

TAREA PROGRAMADA 1

La tarea consiste en extraer información del archivo de texto que se les adjunta. El archivo consiste de un ejemplo de *clave dicotómica* usada por los botánicos para ayudar a la identificación de una especie vegetal. Esencialmente se trata de una gran instrucción *if* en la que se van consultado características específicas del espécimen para ir determinando

```
1 condicionesA
2 condicionesB . . . grupo 1
2' condicionesC . . . grupo 2
1' condicionesD
3 condicionesE . . . grupo 3
3' condicionesF . . . grupo 4
```

El ejemplo anterior corresponde a:

```
if (condicionesA)
    if (condicionesB)      then grupo1()
    else if (condicionesC) then grupo2()
else if (condicionesD)
    if (condicionesE)      then grupo3()
    else if (condicionesF) then grupo4()
```

Como se nota, no hay condiciones else: los grupos deben identificarse explícitamente por el cumplimiento de algunas condiciones.

La tarea consiste en tomar dicho archivo y realizar las siguientes operaciones:

Sustituir abreviaturas

Usar la siguiente tabla para reemplazar las abreviaturas listadas por el texto correspondiente. Las abreviaturas vienen terminadas por un punto, el cual se debe eliminar. Se debe ignorar la diferencia entre mayúsculas y minúsculas al localizar y sustituir las abreviaturas.

Abreviatura	Texto
ca	aproximadamente
carib	caribe
coord	cordillera
pac	pacífico
fl	flor
fls	flores
fr	fruto
frs	frutos
vert	vertiente
verts	vertientes

Extraer y contar las palabras y números del archivo

Una vez sustituidas las abreviaturas, se debe separar el texto en palabras y números, contar las apariciones y listar el resultado en orden lexicográfico.

Antes de separar en palabras se deben eliminar las tildes y convertir todas las letras a minúsculas. La ñe se debe mantener intacta y se debe considerar como una letra.

Los números son secuencias consecutivas de dígitos con un punto opcional en medio de los dígitos. Para que un punto se considere parte de un número, debe ser inmediatamente precedido y seguido por dígitos: en "0.4" el punto se considera parte del número; en "45." el punto no se considera parte del número.

Estructurar los casos de la clave dicotómica

Se deben estructurar los párrafos del texto para hacer explícita la organización de la clave dicotómica. Esto es, para cada caso, se debe generar una línea separada por tabulaciones con la siguiente estructura:

nivel→prima→condiciones→grupo

Se debe asegurar de que el texto no traiga tabulaciones dentro de las condiciones. En el campo prima, se pone un 0 si no aparece y un 1 si aparece; si el caso no tiene asociado un grupo simplemente se deja en blanco pero se incluye la tabulación que le precede.

Los casos están separados en varias líneas consecutivas. Además los casos están separados unos de otros por al menos una línea en blanco. Si un caso incluye un grupo, el grupo estará precedido por varios puntos separados por espacios en blanco.

Por ejemplo:

7' Frs. sostenidos
por una cúpula.

se convierte en:

7**tab**1**tab**Frs. sostenidos por una cúpula.**tab**

Extraer rangos

Finalmente, se deben extraer los rangos numéricos que aparecen en el texto. En el caso más general los rangos tienen una forma como la siguiente: (3-)7-13(-18); esto significa que el rango normal es 7-13, pero opcionalmente se puede dar un rango inferior excepcional 3-7 (sin incluir el 7) y un rango superior excepcional 13-18 (sin incluir el 13).

En algunos casos los rangos pueden ser bidimensionales (longitud x anchura), en cuyo caso los dos rangos se encuentran separados por el carácter "x":

(1-)2.1-3 x 5-8.2(-11.3).

Finalmente, los rangos pueden ser seguidos de una unidad de medida. Dichas unidades si aparecen son uno o dos letras como "m" o "cm".

Se desea representar extraer generando líneas con la siguiente estructura:

tipo→*unidad*→*inf1*→*mín1*→*máx1*→*sup1*→*inf2*→*mín2*→*máx2*→*sup2*

donde:

tipo: es "1" si es un rango de una sola dimensión y "2" si es un rango bidireccional
unidad: texto de la unidad de medida: "cm", "mm", etc.
inf1: límite inferior excepcional del primer sub rango o del único rango
mín1: mínimo usual del primer sub rango o del único rango
máx1: máximo usual del primer sub rango o del único rango
sup1: límite superior excepcional del primer sub rango o del único rango
inf2: para rangos bidimensionales, límite inferior excepcional del segundo sub rango
mín2: para rangos bidimensionales, mínimo usual del segundo sub rango
máx2: para rangos bidimensionales, máximo usual del segundo sub rango
sup2: para rangos bidimensionales, límite superior excepcional del segundo sub rango

Si algún valor no aparece simplemente se omite, pero se preservan las tabulaciones. Pero en caso de un rango tipo 1, los valores asociados con el segundo rango y sus tabulaciones se omiten.

Ejemplos:

```
3-5 mm
1 tab mm tab tab 3 tab 5 tab

0.15-1 (-1.6) m
1 tab m tab tab 0.15 tab 1 tab 1.6

(7-) 9-20 (-26) cm
1 tab cm tab 7 tab 9 tab 20 tab 26

0.4-0.6
1 tab tab 0.4 tab 0.6 tab tab

0.5-2.3 x 0.08-0.5 cm
2 tab cm tab tab 0.5 tab 2.3 tab tab tab 0.08 tab 0.5 tab

(9-) 12-20 (-26) x (3-) 5-9 (-11) cm
2 tab cm tab 9 tab 12 tab 20 tab 26 tab 3 tab 5 tab 9 tab 11
```

Consideraciones finales

Se debe realizar la tarea usando las facilidades de Python para procesar texto.

La tarea es individual y debe entregarse dentro de dos semanas: miércoles 2 de setiembre.

Se deben producir los siguientes tres archivos de texto separados por comas (.CSV). En lugar de comas se usarán tabulaciones.

1. Diccionario: contiene las palabras y números junto con su conteo; ordenado lexicográficamente.
2. Clave: contiene las estructuras de clave identificadas. Debe estar ordenado por el nivel de clave y por su bandera prima.
3. Rangos: contiene las estructuras de los rangos identificados.

Además de estos archivos se debe entregar el código del programa comentado adecuadamente.

