Infrormacija in kodi

Poročilo prve laboratorijske vaje

Avtor: Bor Starčič

Mentorja: izr. prof. dr. Simon Dobrišek, asist. dr. Klemen Grm

Datum: April 1, 2024

Contents

1 Uvod

Entropija igra ključno vlogo pri merjenju stopnje naključnosti ali nepredvidljivosti podatkov. V tem projektu smo se osredotočili na določanje entropije različnih vrst datotek, vključno s tekstovnimi datotekami, slikami in zvočnimi posnetki, pri čemer smo upoštevali različne ravni abstrakcije: od enojnih znakov do kompleksnejših zaporedij, kot so pari, trojice, četverice in petice znakov.

2 Teorija

Entropija H za diskretno naključno spremenljivko je definirana kot:

$$H = -\sum_{i} p(x_i) \log_2 p(x_i)$$

kjer $p(x_i)$ predstavlja verjetnost pojavljanja i-tega znaka. V kontekstu naše analize, kjer se osredotočamo na bite, je osnova logaritma 2, kar odraža dvojiški sistem kodiranja informacij. Entropija je izračunana za zaporedja znakov različnih dolžin (od 1 do 5), kar nam omogoča podrobnejši vpogled v strukturo in kompleksnost podatkov.

3 Metodologija

Za izračun entropije smo uporabili Python skripto, ki prebere datoteko kot zaporedje bajtov, s čimer vsak bajt predstavlja en znak z 256 možnimi vrednostmi. To nam omogoča, da obravnavamo datoteko kot niz 8-bitno kodiranih znakov in izračunamo entropijo za različne vrednosti n.

```
with open(file_path, 'rb') as file:
        b = np.frombuffer(file.read(), dtype=np.uint8)
```

Ta del kode je ključen za pretvorbo vsebine datoteke v obliko, ki je primerna za analizo entropije, saj nam omogoča, da vsak bajt obravnavamo kot diskretno naključno spremenljivko.

4 Rezultati in ugotovitve

Rezultati kažejo, da se z večanjem vrednosti n entropija datotek zmanjšuje. To je pričakovano, saj daljša zaporedja znakov običajno vsebujejo več informacij, hkrati pa se verjetnost njihovega ponovnega pojavljanja zmanjšuje.

Poleg tega smo opazili, da kompresija datotek poveča njihovo entropijo, kar kaže na učinkovitost teh postopkov pri zmanjševanju redundance in povečevanju nepredvidljivosti podatkov. Na splošno kompresirane datoteke kažejo višjo stopnjo naključnosti v primerjavi z originalnimi datotekami. Šifriranje datoteke besedilo.zip ni spremenilo njene entropije, kar predvidevam je zaradi že močne kompresije datoteke besedilo.zip.

Za statistično signifikantno oceno entropije mora biti velikost datoteke dovolj velika, da preseže desetkratnik nabora možnih znakov. To pomeni, da za zanesljiv izračun entropije zaporedij treh bajtov (H3) mora biti datoteka večja od 168 MB. Ker nobena iz preučevanih datotek ni dosegla te velikosti, so izračuni za H3, H4 in H5 manj zanesljivi in ne odražajo natančno entropije datotek.

Table 1: Izračunane entropije za različne datoteke in različne vrednosti n

Datoteka	H1	H2	H3	H4	H5
besedilo.txt	4.5684	3.9941	3.6270	3.3132	3.0330
besedilo.zip	7.9984	7.9387	6.2071	4.6616	3.7294
sifrirano_besedilo.zip	7.9984	7.9387	6.2071	4.6616	3.7294
slika.bmp	7.5885	6.9291	5.9451	4.8110	3.9663
slika.png	7.9871	7.8881	6.4492	4.8682	3.8955
slika.jpg	7.9700	7.6951	5.6730	4.2616	3.4095
posnetek.wav	6.3524	5.9895	5.2759	4.8261	4.2729
posnetek.flac	7.9542	7.9343	7.6126	5.9241	4.7408
posnetek.mp3	7.9580	7.8982	7.0868	5.4049	4.3371

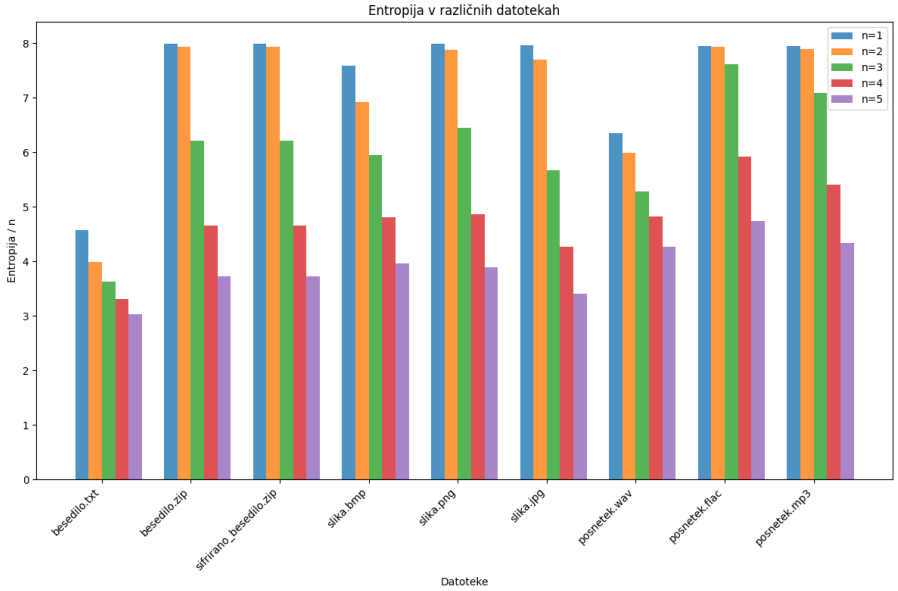

Figure 1: Graf izračunanih entropij

5 Zaključek

Analiza entropije je potrdila, da so metode kompresije učinkovite pri povečevanju naključnosti in zmanjševanju predvidljivosti podatkov. Ugotovitve poudarjajo pomen izbire ustrezne velikosti vrednosti n in velikosti datoteke za natančno in zanesljivo oceno entropije.