Tema 3 · Pràctica: el cas de les lletres boges

Nom i llinatges: _____ **Grup:** _____

Necessites el VS Code i uns 45 minuts. Fes-ho sol o en parella. Llegeix cada pas sencer abans de tocar res, i escriu el que surt a la pantalla, no el que creus que hauria de sortir.

El misteri

Un company t'envia aquesta frase. L'ha copiada d'un fitxer de text:

```
La faÃ§ana de l'Ajuntament Ã©s magnÃfica, i l'Ã via hi passeja cada tarda.
```

Ell jura que l'ha escrita bé. I és veritat: **ningú no s'ha equivocat en escriure-la.** Avui descobriràs qui és el culpable, i com s'arregla sense tornar a escriure res.

> **Recorda (tema 3, apartat 7).** L'ordinador no guarda lletres: guarda nombres. Una **codificació** és la taula que diu quin nombre li toca a cada caràcter. En **ISO-8859-1** (Latin-1) i en **Windows-1252**, cada caràcter ocupa **1 byte**. En **UTF-8**, les lletres d'ASCII ocupen 1 byte i les lletres amb accent, la ç i la ñ, **2 bytes**.

Aposta abans de provar

Contesta ara, sense ordinador. Al final tornaràs aquí a veure si l'has encertada.

a) Quants bytes ocupa la paraula **façana** en UTF-8? I en Latin-1?

---

b) Si un programa llegeix la ç d'UTF-8 com si fos Latin-1, què creus que veuràs: la ç, un altre caràcter o dos caràcters?

Pas 1. Escriu i desa en UTF-8

1. Obre el VS Code i crea un fitxer nou amb **Ctrl + N**.
2. Escriu aquesta frase, exactament igual i **sense prémer Enter al final**: **La façana de l'Ajuntament és magnífica, i l'àvia hi passeja cada tarda per gaudir de l'espectacle.**
3. Mira la **barra d'estat**, la franja de baix de tot. A la dreta hi ha d'haver escrit **UTF-8**.
4. Desa amb **Ctrl + S**. Dins la teva carpeta personal, fes una carpeta nova que es digui **codificacio** i posa al fitxer el nom **text_utf8.txt**.

> **Si a la barra no hi posa UTF-8**, clica on hi posa la codificació, tria **Save with Encoding** i després **UTF-8**. Si tens el VS Code en castellà, l'opció es diu **Guardar con codificación**.

Ara mira quant ocupa. Obre el terminal integrat del VS Code: menú **Terminal › New Terminal** (en castellà, **Terminal › Nuevo terminal**). S'obre a baix. Escriu:

```
cd ~/codificacio
ls -la
```

L'**ls -la** fa una línia per fitxer. La **mida en bytes** és el número que hi ha just abans de la data. Per exemple, en aquesta línia la mida és 57:

```
-rw-r--r-- 1 alumne alumne   57 nov  5 10:12 notes.txt
```

Copia la mida de **text_utf8.txt**:

La frase té **98 caràcters**, comptant els espais i els apòstrofs. Per què el fitxer n'ocupa més? Pista: busca les lletres que no són d'ASCII.

(Si et surt 103 o 104 bytes, has deixat un salt de línia al final. No passa res: tu suma-n'hi 1 o 2 a tots els comptes.)

Pas 2. Posa-li les ulleres equivocades

1. Clica **UTF-8** a la barra d'estat.
2. Tria **Reopen with Encoding** (en castellà, **Volver a abrir con codificación**).
3. A la llista, busca i tria **Western (Windows 1252)**.

Ara el text es veu diferent. Copia què surt en lloc de cada lletra:

Lletra que vas escriure	Què surt ara
ç	Ã§
é	
í	
à	
a, **t**, **l**, l'espai…	

(Comprovació de la primera: la ç es converteix en **Ã§***.)*

> **Compte.** El fitxer **no s'ha espatllat**. No has tocat ni un byte. El que ha canviat és com el VS Code els llegeix: li has posat unes ulleres que no són les seves.

Pas 3. Qui és el culpable?

Exemple resolt 1 — La ç

Quan vas escriure la ç, l'UTF-8 la va desar amb **dos bytes**: **C3** i **A7**.

Ara el VS Code llegeix el fitxer amb Windows-1252. Aquesta codificació no sap que hi ha lletres de dos bytes. Per a ella, **cada byte és una lletra**. Busca el **C3** a la seva taula i hi troba la **Ã**. Busca l'**A7** i hi troba la **§**.

Per això d'una lletra en surten dues.

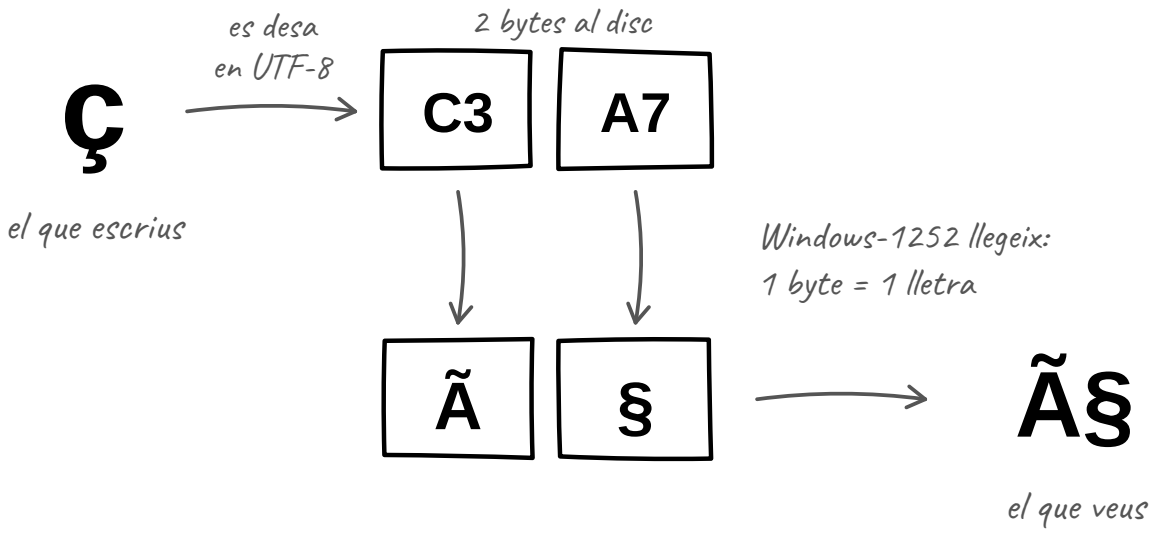

FIG 3.15 Una ç desada en UTF-8 són dos bytes. Llegits amb Windows-1252, són dues lletres.

Comprovació: fes el camí al revés. En Windows-1252, la **Ã** és el byte **C3** i la **§** és l'**A7**. Junts, **C3 A7**, en UTF-8 són la ç. Quadra.

> **Per què totes comencen per Ã?** Perquè en UTF-8, **totes** les lletres de la taula de sota comencen pel mateix byte, el **C3**. I el **C3**, en Windows-1252, és la **Ã**. Si veus moltes **Ã** en un text, ja saps què ha passat.

La teva xuleta

Com desa l'UTF-8 les lletres que fem servir en català i en castellà:

à	ç	è	é	í	ñ	ò	ó	ú	ü
C3 A0	C3 A7	C3 A8	C3 A9	C3 AD	C3 B1	C3 B2	C3 B3	C3 BA	C3 BC

Què mostra Windows-1252 quan troba cadascun d'aquests bytes:

Byte	A0	A7	A8	A9	AD	B1	B2	B3	BA	BC	C3
Mostra	espai	§	¨	©	res	±	²	³	º	¼	Ã

L'**A0** és un espai especial i l'**AD**, un guionet que no es veu. Per això la **à** i la **í** semblen una **Ã** tota sola.

Guiat 1 — Completa la taula

Fes-ho amb les dues xuletes: primer els bytes de l'UTF-8, després què mostra Windows-1252.

Lletra	Bytes en UTF-8	Llegit amb Windows-1252
ç	**C3 A7**	Ã§
ò	**C3**	Ã
è		
ñ		

Practica 1 — Fes de detectiu

Aquests textos s'han llegit amb Windows-1252, però eren UTF-8. Què hi posava de debò?

a) **cafÃ¨** = ...

b) **aÃ§Ã²** = ...

c) **canÃ§Ã³** = ...

d) **CatalÃ** (compte: el que sembla un espai al final no ho és) = ...

(Comprovació del primer: **Ã¨** *és* **C3 A8***, que en UTF-8 és la è. Hi posava* **cafè***.)*

Practica 2 — Ara endevina tu el desastre

Aquestes paraules estan desades en UTF-8. Què es veuria si les obrissis amb Windows-1252?

a) **però** = ...

b) **Lleó** = ...

c) **Hola** = ...

d) **pingüí** = ...

(Comprovació del primer: la ò són els bytes **C3 B2***, que es veuen com* **Ã²***. Surt* **perÃ²***.)*

Pas 4. Al revés: desa en Latin-1 i obre en UTF-8

1. Torna a obrir **text_utf8.txt** amb **UTF-8** (barra d'estat, **Reopen with Encoding**, **UTF-8**). Comprova que la frase torna a sortir bé.
2. Fes **Ctrl + Maj + S** (Desa com) i posa-li el nom **text_latin1.txt**.
3. Clica la codificació a la barra d'estat, tria **Save with Encoding** i després **Western (ISO 8859-1)**.

Torna al terminal i fes **ls -la** una altra vegada. Ara hi ha dos fitxers. Mida de **text_latin1.txt**:

Per què ocupa 4 bytes menys que **text_utf8.txt**?

Ara posa-li les ulleres equivocades: barra d'estat, **Reopen with Encoding**, **UTF-8**. Què surt on hi havia les lletres especials?

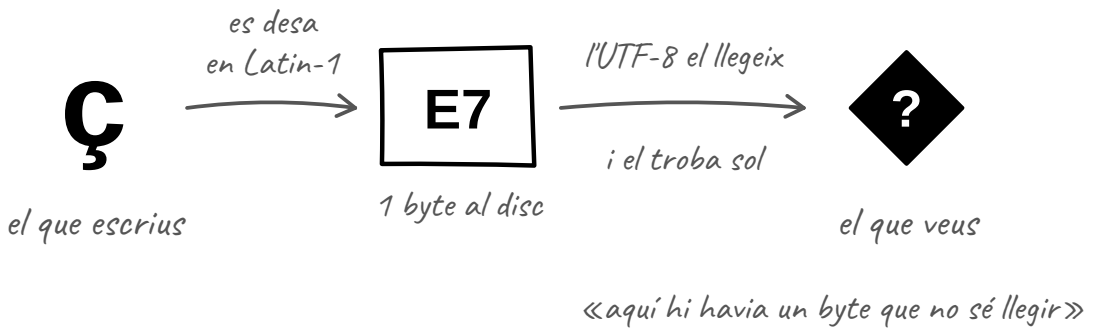

FIG 3.16 En Latin-1 la ç és un sol byte, l'**E7**. L'UTF-8 no l'entén tot sol i hi posa el rombe.

Aquest rombe és el **caràcter de substitució**. Vol dir: «aquí hi havia un byte que no sé llegir». En UTF-8, un byte més gran que 127 sempre ha d'anar en grup amb un altre. L'**E7** hi va sol, i per això l'UTF-8 no hi troba cap lletra.

No hi ha preguntes tontes

Si el fitxer no s'espatlla, per què es veu malament? Perquè un fitxer de text **no diu** en quina codificació està. Només hi ha bytes. El programa ha d'endevinar com llegir-los, i a vegades s'equivoca.

Per què la a o la t sempre surten bé? Perquè són d'ASCII. Valen el mateix i ocupen 1 byte en ASCII, en Latin-1, en Windows-1252 i en UTF-8. Només fallen les lletres de fora de l'ASCII.

Quina codificació he de fer servir? UTF-8, gairebé sempre. És la de les pàgines web, la de Linux i la que fa servir el VS Code si no li dius res. Latin-1 i Windows-1252 només te les trobaràs en fitxers antics.

Pas 5. Resol el cas

a) El company del principi t'envia el fitxer. Com l'arregles **sense tornar a escriure res**?

b) Explica amb les teves paraules per què els caràcters no es veuen bé quan canvies la codificació. Fes servir les paraules **byte**, **UTF-8** i **Windows-1252**.

c) Torna a l'aposta del principi. L'havies encertada? Què t'ha sorprès?

Pas 6. I els emojis?

Obre **text_utf8.txt** amb **UTF-8** i afegeix un emoji al final de la frase. Com es posa un emoji al Debian:

1. Obre l'aplicació **Caràcters** (a **Activitats**, escriu *caràcters*).
2. Busca *cara* i clica l'emoji que vulguis: es copia.
3. Torna al VS Code i enganxa'l amb **Ctrl + V**.

Si no tens l'aplicació, copia'n un d'aquí mateix, d'aquest PDF. Selecciona'n **només un**:



I si al VS Code es veu un quadradet en lloc del dibuix, no passa res: és que falta la tipografia, però els bytes hi són igual.

Desa i fes **ls -la** al terminal. Quants bytes ha crescut **text_utf8.txt**?

Ara obre'l amb **Western (Windows 1252)**. Quants caràcters estranys surten en lloc de l'emoji? Per què?

El que has de saber sí o sí

- Un fitxer de text només guarda **bytes**. La codificació és la taula per llegir-los.
- Si l'obres amb una codificació que no és la seva, **el fitxer no s'espatlla**: només es veu malament.
- **Moltes Ã** en un text: era UTF-8 i s'ha llegit com a Latin-1 o Windows-1252.
- **Rombes amb interrogant**: era Latin-1 i s'ha llegit com a UTF-8.
- S'arregla tornant-lo a obrir amb la codificació bona: **Reopen with Encoding**.
- Davant del dubte, **UTF-8**.