



Procesamiento de strings

Análisis y diseño de algoritmos avanzados

Dra. Valentina Narváez Terán



Tecnológico
de Monterrey

Codificación de datos: Huffman codes

Los códigos de Huffman son una forma de codificar strings

Cada carácter de la cadena original se transforma en un código de longitud variable

Se asignan **códigos mas cortos a los caracteres mas comunes** en la cadena original

¿Cómo funciona?

Con un tipo de árbol binario, llamado **Huffman tree**



Codificación de datos: Huffman codes

El árbol binario sirve como soporte para construir los códigos

Para asignar códigos mas cortos a los caracteres mas comunes, **el árbol se construye con las probabilidades de ocurrencia** de cada carácter

¿De donde salen las probabilidades para cada carácter del texto?

$$\frac{\text{número de ocurrencias}}{\text{total de caracteres}}$$



Codificación de datos: Huffman codes

A	B	C	D	E
0.35	0.1	0.2	0.2	0.15

Cada carácter es la raíz de su propio árbol binario, así que inicialmente, hay $n = 5$ árboles

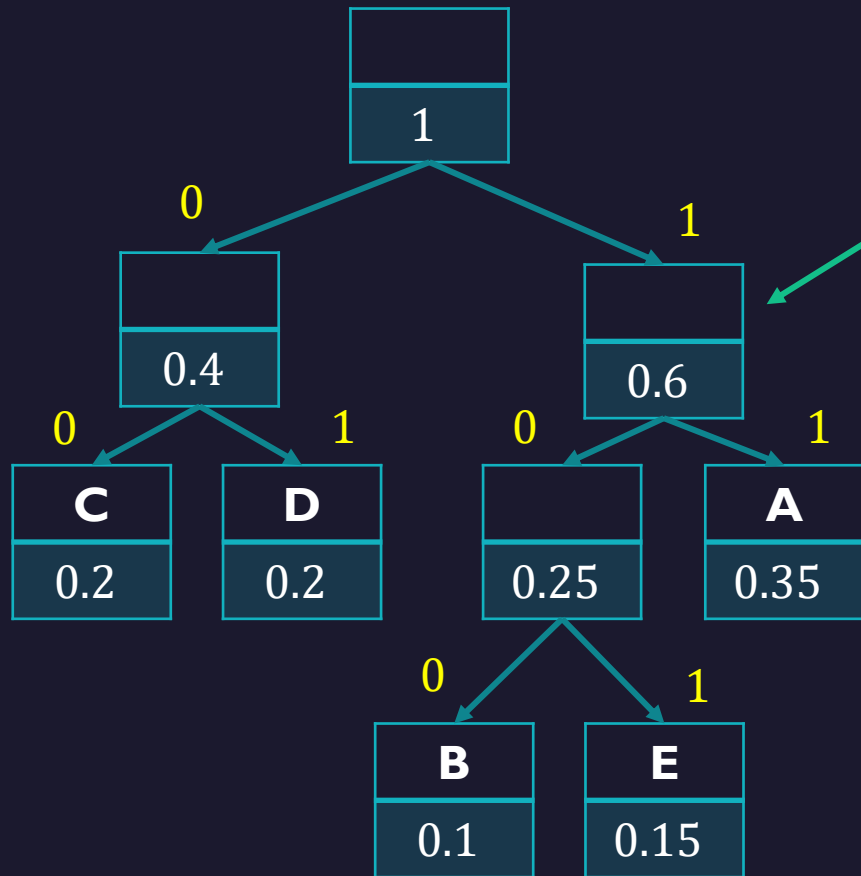
A	B	C	D	E
0.35	0.1	0.2	0.2	0.15

Ejemplo

Supongamos que queremos codificar un texto cualquiera, formado solo con el alfabeto $\{A, B, C, D, E\}$

Eventualmente, todos se volverán parte de un mismo árbol

Huffman codes: el árbol



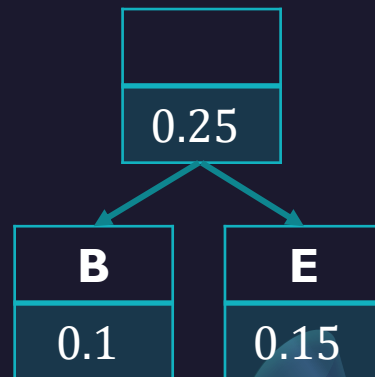
Eventualmente, el árbol final quedaría así

Nota que los caracteres del alfabeto están en nodos hoja

Ejemplo de encoding ¿como funciona?

11 01 11 = ADA

Huffman codes: construcción del árbol

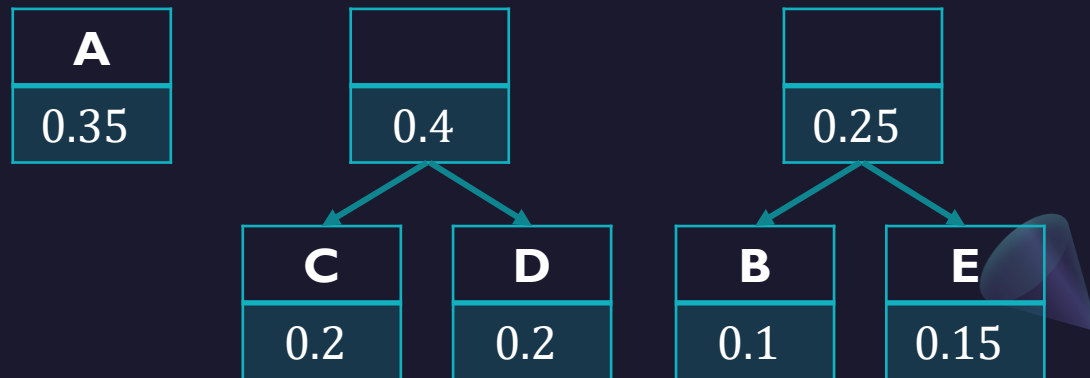
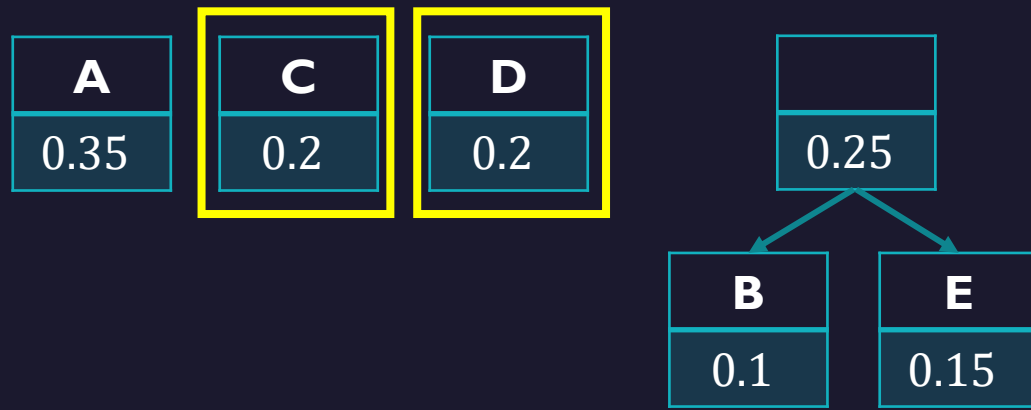


Árbol fusionado

El algoritmo es simple:

1. Encuentra los dos árboles con la menor probabilidad
2. Crea una raíz nueva, cuyo valor es la suma de las dos probabilidades del paso 1
3. Agrega a los dos árboles del paso 1 como hijos de la raíz
4. Y repite hasta que todos sean un único árbol

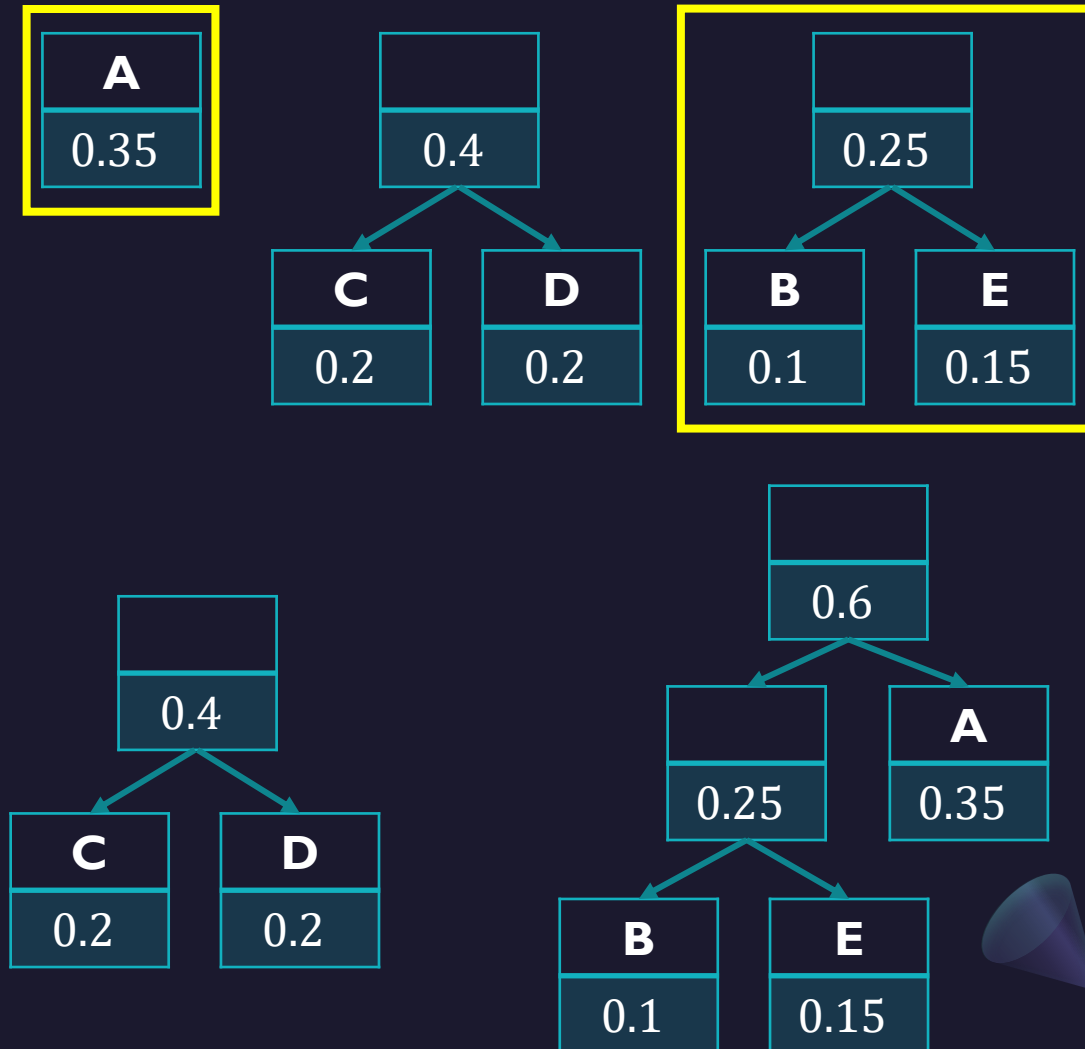
Huffman codes: construcción del árbol



El algoritmo es simple:

1. Encuentra los **dos arboles con la menor probabilidad**
2. Crea una raíz nueva, cuyo valor es la suma de las dos probabilidades del paso 1
3. Agrega a los dos arboles del paso 1 como hijos de la raíz
4. Y repite hasta que todos sean un único árbol

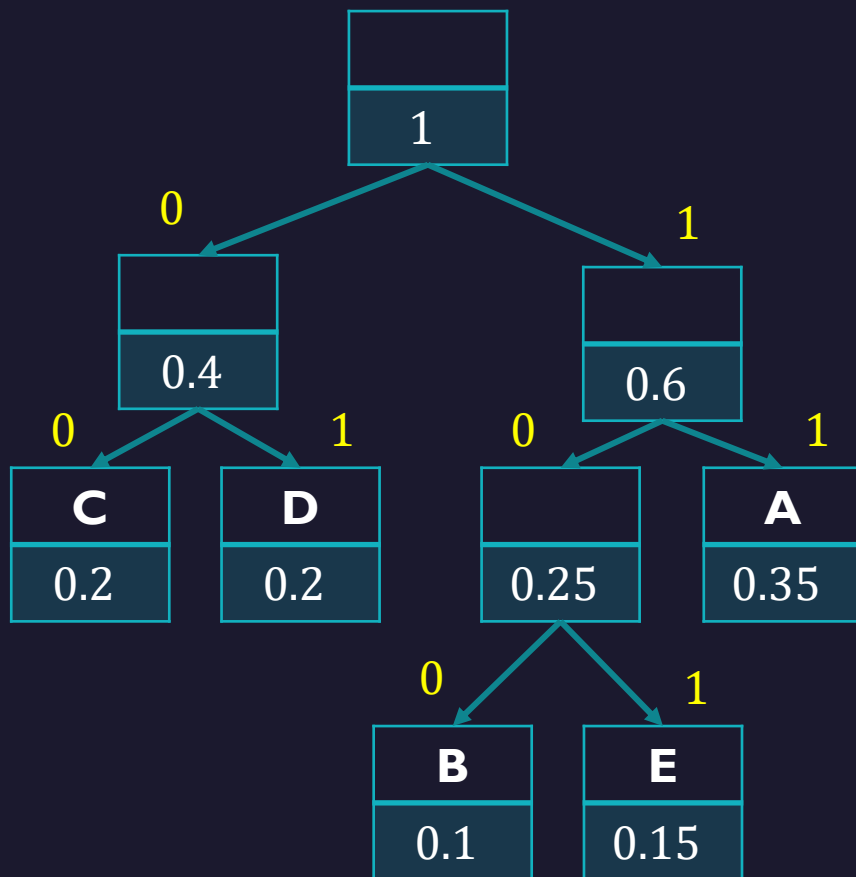
Huffman codes: construcción del árbol



El algoritmo es simple:

1. Encuentra los dos arboles con la menor probabilidad
2. Crea una raíz nueva, cuyo valor es la suma de las dos probabilidades del paso 1
3. Agrega a los dos arboles del paso 1 como hijos de la raíz
4. Y repite hasta que todos sean un único árbol

Huffman codes: construcción del árbol



Al final tendremos un árbol cuya raíz tiene probabilidad 1

¿Cómo se usa este árbol para crear un código?

Con **etiquetas** para los nodos:

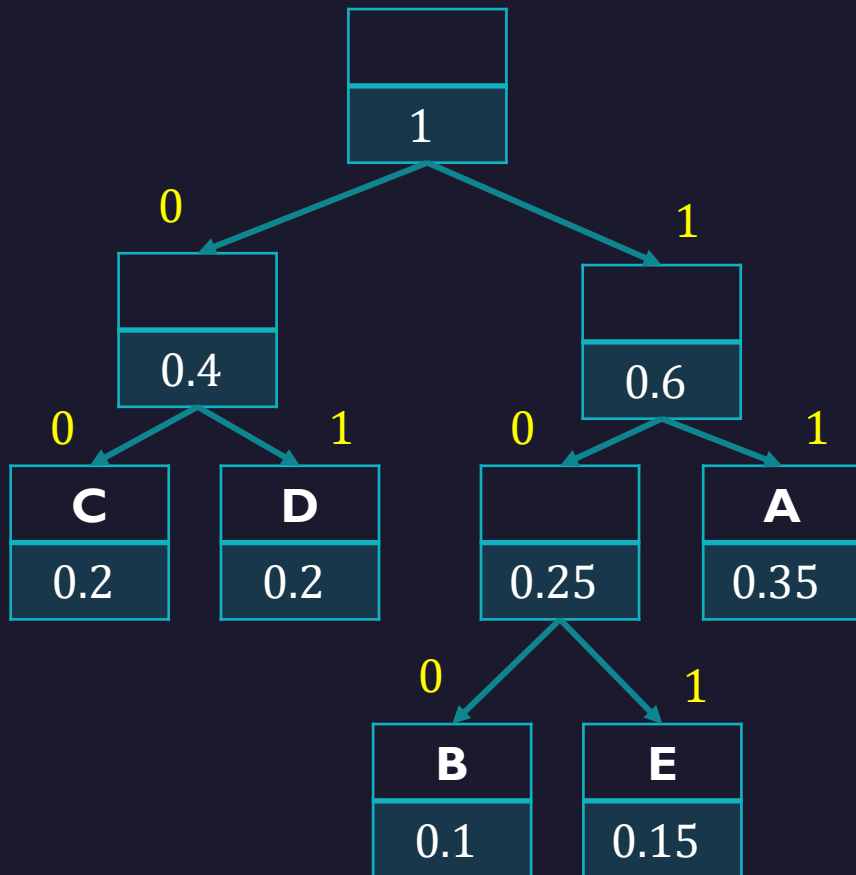
Todos los **hijos izquierdos** tienen **etiqueta 0**.

Todos los **hijos derechos** tienen **etiqueta 1**

Para para codificar, seguimos el camino de la raíz hasta el nodo algún nodo hoja

Su **código** son las **etiquetas** de los nodos por donde pasas

Huffman codes: codificación



A	11
B	100
C	00
D	01
E	101

Códigos de cada carácter

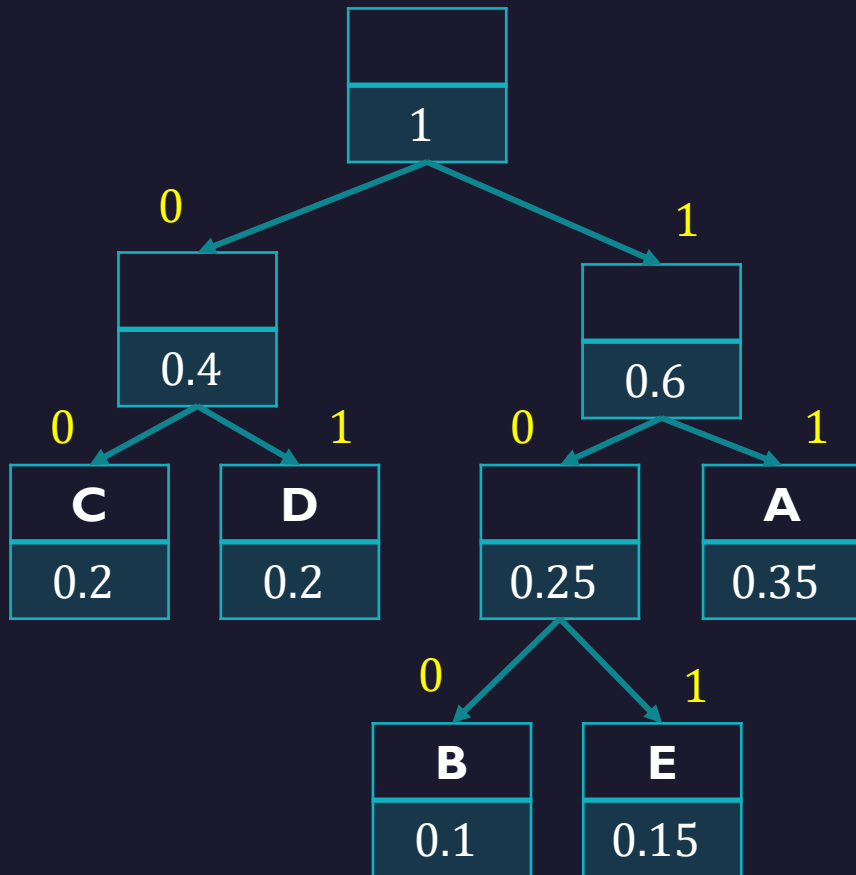
Se usan para codificar la cadena original, sustituyendo cada carácter por su código

Usando el árbol, esta tabla se puede obtener de diferentes formas

¿Cómo lo harías?



Huffman codes: decodificación



El algoritmo para decodificar

Iniciar en la raíz

Por cada carácter del código:

Desplazarse por los nodos según las etiquetas del árbol

Al llegar a una hoja, anotar el carácter y volver a la raíz

Decodifica:

01 01 11 11 01 00 01

Huffman codes: **decodificación**

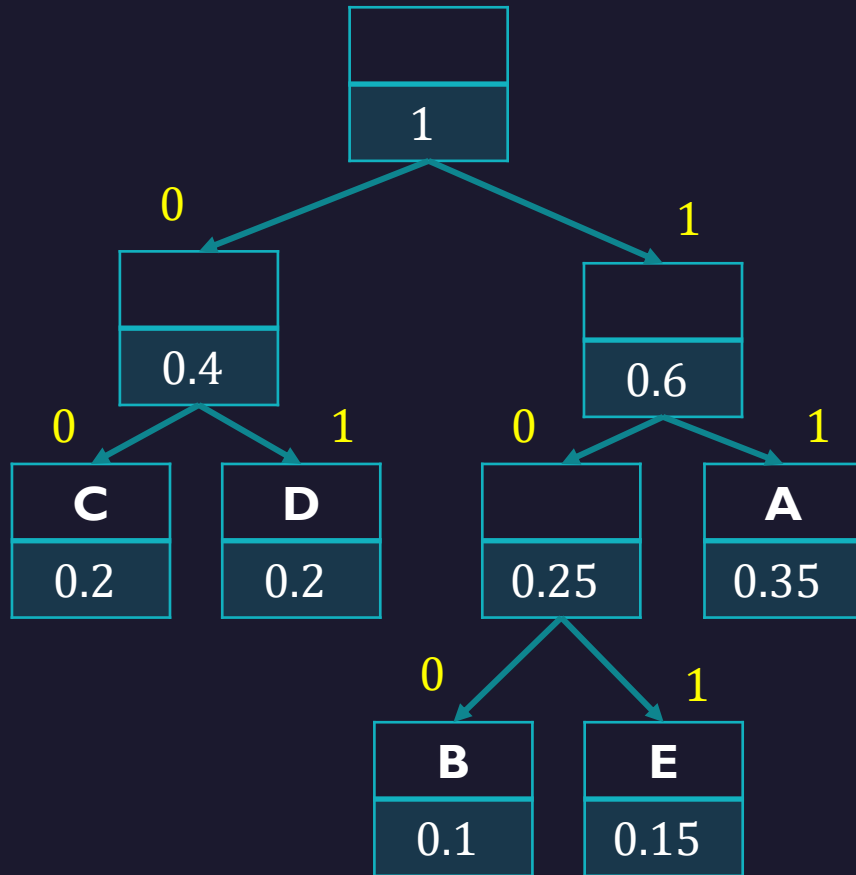
El algoritmo para decodificar

Iniciar en la raíz

Por cada carácter del código:

Desplazarse por los nodos según las etiquetas del árbol

Al llegar a una hoja, anotar el carácter y volver a la raíz



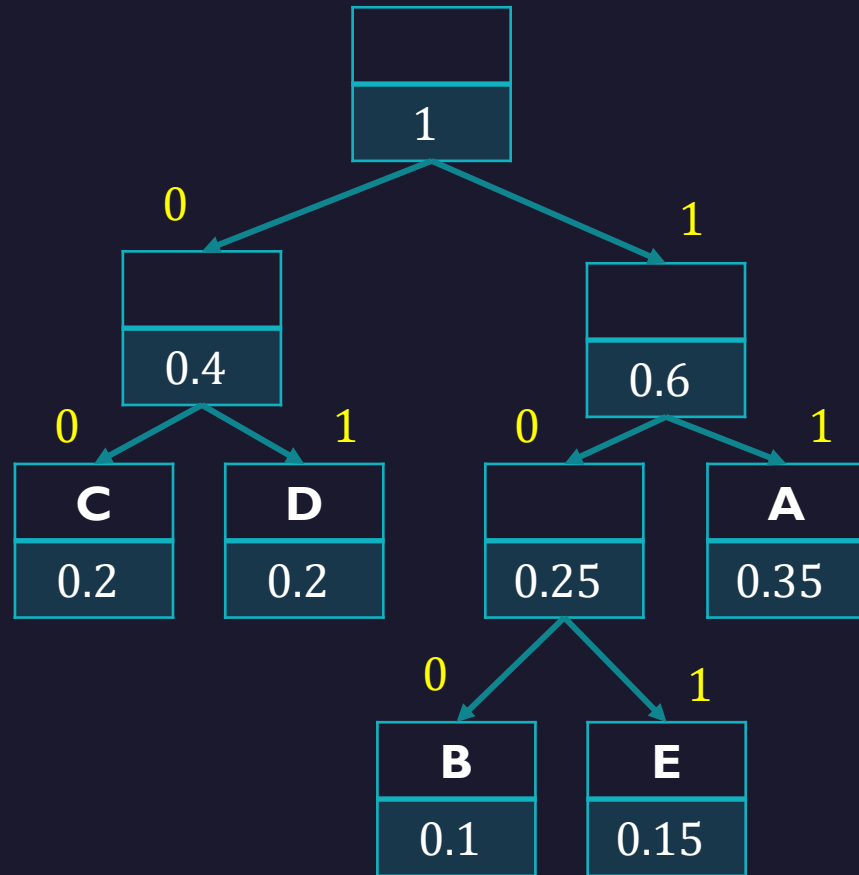
Decodifica:

01 01 11 11 01 00 01

Respuesta:

D D A A D C D

Huffman codes: decodificación

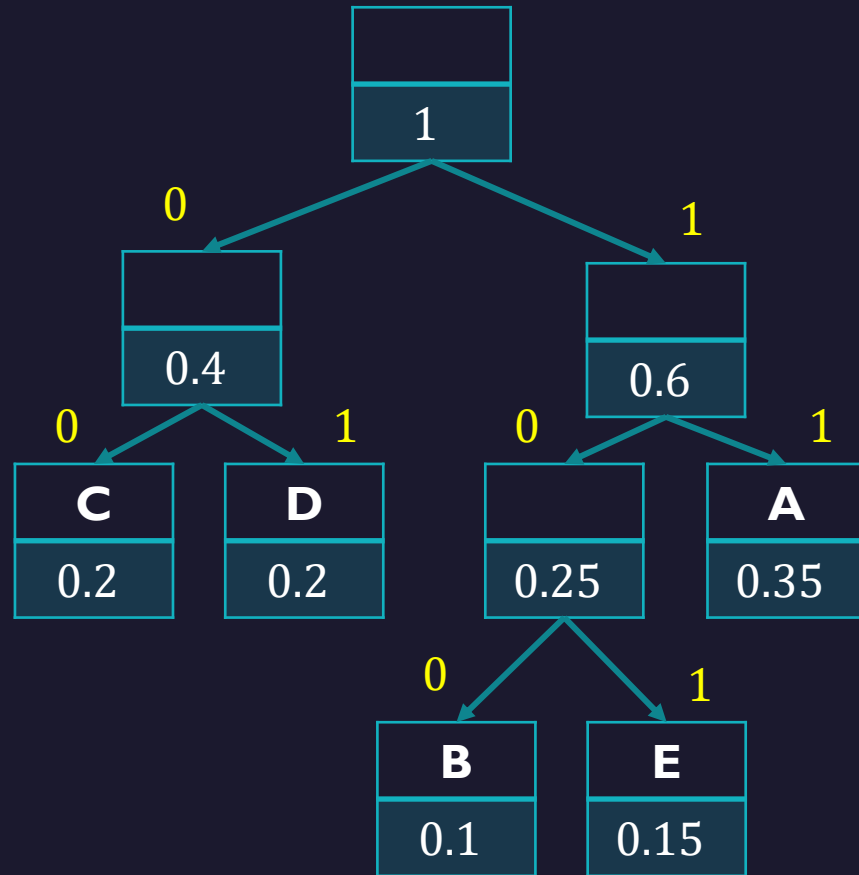


Ejercicios:

11 11 00 00 01

11 11 101 01 01 0

Huffman codes: decodificación



Ejercicios:

11 11 00 00 01
AACCD

11 11 101 01 01 0
AAEDD ... ?

Si hay caracteres que sobran, es un indicador de que el código es incorrecto: tuvo **errores de copiado, de transmisión, o fue intervenido**

Huffman codes: actividad 2.2

Usando el texto “**Instrucciones para subir una escalera**” de Julio Cortázar...

Con base el Notebook de *clase 11* (en Temas), crea un programa de C++/Python que pueda:

1. Abrir el archivo del texto (en .txt) y calcular las probabilidades de cada carácter.
Elimina los acentos para evitar problemas.
2. Crear el árbol de Huffman, con base en las probabilidades
3. Codificar el texto y guárdalo en formato binario (.bin, un archivo de bytes)
4. Abrir y decodificar el archivo binario creado. ¿Deberías guardar el árbol para decodificar?
5. Incluye un menú con opciones para:
 1. Codificar / decodificar archivos
 2. Mostrar una lista de los caracteres, con sus ocurrencias, probabilidades y códigos
 3. Codificar lo que el usuario ingrese
 4. Decodificar lo que el usuario ingrese

Modalidad: individual

Crea tu código en un Colab/Replit o similar y comparte con permisos de edición.

Mas detalles en Canvas.

