f. Cake 6: flour, baking soda, sea salt, butter, white sugar, brown sugar, eggs, vanilla extract, Chooocolate chips, canola oil.
   g. Cake 7: wafers, cream cheese, sugar, eggs, vanilla extract, cherry pie filling.

8. Insertar comas entre grupos de cada tres dígitos para las siguientes poblaciones por país. **(4 puntos)**
   a. China 1361220000
   b. India 1236800000
   c. United States 317121000
   d. Indonesia 237641326
   e. Brazil 201032714
   f. Pakistan 184872000
   g. Nigeria 173615000
   h. Bangladesh 152518015
   i. Russia 143600000

9. Simplificar direcciones IPv6. **(4 puntos)**

   Una dirección IP está compuesta de 8 bloques de números hexadecimales. Cada bloque está de cuatro dígitos y los bloques están separados por el signo “:” (dos

puntos). Adicionalmente, existen las siguientes reglas para simplificar y reducir el tamaño de una dirección IPv6:

a) Los bloques compuestos únicamente de ceros pueden ser omitidos.
b) Los ceros al principio de un bloque pueden ser omitidos.

A continuación, algunos ejemplos:

Dirección original (extendida): 2001:0db8:0000:0000:0000:ff00:0042:8329
Despúes de remover ceros al inicio de un bloque: 2001:db8:0:0:0:ff00:42:8329
Despúes de eliminar bloques de ceros: 2001:db8::ff00:42:8329
Dirección final (simplificada): 2001:db8::ff00:42:8329

Dirección original (extendida): 2607:f0d0:1002:0051:0000:0000:0000:0004
Despúes de remover ceros al inicio de un bloque: 2607:f0d0:1002:51:0:0:0:4
Despúes de eliminar bloques de ceros: 2607:f0d0:1002:51::4
Dirección final (simplificada): 2607:f0d0:1002:51::4

PARTE 2. NORMALIZACIÓN DE TEXTOS

Los textos en español e inglés a analizar y procesar son los incluidos en el Anexo B y deben ser guardados en dos archivos, uno por cada documento, para ser leídos y cargados en memoria desde el programa a desarrollar.

El reporte y el código, respectivamente, deberá incluir lo siguiente:

1. **Investigación (03 puntos).** Identificar los módulos de Python adecuados para procesar textos en español e inglés, respectivamente. En el reporte se debe señalar las diferencias funcionales entre los respectivos módulos e indicar las razones por las cuales el alumno los eligió para procesar los textos provistos.

2. **Análisis exploratorio de texto (15 puntos).** Se deberá generar el código necesario para explorar los textos y familiarizarse con el contenido de los mismos. De forma enunciativa más no limitativa se debe:

   a) Identificar el número total de tokens en el texto.
   b) Identificar el número de tokens únicos en el texto.
   c) Desplegar en un histograma los 15 tokens más comunes.
   d) Desplegar en un histograma los 15 tokens más comunes.

3. **Normalización de texto (15 puntos).** General el código necesario para remover términos no relevantes (stop words), “lematizar” y hacer “stemming” de cada uno

de los textos. Adicionalmente, y con la base en lo observado en el punto anterior, el alumno debe **proponer, justificar y ejecutar al menos tres técnicas de normalización adicionales** para cada texto. El total de tareas de normalización debe ser de al menos 6.

El orden en que se ejecutan las técnicas de normalización impacta el documento resultado de dicho flujo. Por tanto, el reporte debe incluir **la justificación del orden en que el alumno decidió ejecutar cada una de las tareas de normalización**.

4. **Exploración de texto posterior a normalización (0 puntos).** Realizar lo mismo solicitado en el punto 2 pero esta vez sobre los textos obtenidos después de la etapa de normalización.

5. **Análisis y conclusiones (25 puntos).** Se deben contrastar los resultados obtenidos en los puntos 2 y 4 para identificar las diferencias entre el texto original y el documento resultado después de haber sido normalizado. En específico, se debe indicar y argumentar el efecto que pudo o no haber tenido cada una de las técnicas de normalización aplicadas sobre los respectivos documentos.

Reporte formal y comentarios en código (20 puntos). El reporte debe ser formal por lo que **al menos** debe estar divido en secciones y nombrar todas las tablas e imágenes incluidas en el documento. Además, cada sección debe: describir su propósito y el resultado de las tareas ejecutadas sobre el texto o cuerpo de documentos; así como incluir evidencia de dichos resultados. Asimismo, el código deberá incluir una cabecera y comentarios relevantes conforme a lo detallado en el documento “ReglasEvaluación” publicado al inicio del semestre.

Notas:

1. Las imágenes de código no son evidencia de funcionamiento.
2. En caso de no entregar reporte se asignará cero en la calificación de la práctica.

ANEXO "A".
Cadenas de texto a analizar para la Parte 1 de la práctica

Amount:Category:Date:Description
5.25:supply:20170222:box of staples
79.81:meal:20170222:lunch with ABC Corp. clients Al, Bob, and Cy
43.00:travel:20170222:cab back to office
383.75:travel:20170223:flight to Boston, to visit ABC Corp.
55.00:travel:20170223:cab to ABC Corp. in Cambridge, MA
23.25:meal:20170223:dinner at Logan Airport
318.47:supply:20170224:paper, toner, pens, paperclips, tape
142.12:meal:20170226:host dinner with ABC clients, Al, Bob, Cy, Dave, Ellie
303.94:util:20170227:Peoples Gas
121.07:util:20170227:Verizon Wireless
7.59:supply:20170227:Python book (used)
79.99:supply:20170227:spare 20" monitor
49.86:supply:20170228:Stoch Cal for Finance II
6.53:meal:20170302:Dunkin Donuts, drive to Big Inc. near DC
127.23:meal:20170302:dinner, Tavern64
33.07:meal:20170303:dinner, Uncle Julio's
86.00:travel:20170304:mileage, drive to/from Big Inc., Reston, VA
22.00:travel:20170304:tolls
378.81:travel:20170304:Hyatt Hotel, Reston VA, for Big Inc. meeting
1247.49:supply:20170306:Dell 7000 laptop/workstation
6.99:supply:20170306:HDMI cable
212.06:util:20170308:Duquesne Light
23.86:supply:20170309:Practical Guide to Quant Finance Interviews
195.89:supply:20170309:black toner, HP 304A, 2-pack
86.00:travel:20170317:mileage, drive to/from Big Inc., Reston, VA
32.27:meal:20170317:lunch at Clyde's with Fred and Gina, Big Inc.
22.00:travel:20170317:tolls
119.56:util:20170319:Verizon Wireless
284.23:util:20170323:Peoples Gas
8.98:supply:20170325:Flair pens

ANEXO "B"

1. Documento en español

Por lo general, no pensamos en las complejidades de nuestros propios lenguajes. Es un comportamiento intuitivo que se utiliza para transmitir información y significados con señales semánticas, como palabras, signos o imágenes. Se dice que es más fácil aprender un idioma nuevo cuando somos adolescentes porque se trata de un comportamiento repetible y entrenado, casi como caminar. Asimismo, el idioma no sigue un conjunto de reglas estricto, ya que las excepciones son innumerables, por ejemplo: los sustantivos que terminan con 'a' son femeninos, pero no es el caso del sustantivo 'el problema'. Sin embargo, a los humanos nos resulta natural es extremadamente difícil para las computadoras, ya que tienen que lidiar con una gran cantidad de datos no estructurados, la ausencia de reglas formales y la falta de un contexto o una intención real. Es por eso que el aprendizaje automático y la inteligencia artificial (IA) ganan fuerza y llaman la atención, puesto que los humanos dependen cada vez más de los sistemas informáticos para comunicarse y realizar tareas. A medida que la IA se vuelve más sofisticada, también lo hace el procesamiento del lenguaje natural (PLN).

El Procesamiento del Lenguaje Natural es el campo de conocimiento de la Inteligencia Artificial que se ocupa de la investigar la manera de comunicar las máquinas con las personas mediante el uso de lenguas naturales, como el español, el inglés o el chino. Virtualmente, cualquier lengua humana puede ser tratada por los ordenadores. Lógicamente, limitaciones de interés económico o práctico hace que solo las lenguas más habladas o utilizadas en el mundo digital tengan aplicaciones en uso.

Pensemos en cuántas lenguas hablan Siri (20) o Google Assistant (8). El inglés, español, alemán, francés, portugués, chino, árabe y japonés (no necesariamente en este orden) son las que cuentan con más aplicaciones que las entienden. Google Translate es la que más lenguas trata, superando el centenar… pero hay entre 5000 y 7000 lenguas en el mundo.

2. Documento en inglés

My own journey toward language was sparked in 1996 when I discovered Keith Basso's astonishing book Wisdom Sits in Places. Writing about the unique place-making language of the Western Apache, Basso described language in a way that I'd never considered before, as roots and fragments strung together to sing of the land. This idea intrigued me so much that I began carrying Donald Borror's classic little book, the Dictionary of Word Roots and Combining Forms, with me on all my hikes (a practice which I've continued on a daily basis for nearly twenty years and on thousands of miles of trails) in order to learn

the meaning and origin of word elements at the moment they occurred to me while walking in wild landscapes.

For many years this seemed little more than a quirky hobby, with no real intent or direction, but then a friend introduced me to Calvert Watkins's magisterial survey of Indo-European poetics, How to Kill a Dragon. In a flash I realized that there might be untapped ways for the English language to speak of the magic of the land and the depths of the human spirit, so I began a four-year quest to read every book I could find on the history, formation, and word-making processes of the English language.

What you hold here is the result of my investigation: 76 sections that explore some of the many pieces and processes that have gone into shaping the English language as we use it today. As I researched and wrote each section of this book I carried these ideas with me on long hikes in wild places and held them up against the natural world to see which ideas resonated and which ideas took on a life of their own. This book emerges from and reflects these hikes, and because I also lead walks as a naturalist in my professional life this book is modeled on the metaphysic that I know best—the flow of ideas and observations that arise spontaneously when humans encounter the world with curiosity and wonder.