KOMPRESIJA TEKSTA

- Tekst je još jedan značajan medium za predstavljanje multimedijalnih informacija.

- Ustvari, to je **najstariji medijum** za skladištenje i prenos informacija. **Najveći broj današnjih informacija jos uvek se čuva u tekstualnom obliku.**

- U ovom poglavlju daćemo najpre kratak uvod u predstavljanje teksta u digitalnoj formi, a zatim će biti reči o nekim tehnikama kompresije teksta.

Predstavljanje teksta u digitalnoj formi

- Za razliku od audio i video podataka, kod teksta nema odmeravanja u vremenskom domenu niti kvantizacije.
- Za podatke o tekstu može se pretpostaviti da su stvoreni **iz diskretnog informacionog izvora** koji emituje simbole, a ovi predstavljaju slova koja odgovaraju nekoj azbuci (alfabetu).
- Pošto se tekstualni podatak široko koristi za predstavljanje mnogih dokumenta (knjige, novine, periodike), važno je sprovesti **efikasno predstavljanje kako bi se smanjio zahtev za skladištenim prostorom (memorijom).**

- Pošto izvor ima svoju azbuku, to znači da on može da emituje bilo koji simbol (poruku) koji pripada datoj azbuci.

- Da bi opisali izvor, možemo uzeti u obzir sve moguće simbole koje izvor može da emituje.
- To znači da će izvor biti mnogo bolje opisan nekom **srednjom količinom informacije** koju on emituje. Ova srednja količina informcije naziva se **entropija izvora**.

- **Entropija** predstavlja prosečnu količinu informacija po simbolu kojom izvor snabdeva posmatrača, odnosno prosečni izvor neizvesnosti koju posmatrač poseduje pre nego sto izvrši ispitivaje izlaza iz izvora.

- Neka diskretni izvor informacija emituje simbole iz azbuke tj. skupa simbola S
- $S = \{s1, \ldots, si, \ldots, sj \ldots, sq\}$
- pri čemu se svaki simbol pojavljuje sa izvesnom verovatnoćom
- $P(s1), \ldots, P(si), \ldots, P(sj) \ldots, P(sq)$.
- Ukoliko pojava datog simbola ne zavisi od prethodnih simbola, entropija ovakvog izvora informacije data je izrazom

$$H(s) = \sum_{S} P(s_i) \log_2 \frac{1}{P(s_i)}, \quad s_i \in S . \tag{6.1.1}$$

Maksimalnu entropiju imaće izvor kod koga su verovatnoće emitovanja svih simbola date azbuke jednako verovatne. Tada je neizvesnost najveća pošto nijedan simbol nema prioritet. Ako azbuka ima q simbola čija je verovatnoća pojavljivanja p = 1/q onda je maksimalna entropija

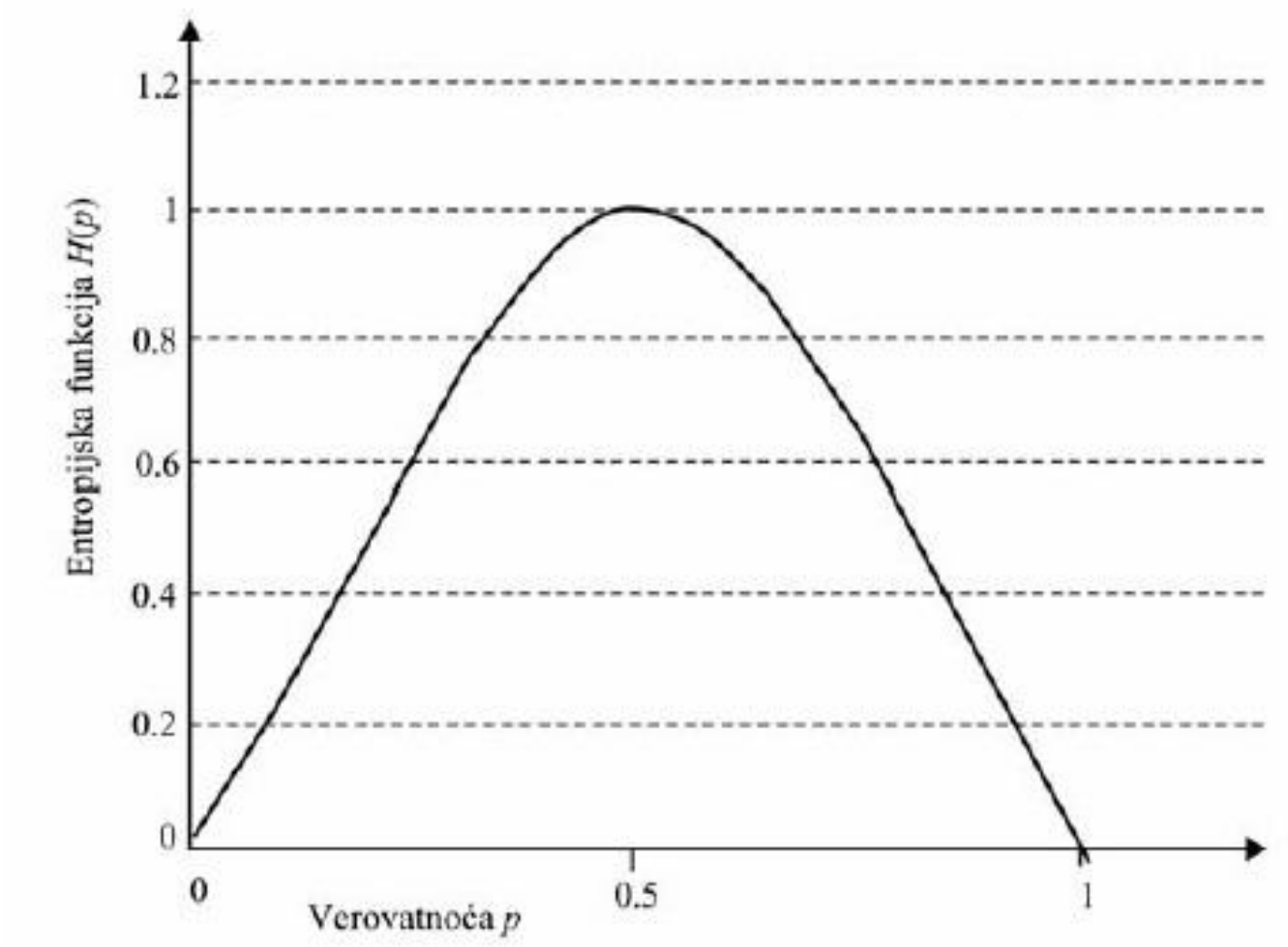
$$H_{\max} = \sum_{i=1}^{\Re} P(s_i) \log_2 \frac{1}{P(s_i)} = \sum_{i=1}^{q} \frac{1}{q} \log_2 (q) = \log_2 (q) \tag{6.1.2}$$

jer je očigledno $\sum_{i=1}^{q} \frac{1}{q} = 1$.

- Kao što se vidi, **maksimalna entropija** brojno je jednaka količni informacija.
- **Minimalna entropija** nastaje kada izvor ne emituje nikakvu informaciju. To je slučaj kada izvor emituje samo jedan simbol, tj. kada je verovatnoća pojavljivanja jednog simbola jednaka jedan, a svih ostalih je nula.
- Na slici 6.1.1 prikazana je entropija binarnog izvora informacija u funkciji verovatnoce pojave 0 ili 1.

- Funkcija H(p) naziva se entropijska funkcija. Izvori koji emituju simbole koji su međusobno statistički nezavisni nazivaju se izvori bez memorije.

- Najzad na osnovu svega do sada naloženog može se izvući zaključak da je **izvor informacija definisan: azbukom (listom simbola) koju emituje, verovatnoćom pojavljivanja simbola, brzinom emitovanja simbola I prosečnom brzinom emitovaja informacije.**



Slika 6.1.1—Entropija binarnog izvora informacija

Principi kompresije teksta, statistička suvišnost

- Tipični podaci o tekstu nose suvišnu informaciju. **Tehnike kompresije teksta** smanjuju ili eliminišu tu suvišnost čime se postiže kompresija.

- Neka je izvor informacija pisani jezik. On se kao i govorni može posmatrati kao stohastički proces u kome su nizovi slovnih znakova determinisani zakonima verovatnoće.

- Međutim opisivanje jezika kao stohastičkog procesa je otežano zbog toga sto su različiti elementi jezika (slova, grupe slova i reči u jeziku) međusobno zavisni.

- To znači da upravo emitovanje simbola zavisi od niza prethodno emitovanih simbola.
- Ukoliko verovatnoća pojavljivanja datog slova zavisi od “m” prethodnih simbola, tj ako je uslovna verovatnoća da će posle sj1, sj2, …, sjm biti emitovan simbol si data izrazom
- P (si |sj1, sj2, …, sjm)
- gde je
- i = 1, 2, … , q, jp = 1, 2, … , q,
- kažemo da se radi o **Markovljenim izvorima** m-tog reda koji generišu simbole sa liste
- s = (s1, s2,… sq), pri čemu su pored već opisanih uslovnih verovatnoća poznate i verovatnoće
- P(si), dok se entropija ovog izvora određuje pomoću izraza

$$H(S) = \sum_{i=1}^{q^{m+1}} P\left(s_{j1}, \ldots, s_{jm}, s_i\right) \times \log_2 \frac{1}{P\left(s_i \mid s_{j1}, s_{j2}, \ldots, s_{jm}\right)}.$$

- Niz od prethodno emitovanih “m” simbola predstavlja stanje izvora. Ukupan broj stanja iznosi qm, pa je broj uslovnih verovatnoća qm+1. **Postojanje statističke zavisnosti između simbola smanjuje neizvesnost pa, prema tome, i količinu informacije.**

- **To znači da će entropija izvora koji emituje statistički zavisne impulse uvek biti manja od entropije istog takvog izvora, ali koji emituje statistički nezavisne impulse.**

- Označimo sa Hp vrednost entropije neke azbuke pod pretpostavkom da je verovatnoća pojavljivanja svih slova (simbola) date azbuke jednaka i da ne postoji nikakva statistička zavisnost pri pojavi slova (**H0 = Hmax**).
- Neka je **H∞ stvarna vrednost** entropije u slučaju da su uzete u obzir sve statističke zavisnosti jezika koje se izrazavaju pomoću date azbuke. Tada **odnos H∞/H0 predstavlja stepen iskoriscenja simbola (slova) ili relativnu entropiju.**

Statistička suvišnost odnosno redundantnost predstavlja meru za ograničenja koju tekstu nameće statistička struktura jezika. Definiše se preko stepena iskorišćenja slova, tj.,

$$R = 1 - \frac{H_\infty}{H_0} \qquad (6.2.2)$$

ili u procentima

$$R\,(\%) = 100\,(H_0 - H_\infty)/H_0. \qquad (6.2.3)$$

Shannon-ova teorema kodovanja za bešumni izvor

- Funkcija gustine verovatnoće i entropija izvora informacija mogu se primeniti kod komprimovanja teksta. U tom cilju pomoći će nam Shannon-ova teorema kodovanja za bešumni izvor.

- Neka je **S izvorni tekst** sa **alfabetom veličine K** i **entropijom H(S)**. Posmatrajmo blokove kodovanja od N izvornih simbola u binarne kodne reči. Shannon-ova teorema kodovanja za bešumni izvor kaže da je za bilo koje $S > 0$, moguće za N dovoljno veliko konstruisati kod tako da **prosečan broj bita po originalnom izvornom simbolu R** zadovoljava sledeću relaciju:

- $$H(S) \leq R \leq H(S) + S$$

- **Drugim rečima izvor može biti kodovan bez gubitaka sa prosečnim brojem bita bliskim njihovoj entropiji ali ne manjoj od entropije.** Entropija izvora ograničena je sa 0 i log2K. To znači, drugim rećima da je
- $0 \leq H(S) \leq \log_2 K.$
- U ovoj relaciji leva strana važi ako je p[k] nula za sve izvorne simbole sk izuzev jedinice. U tom slučaju je izvor apsolutno prediktabilan. Desna strana važi kada svaki izvorni simbol sk ima istu verovatnoću pojavljivanja. Redundansa izvora je definisano kao
- $\text{REDUNDANSA} = \log_2 K - H(S).$
- **Ova relacija pokazuje da ukoliko izvor ima alfabet velicine K, maksimalna entropija izvora je log2K. Ako je entropija iznosi log2K za izvor se može kazati da je nulte suvišnosti. U najvećem broju slučajeva informacija sadrži suvišnost.**

Huffman-ovo kodovanje

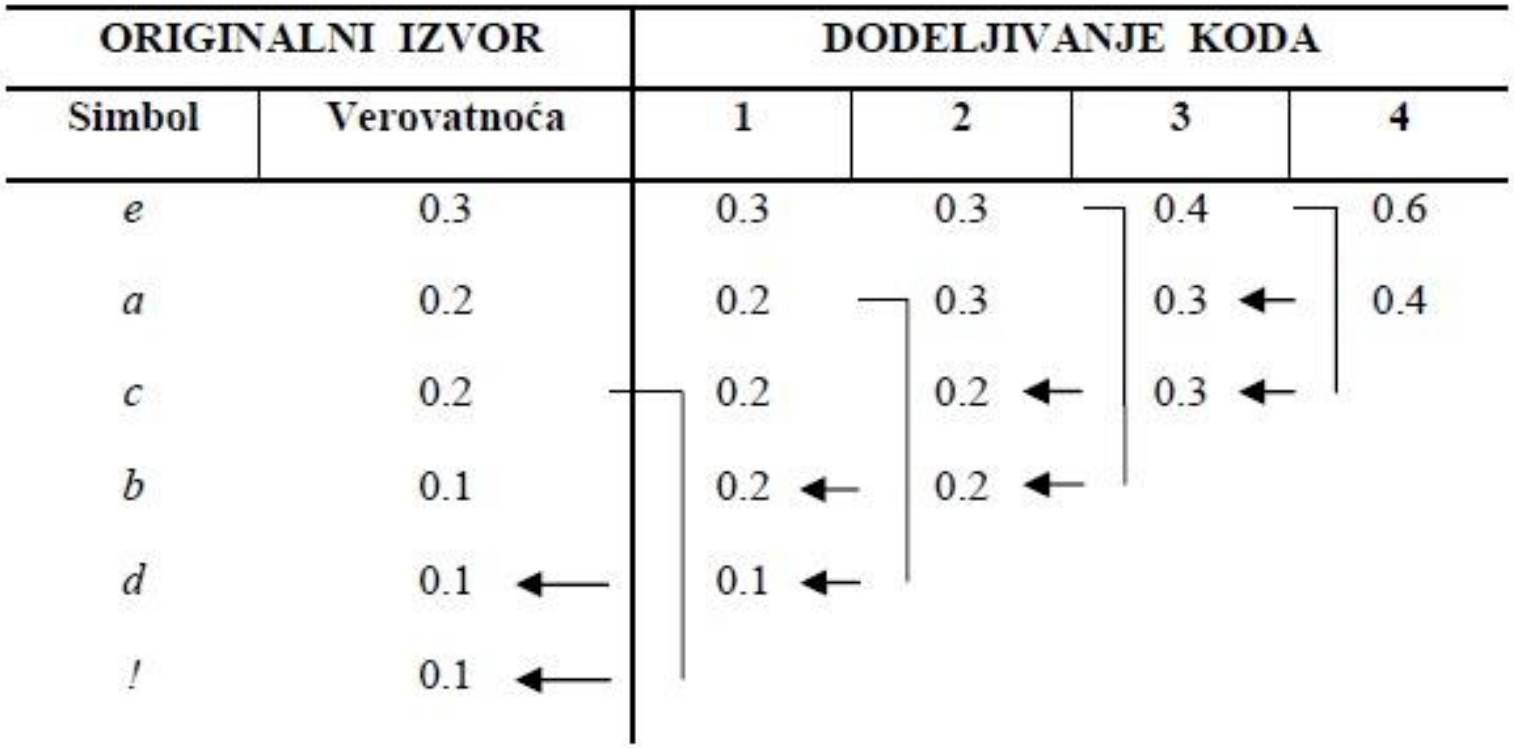
- Prema Shannon-ovoj teoremi o bešumnom kodovanju, prosečan broj bita R za kodovanje izvorne informacije je ograničen sa donje strane entropijom izvora. Međutim Shannon-ova teorema ne utvrđuje način projektovanja kodera koji će korigovati izvornu informaciju sa prosečnim brojem bita R . **Huffman je dao praktičnu metodu za projektovanje kodera koji daje broj bita blizak entropiji.** Ovaj metod daje kod promenljive dužine (eng. variable length code – VLC) za svaki izvorni simbol i to tako da je **broj bita u kodu približno obrnuto proporcionalan verovatnoći nastajanja simbola.**

- Tabela 6.4.1 prikazuje alfabet izvora zajedno sa verovatnoćama individualnih simbola. Izvor informacija sa 6 simbola zahteva najviše log2 6 = 2.58 bita. Entropija izvora je u ovom slučaju H = − (0.3 · log2 0.3 + 2 · 0.2 log2 0.2 + 3 · 0.1 log2 0.1) = 2.44 bit/simbol. To drugim rećima znači da je moguće projektovati koder koji će obavljati svoju funkciju sa 2.44 bit/simbol u proseku.

Tabela 6.4.1—Primena modela za alfabet {*a*, *b*, *c*, *d*, *e*, *!*}

Simboli	Verovatnoća	Huffman-ov kod
a	0.2	10
b	0.1	011
c	0.2	11
d	0.2	0100
e	0.3	00
!	0.1	0101

- Huffman-ov kod za izvor informacije je generisan u dva koraka: **redukcija izvora i naznaka koda.** Proces redukcije izvora za informacioni izvor koji odgovara Tabeli 6.4.1. objasnićemo korišćenjem Slike 6.4.1.
- Postoji nekoliko koraka. **U prvom** koraku simboli su poređani po opadajućoj verovatnoći nestajanja.
- Na primer ovde je p(b) = p(d) = p(!) = 0.1. Izabrali smo redosled “b, d, !”, ali je mogućan bilo koji redosled..
- Kada se simboli razvrstavaju, u **sledećem koraku** novi simbol je kreiran kombinacijom dva najmanja verovatna simbola. U našem primeru stavili smo “d” i “!”. Stoga, novi simbol ima verovatnoću nastajenja 0.2 (0.1 i 0.1). Sada vidimo da se broj simbola smanjio za jedan.
- U koloni dva, izvor ima četiri simbola umesto pet. Verovatnoće ta četiri simbola su raspoređene po opadanju, kao što je to prikazano u koloni tri. U **sledećem koraku** dva najmanje verovatna simbola su ponovo spojena, pa su zatim nove verovatnoće ponovo raspoređene po opadanju.
- Proces se ponavlja dok se izvor ne svede na dva simbola. U našem primeru dve poslednje verovatnoće su 0.6 i 0.



Slika 6.4.1—Huffman-ov proces redukcije izvora

ORIGINALNI IZVOR			DODELJIVANJE KODA							
Simbol	Verovatnoća	KOD	1		2		3		4	
e	0.3	00	0.3	00	0.3	00	0.4	1	0.6	0
a	0.2	10	0.2	10	0.3	01	0.3	00 ←	0.4	1
c	0.2	11	0.2	11	0.2	10 ←	0.3	01 ←		
b	0.1	001	0.2	010 ←	0.2	11 ←				
d	0.1	0100 ←	0.1	011 ←						
!	0.1	0101 ←								

Slika 6.4.2—Proces dodeljivanja Huffman-ovog koda

- Za prva dva simbola dodeljene su kodne reči 0 i 1. U tom slučaju dodeljujemo “0” simbola sa verovatnoćom pojavljivanja 0.6 i “1” simbolu sa verovatnoćom pojavljivanja 0.4.
- Verovatnoća pojavljivanja 0.6 dobijena je spajanjem dve verovatnoće 0.3 i 0.3. Kodovi dodeljeni ovim verovatnoćama su “0 praćena 0” i “0 praćena 1”. Posle tog dodeljivanja postoje tri simbola sa verovatnoćama 0.4, 0.3 i 0.3 i kodovima 1, 00 i 01 respektivno.
- Verovatnoća 0.4 dobijena je spajanjem dve verovatnoće 0.2 i 0.2. Na tom koraku kodovi dodeljeni ovim verovatnoćama (0.2 i 0.2) su “1 praćen 0” i “1 praćen sa 1”. Posle tog dodeljivanja postoje 4 simbola sa verovatnoćama 0.3, 0.3, 0.2 i 0.2 i kodovima 00, 01, 10 i 11 respektivno.

- U sledećem koraku, druga verovatnoća 0.3 dobijena je spajanjem dve verovatnoće 0.2 i 0.1. Kodovi dodeljeni ovim verovatnoćama (0.2 i 0.1) su “01 praćen 0” i “01 praćen sa 1”. Posle tog dedeljivanja postoji pet simbola sa verovatnoćama pojavljivanja 0.3, 0.2, 0.2, 0.2 i 0.1 sa kodovima 00, 10, 11, 010 i 011 respektivno. Četvrta verovatnoća 0.2 dobijena je spajanjem dve verovatnoće 0.1 i 0.1. U ovom koraku kod dodeljen ovim verovatnoćama (0.1 i 0.1) je “010 praćen sa 0” i “010 praćen sa 1“. Posle tog dodeljivanja postoji šest simbola sa verovatnoćama 0.3, 0.2, 0.2, 0.1, 0.1 i 0.1 i odgovarajućim kodovima 00, 10, 11, 011, 0100 i 0101 respektivno. Na taj način dobili smo Huffman-ove kodove za sve originalne izvorne simbole.

Aritmetičko kodovanje

- Da bi se postigla razumna efikasnost primenom Huffman-ovog kodovanja, niz koji generiše izvor deli se na blokove, a svakom bloku se dodeljuje kodna reč promenljive dužine.
- Na strani dekodera, primenjeni niz je raščlanjen na blokove različite dužine koji odgovaraju individualnim kodnim rečima. Pri tome, postoji korespondencija jedan prema jedan između blokova kodne reci i blokova koji se odnose na izvorni niz.

- **Kod aritmetičkog kodovanja, neznatno različiti izvorni nizovi daju značajno različite kodne nizove** Kao entropijska kodna tehnika, aritmetičko kodovanje postiže veću kompresiju nego Haffman-ovo kodovaje

- Za razliku od Huffman-ovog kodovanja, kod aritmetickog kodovanja **prosečan broj bita se približava teorijskoj granici, entropiji, za bilo koji proizvoljan izvor**.
-
- KodHuffman-ovog kodovanje svakom simbolu se pripisuje kodna reč promenljive dužine.
- U procesu kodovanja, simboli su nezavisno kodovani. Kod aritmetičkog kodovanja, kodnoj reči je pripisanacela ulazna poruka.

Pitanja:

1. Čime se karakteriše predstavljanje teksta u digitalnoj formi?
2. Šta je to entropija izvora informacija?
3. Po čemu se tekstualni podaci razlikuju od audio i video podataka?
4. Kada izvor informacija ima maksimalnu, a kada minimalnu entropiju?
5. Šta su to izvori informacija bez memorije?
6. Šta je to relativna entropija?
7. Kako glasi Shannon-ova teorema kodovanja za bešumni izvor?
8. Čime se karakterise Huffman-ova metoda za projektovanje kodera?
9. Koja je glavna karakteriska aritmetičkog kodovanja?
10. U čemu se razlikuje aritmetičko od Huffman-ovog kodovanja?