ANÀLISI DELS COMPRESSORS BZ I GZ

Probabilitat i Estadística

PROJECTE BLOC 7
FACULTAT D'INFORMÀTICA DE BARCELONA - UPC '16

1. RESUM

1.1. Objectius

Comparar dos formats de compressió Gzip i Bzip2 per tal d'esbrinar quin dels dos es comporta millor a l'hora de comprimir projectes relacionats amb la informàtica i, com a objectiu secundari analitzar el comportament de l'arxivador d'arxius Tar.

1.2. Mètode

S'ha creat un script amb el llenguatge python per tal d'obtenir projectes de manera aleatòria d'entre els repositoris públics de GitHub. La mostra final consta de cinc cents projectes.

1.3. Resultats

Als resultat s'ha vist que la mitjana de la ràtio %Gzip - %Bzip2 és negativa i que els dos extrems de seu interval de confiança també son negatius.
Altrament el format Tar, de mitjana, quasi duplica la mida original de l'arxiu, però per a arxius petits aquesta mida és pot arribar a multiplicar per set.

Lo normal es incluir el resultado numérico, "con números"

Esta afirmación se podría objetar mucho, ya que el ratio depende fuertemente del tamaño inicial

1.4. Discussió

El format Bzip2 es capaç de generar arxius més petits de mitjana, i per tant és millor en terme d'espai, que era l'objectiu a analitzar. Per altre banda, el format Tar pot arribar a multiplicar per set la mida d'un arxiu petit, però com més gran és aquest arxiu més disminueix la ràtio, tot i que en valors absoluts la diferencia entre la mida inicial i la del Tar augmenta amb arxius més grans.

Demasiada importancia a un resultado secundario pobremente analizado.

2. INTRODUCCIÓ

Com a estudiants d'enginyeria informàtica, sovint ens veiem en la necessitat de treballar amb projectes que ocupen molt espai. Això en dificulta la distribució i emmagatzematge. Per això és molt útil l'ús de software destinat a la compressió d'aquest conjunt d'arxius. Com en realitzem un ús freqüent hem destinat aquest treball a comparar dos d'aquests mètodes de compressió.

En concret el Bzip2 i Gzip. Ho hem fet segons la seva ràtio de compressió, obviant el temps que es triga en generar l'arxiu. D'aquesta manera podrem prendre una millor elecció alhora de triar un d'aquests dos compressors. Paral·lelament hem analitzat el comportament del format Tar alhora d'agrupar arxius.

También se podría justificar por qué estos dos, si hay otras opciones (o por qué se han desechado esas otras)

3. OBJECTIU

Determinar quin dels dos compressors, Gzip i Bzip2 és millor, entenent com a millor aquell que aconsegueix reduir més l'arxiu d'entrada, sense tenir en compte el temps necessari per a generar-lo.

Com a objectiu secundari es vol determinar el comportament del format Tar per a diferents tipus d'arxius.

5. MATERIAL I MÈTODES

Per a dur a terme l'estudi s'ha creat un script amb el llenguatge de programació Python que utilitza el generador de nombres aleatoris **random.randint()**, amb un rang entre 1 i el total de repositoris públics de GitHub per tal d'obtenir l'índex d'un projecte. Aquest índex és únic i s'utilitza per a fer la següent petició al servidor:

https://api.github.com/repositories

El mateix script que descarrega el projecte del servidor, el comprimeix utilitzant els formats de compressió Gzip, Bzip2 i l'arxivador Tar. També n'obté les mides en KB i genera un document Excel on hi escriu totes les dades. Aquest procés s'ha executat cinc-centes vegades per tal d'obtenir la mostra inal.

Los apartados 6, 7, 8 y 9 se pueden considerar subapartados de M & M.

6. VARIABLES

Per trobar un indicador que ens serveixi per a comparar quin dels dos compressors es millor, s'ha obtingut el percentatge de compressió en els dos formats i s'ha realitzat la diferencia entre ells (%Gzip - %Bzip2). De tal manera que obtenir un resultat positiu indica que per aquell projecte, la compressió mitjançant el Gzip ha estat millor. Per altre banda, un resultat negatiu indica una millor compressió per part del format Bzip2.

Per a l'anàlisi del objectiu secundari, els nostres indicadors són la ràtio d'augment respecte l'arxiu inicial un cop arxivat amb el format Tar, i la seva diferència absoluta.

7. PROVA DE SIGNIFICACIÓ

Mejor: "modelo estadístico", o similar.

És unilateral i volem comprovar que la ràtio de compressió d'un dels compressors és millor que el de l'altre. On Ψ és la diferencia entre les ràtios de compressió dels dos formats (%Gzip - %Bzip2).

No hay razones para ello

$$H0: \Psi = 0$$
$$H1: \Psi < 0$$

Per a la resta de l'estudi s'utilitzarà la lletra grega Ψ per referir-se a la variable %Gzip - %Bzip2.

Hum, no suelen usarse variables griegas para las variables observadas, pero contra gustos ... Por cierto, más abajo aparece como "GB"

8. PREMISES

La prova és unilateral. No es una premisa como tal

"las observaciones dentro de cada muestra son indep."

Les mostres són independents.

Les mostres son aparellades.

9. ESTADÍSTIC I DISTRIBUCIÓ

μ	$H_0 : \mu_1 = \mu_2$	$\hat{t} = \frac{\overline{D} - \mu_0}{S_D / \sqrt{n}}$	D ~ N m.a. aparellada	$\hat{t} \sim t_{n-1}$	Rebutjar si $\lvert \hat{t} \rvert > t_{n-1,\, 1-\alpha/2}$

Fig. 1 : Hipòtesis, premisses i distribució de l'estudi.

las figuras se han de referenciar (y explicar)

10. RESULTATS

Mitjana de %Gzip:	0.3636684
Desviació Estàndard de %Gzip:	0.1859363
Mitjana de %Bzip2:	0.3752131
Desviació Estàndard %Bzip2:	0.1906292

Si son porcentajes, deben ser: 36.36, 18.59, 37.52, etc. Pero nos interesan los indicadores de la diferencia (relativa).

S'han obtingut els resultats anteriors mitjançant el programa R. Es pot observar que tot i que les mitjanes són molt similars, la del Bzip2 és lleugerament superior a la del Gzip fet que indica que és millor compressor. Tot i així també presenta una desviació més gran.

Per veure com es reparteix la diferència entre ambdós percentatges respecte a la mida d'entrada dels arxius s'ha creat el següent plot on veiem que per a arxius petits el compressor Bzip2 clarament comprimeix més. Amb arxius d'altres mides, tot hi haver-hi més arxius on el compressor Gzip es comporta lleugerament millor, aquells en els que el Bzip2 comprimeix millor la diferència sol ser molt més marcada. A part d'això podem veure alguns "outliers" on el compressor Bzip2 es comporta millor del normal respecte al Gzip.

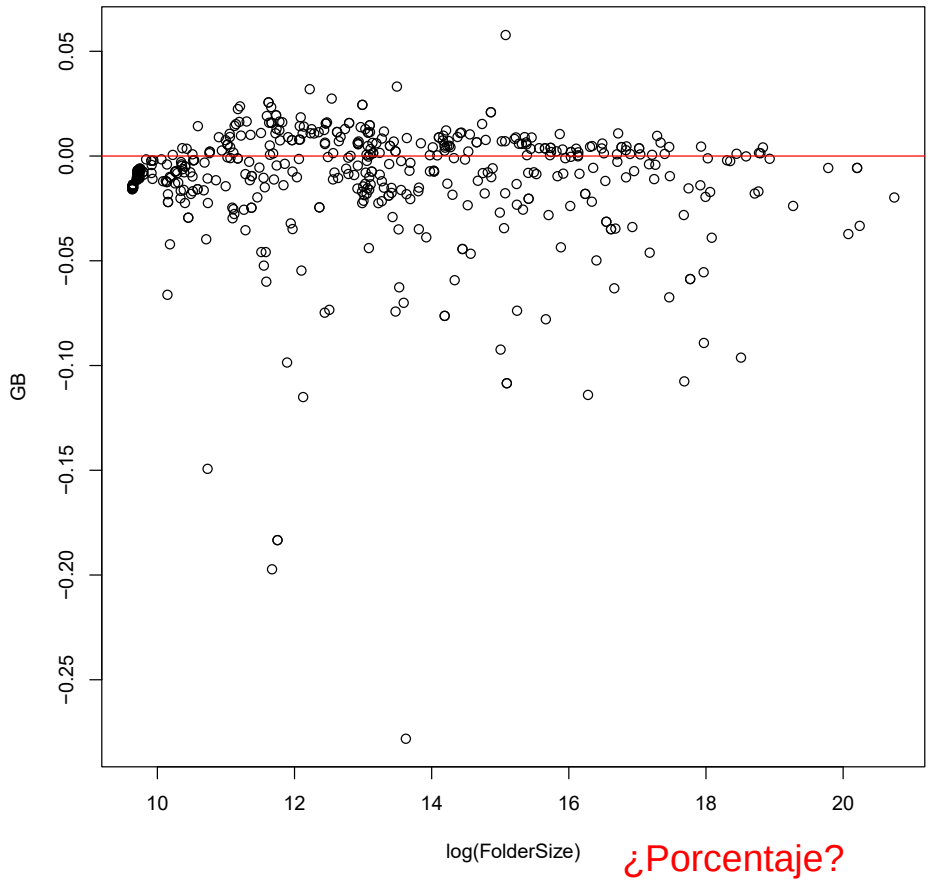

Fig. 2 : Plot mostrant la diferència entre el percentatge de compressió del Gzip i el del Bzip2

Mitjana de Ψ -0.0115447

OK

Desviació Estàndard de Ψ 0.03089752

¿? Creo que la media y la sd no tienen esa capacidad

Les conclusions anteriors també s'observen quan s'obté la mitjana i la desviació de la variable a estudiar.
Per tal d'observar la variable Ψ de de forma gràfica n'hem obtingut el seu histograma, on hi podem veure que, tot i que es segueix veient una alta semblança, podem començar a intuir que la variable es decanta cap a valors negatius cosa que significa una millor compressió per part del Bzip2.

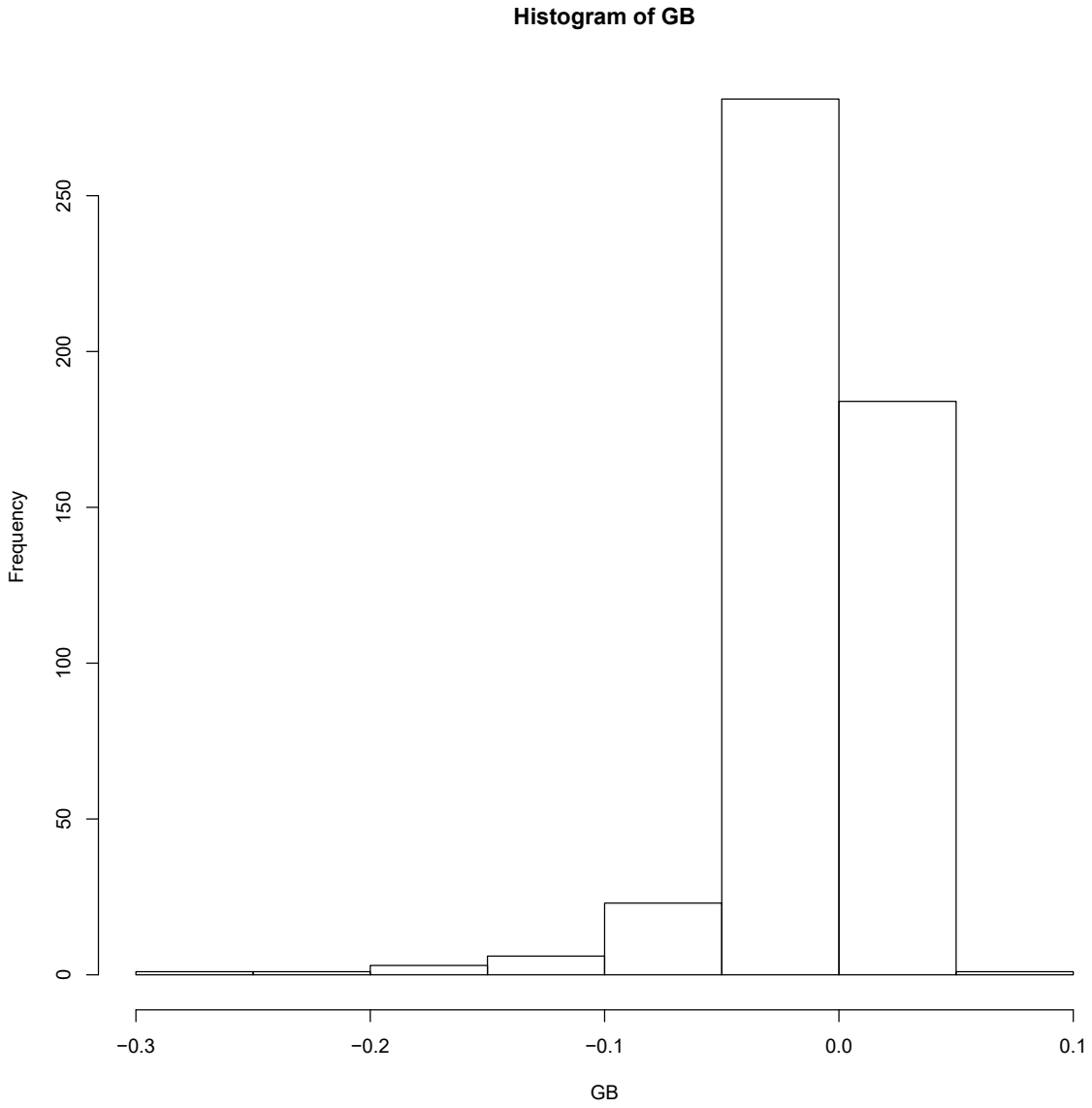

Fig. 3 : Histograma de la variable Ψ.

A continuació és comprova la premissa de normalitat de la variable mitjançant la funció QQNorm de R i es pot afirmar que la variable compleix la normalitat tot i la presencia d"'outliers".

Claramente, no. Sin embargo, a partir de -1 cuadra bastante bien.

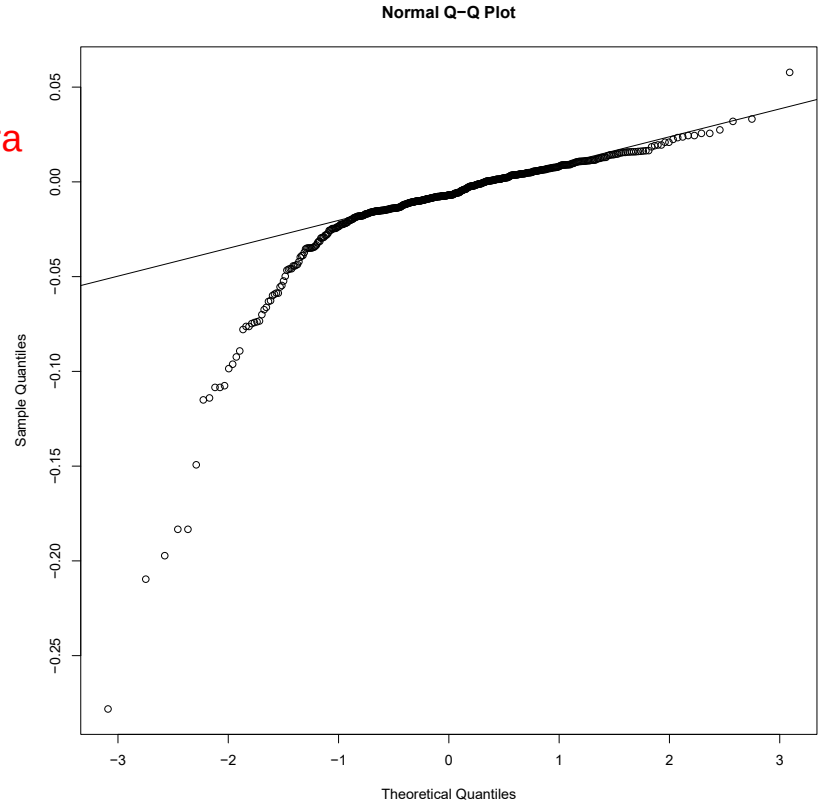

Fig. 4 : QQPlot de la variable Ψ.

Finalment, s'ha dut a terme la prova d'hipòtesi mitjançant la comanda t.test de R obtenint els següents resultats:

data: GB

t = -8.355, df = 499, p-value = 6.53e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

-0.01425952 -0.00882988

sample estimates:

mean of x

-0.0115447

S'ha vist que el P-valor obtingut és més petit que el valor de risc del 5%, i per tant és pot rebutjar la hipòtesi nul·la i es confirmar la hipòtesi alternativa.

Quan s'observa l'interval de confiança del 95% es pot veure que ambdós límits estan en valors negatius i significativament allunyats del origen, cosa que indica un millor comportament per part del compressor Bzip2.

11. DISCUSSIÓ

La conclusió del estudi realitzat mostra que el compressor Bzip2 és capaç de generar arxius més reduïts que el compressor Gzip. Concretament el percentatge de compressió del Bzip2 és un 1.15447% millor que el percentatge de compressió del Gzip. Aquesta millora només contempla la mida resultant del arxiu sense tenir en compte el temps de compressió (on veuríem que el Gzip és més ràpid).

IC: 0.9 - 1.4%

Analitzant els continguts dels "outliers" de la figura 2, i una mostra aleatòria d'entre la resta, hem observat que en els casos "outliers" una majoria de la mida total està formada per fotografies amb format Png, i a la resta de repositoris de la mostra només hi trobem arxius de codi. Això ens portaria a una conclusió prematura de que el compressor Bzip2 funciona significativament millor que el Gzip alhora de comprimir fotos en aquest format. Tot i així faria falta un estudi en més detall per tal d'arribar a una conclusió més sòlida.

A la pràctica si escollíssim un projecte aleatori de GitHub, al comprimir-lo amb el format Bzip2 podríem esperar en un 95% del casos que aquest arxiu sigui comprimit entre un 35.84% i un 39.20%. En canvi si aquest mateix arxiu el comprimíssim amb Gzip podríem esperar una compressió entre un 34.77% i un 38.04%.

Esto es bastante interesante, aunque se podría documentar mejor

¿Limitaciones?

ANÀLISI SECUNDARI

En la realització de l'estudi ens hem adonat que el format Tar augmenta la mida dels projectes que arxiva. Partint d'aquesta observació, s'ha analitzat, com a objectiu secundari el comportament d'aquest format. Per fer-ho s'ha calculat la ràtio de creixement del arxiu original respecte al resultat per a cada una de les cinc-centes mostres i la diferencia absoluta:

Ràtio = Mida Resultant / Mida Original

Diferència = Mida Resultant - Mida Original

Observant la ràtio veiem que en arxius de mides petites poden arribar a augmentar la seva mida fins a set vegades, en canvi com més grans són aquests arxius més petit es aquest augment. Com podem veure en la figura 5.

En canvi observant la diferència entre veiem que aquesta és més gran com més gran són els arxius, tal com s'aprecia en la figura 6.

analizar estos datos con un modelo lineal hubiera sido muy positivo. Lástima.

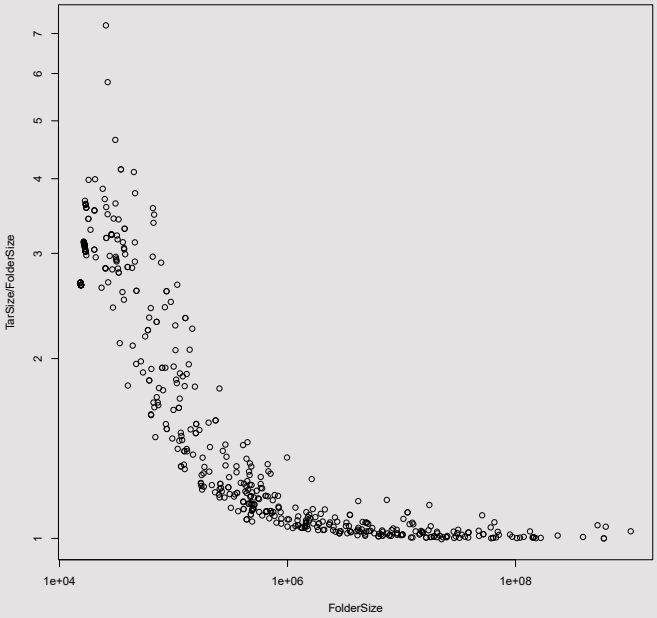

Fig. 5 : Plot de la ràtio respecte el logaritme de la mida original del arxiu

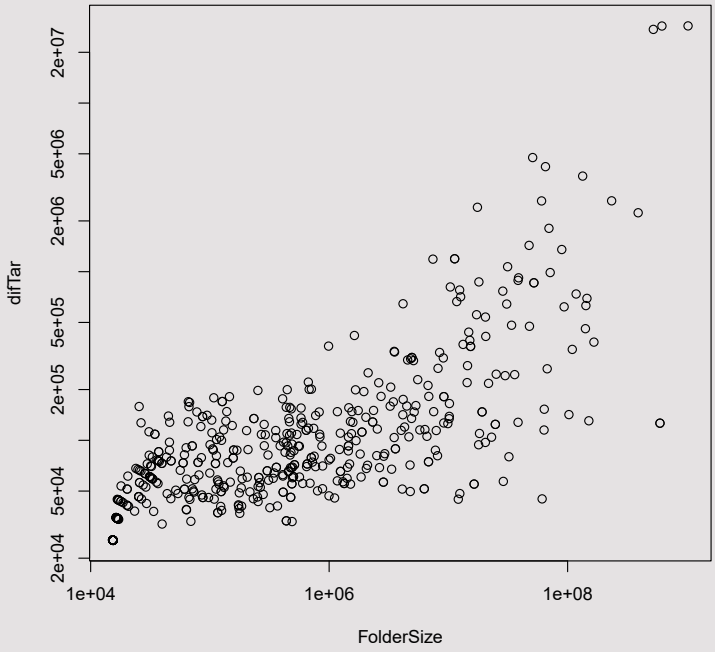

Fig. 6 : Plot de l'augment de mida respecte el logaritme de la mida original de l'arxiu.

Per ratificar aquesta conclusió s'ha calculat la mitjana i la desviació d'aquesta ràtio:

Mitjana de la Ràtio: 1.739875

Desviació de la Ràtio: 0.9244396

PROJECTE B7 PROBABILITAT I ESTADÍSTICA
FACULTAT D'INFORMÀTICA DE BARCELONA - UPC '16

JOSEP SANS GERARD SERRAMIÀ CARLES ROJAS