Treball de CDI-GEI

Curs 2022/2023

6 de març de 2023

El treball consisteix en dissenyar i implementar en Python un compressor sense pèrdua pensat per a fitxers que contenen text, i el descompressor corresponent. Entre altres, es poden fer servir *codis de Huffman* o *codificació aritmètica* (millor en versió adaptativa), *mètodes de diccionari*, *Burrows-Wheeler*, *run-length*, *prediction by partial matching*, etc. En aquesta *pàgina web* podeu trobar una llista d'algorismes i tècniques de compressió; tots ells estan explicats detalladament en els llibres de la bibliografia de l'assignatura.

Us pot ser útil consultar la *pàgina* on es comparen mètodes i es fa una classificació en funció de la compressió assolida en un mateix fitxer. Vegeu també el *premi Hutter*.

La pràctica es farà en grups de (fins a) 4 estudiants. Podeu organitzar-vos entre vosaltres durant les dues primeres classes de laboratori i dir-me la composició dels grups.

Requisits:

- Es pot usar *qualsevol mètode o combinació de mètodes*.
- El compressor ha de funcionar correctament *sobre qualsevol fitxer* que contingui caràcters Unicode en format UTF-8. En descomprimir un fitxer prèviament comprimit el resultat ha de ser *exactament* el fitxer de partida.
- El rendiment del compressor/descompressor es mesurarà sobre els fitxers de textos que teniu a Atenea. La unitat de mesura seran els bits per símbol Unicode comprimit, segons la fórmula següent:

$$8 \times \frac{\text{nombre de bytes del fitxer comprimit}}{\text{nombre de caràcters Unicode del fitxer original}}.$$

  Observeu que el denominador no sempre és el nombre de bytes del fitxer a comprimir: hi ha caràcters Unicode que s'escriuen usant més d'un byte.
- El temps d'execució per a la compressió o la descompressió de qualsevol dels fitxers de textos d'Atenea ha de ser inferior a cinc minuts en un dels ordinadors del laboratori on es fan les classes.

Entrega:

- Tres fitxers amb **exactament** els noms que es donen a continuació:
  - Dos scripts de Python amb els noms `compressor.py` i `descompressor.py` que contenen el codi Python del compressor i descompressor. Han de funcionar en els ordinadors de la classe de laboratori, de manera que les comandes següents facin exactament el que s'indica a continuació:
    - `python compressor.py nomfitxer`
      llegeix un fitxer de nom `nomfitxer.txt`, comprimeix el seu contingut, i crea un fitxer comprimit amb el nom `nomfitxer.cdi`; a més, imprimeix per pantalla dues línies on es dona la compressió i el temps:
      `Bits per símbol =` el valor en bits/símbol
      `Temps compressió =` el temps en segons.
    - `python descompressor.py nomfitxer`
      llegeix el fitxer de nom `nomfitxer.cdi`, el descomprimeix i escriu el resultat en un fitxer de nom `nomfitxer-desc.txt`; a més imprimeix per pantalla:
      `Temps descompressió =` el temps.
  - Un fitxer PDF amb el nom `readme.pdf` on hi hagi una descripció molt breu del mètode o mètodes usats, la bibliografia o webgrafia consultada i si es fa servir algun package de Python no estàndard i per a què. També es poden explicar les dificultats trobades i com s'han solucionat.
    A més, el fitxer ha de contenir una taula on hi hagi les dades per als fitxers de text d'Atenea: per a cadascun la compressió assolida (en bits per símbol) i els temps de compressió i de descompressió.
    Mida màxima del fitxer: 2 pàgines.
- Data límit per a l'entrega: **dilluns 22 de maig a les 23:59 hores.** Es pot canviar o millorar la versió presentada tantes vegades com es vulgui fins a aquesta data.

Avaluació:

- Es valorarà sobretot la capacitat de comprimir les dades mesurada en bits per símbol sobre els fitxers de text d'Atenea però també es podrà tenir en compte el temps que tarden el compressor i el descompressor, la dificultat del mètode usat, etc.
- Per aprovar la pràctica la implementació ha de funcionar correctament d'acord amb els requisits i la compressió ha de ser inferior a 1.25 vegades la mitjana de les compressions assolides per tots els treballs presentats.
- Almenys els tres treballs amb millor compressió tindran un excel·lent.
- Almenys la meitat dels altres treballs aprovats tindran un notable.
- Durant les dues darreres classes de laboratori (25 de maig i 1 de juny) es podrà demanar als autors explicar alguns dels treballs; els autors han de saber explicar el(s) mètode(s) utilitzat(s) i explicar i justificar la seva implementació.