دانشکده آموزش های الکترونیکی دانشگاه شیراز

سیستم های چند رسانه ای

فشرده سازی متن

وحید اعتمادی

بهار 94

Entropy

- Which sentence contains the most information?
  - Shiraz temperature in Tir is over 30.
  - Shiraz temperature in Tir is under 20.
- The first statement does not tell us anything new
- The second statement is improbable, giving us a great deal of information
- Statement with unpredictability factor conveys the most new information and surprise

Entropy

- I(A): Amount of information that event A provides
- The amount of information associated with an event should be inversely related to the probability of the event.

$$\text{Definition: } I(A) = \log(1/p_A) = -\log(p_A)$$

- Note: with this definition, an event that is sure to happen (pA = 1) provides I(A) = 0 bits of information

مقدمات تئوري اطلاعات

- آنتروپي η از يک منبع اطلاعاتي $S=\{s_1, s_2, \ldots, s_n\}$ برابر است با:

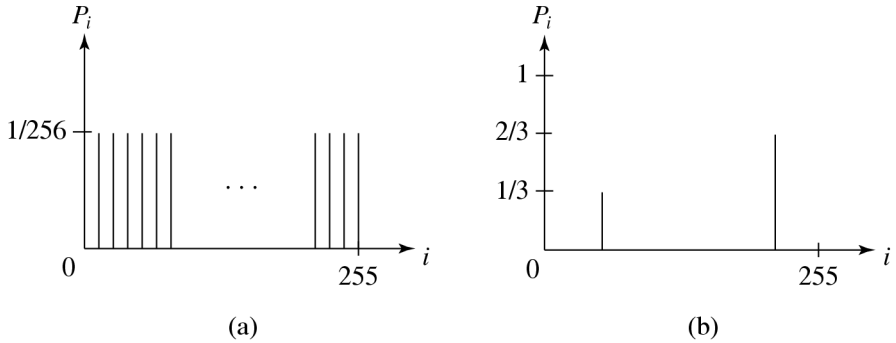
$$\eta = H(S) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}$$

$$= -\sum_{i=1}^{n} p_i \log_2 p_i$$

pi - احتمال رخداد s_i در S

$\log_2 \frac{1}{p_i}$ مشخص کننده ميزان اطلاعات موجود در s_i موجود است، بنابراين مي‌تواند تعيين کننده تعداد بيت‌هاي مورد نياز براي کد کردن s_i باشد.

مقدمات تئوري اطلاعات



هيستوگرام سطح خاكستري دو تصوير

- (a) هيستوگرام يک تصوير با شدت‌هاي توزيع سطح خاكستري يک‌نواخت و مساوي را نشان مي‌دهد، به ازاي هر i داريم:

$p_i = 1/256$ بنابراين آنتروپي اين تصوير برابر است با:

$\log_2 256 = 8$

$$\sum_{i=1}^{256} \frac{1}{256} \log_2 256 = 8$$

- (b) هيستوگرام يک تصوير با دو مقدار ممكن را نشان ميدهد. آنتروپي آن 0.92 است.

$$\frac{1}{3}\log_2 3 + \frac{2}{3}\log_2 \frac{3}{2}$$

کدینگ Variable-Lenght (VLC)

- **الگوریتم شانون-فانو (روش بالا به پایین):**

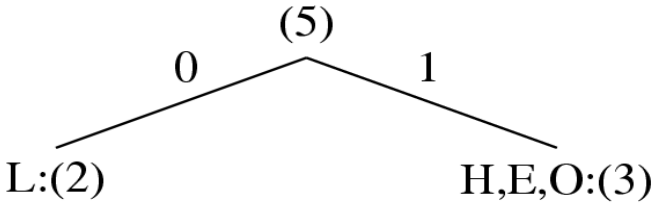
1. نمادها را بر اساس شمار فراواني رخدادشان مرتب کنید.
2. نمادها را به صورت بازگشتي به دو قسمت تقسیم کنید، هر کدام با تعداد شماره‌هاي تقریبا برابر، تا جايي‌که هر قسمت فقط یک نماد داشته باشد.

- مثال: کدینگ کلمه HELLO

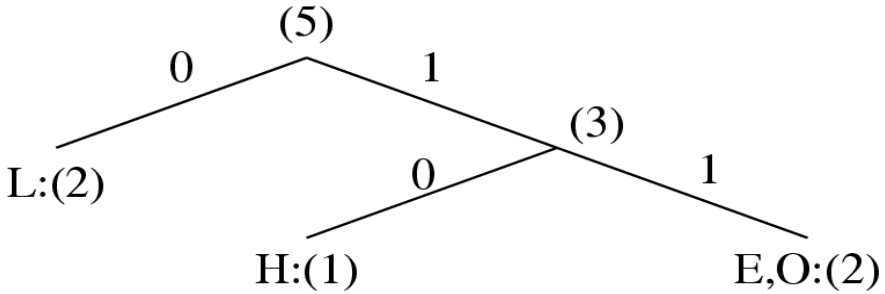
نماد	H	E	L	O
تعداد	1	1	2	1

شمار فراواني نمادها در HELLO

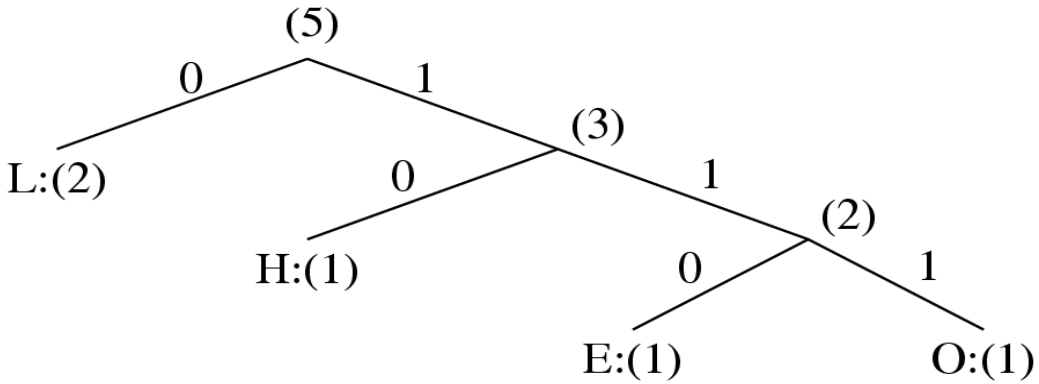
VLC



(a)



(b)



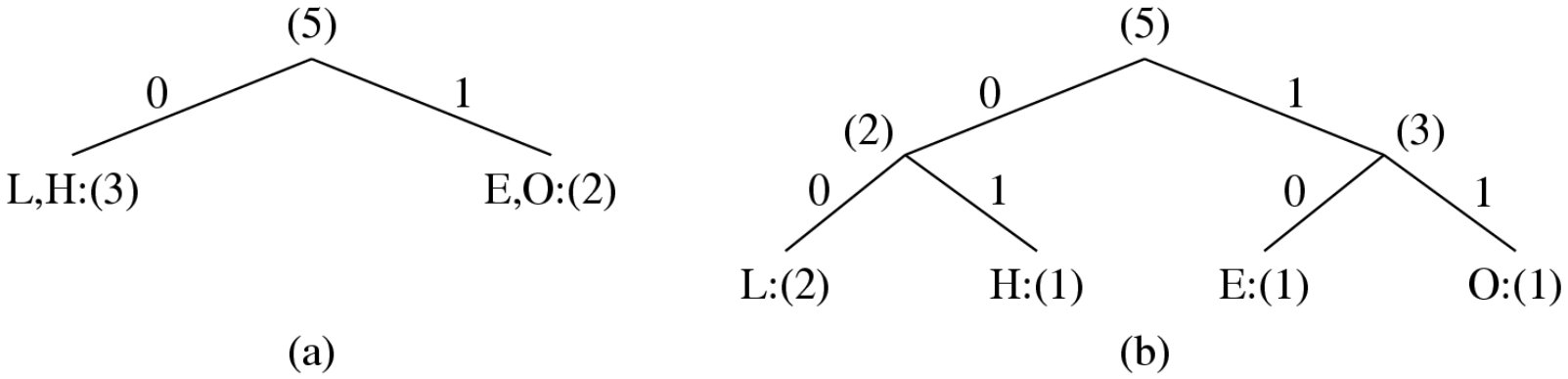
(c)

VLC

نتايج اجراي شانون-فانو بر روي HELLO

Symbol	Count	$\text{Log}_2 \frac{1}{p_i}$	Code	# of bits used
L	2	1.32	0	1
H	1	2.32	10	2
E	1	2.32	110	3
O	1	2.32	111	3
			TOTAL # of bits:	10

VLC



- درخت کدینگ دیگري توسط شانون-فانو براي HELLO

VLC

- نتايج ديگر از اجراي شانون-فانو بر روي HELLO

Symbol	Count	$\text{Log}_2 \frac{1}{p_i}$	Code	# of bits used
L	2	1.32	00	4
H	1	2.32	01	2
E	1	2.32	10	2
O	1	2.32	11	2
			TOTAL # of bits:	10

VLC

- **الگوريتم كدينگ هافمن** (Huffman) روند پايين به بالا:

1. تمام نمادها را بر اساس شمار فراواني آن‌ها در يک ليست مرتب كنيد.
2. تا زماني‌كه فقط يک نماد باقي بماند، ادامه دهيد:

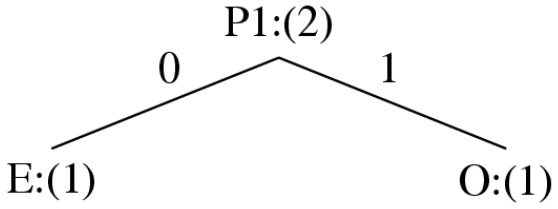
(1) دو نمادي كه داراي كمترين فراواني هستند را از ليست برداريد. يک زير درخت هافمن ايجاد كنيد كه اين دو نماد را به عنوان گره فرزند قرار داده و يک گره پدر ايجاد كنيد.

(2) جمع تعداد فرزندان را به نود پدر اختصاص داده و در ليست به عنوان ترتيب قرارگيري وارد كنيد.

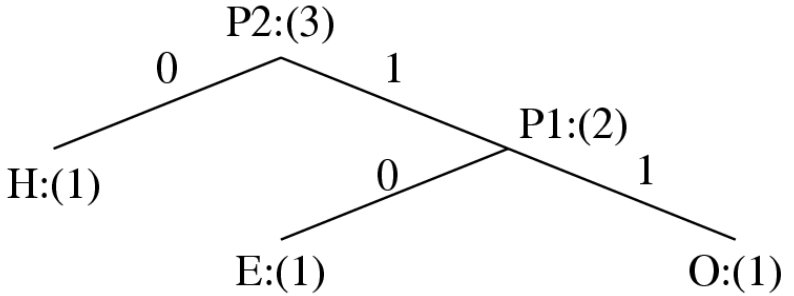
(3) فرزندان را از ليست حذف كنيد.

(4) يک كليدواژه براي هر برگ بر اساس مسير از ريشه تعيين مي‌شود.

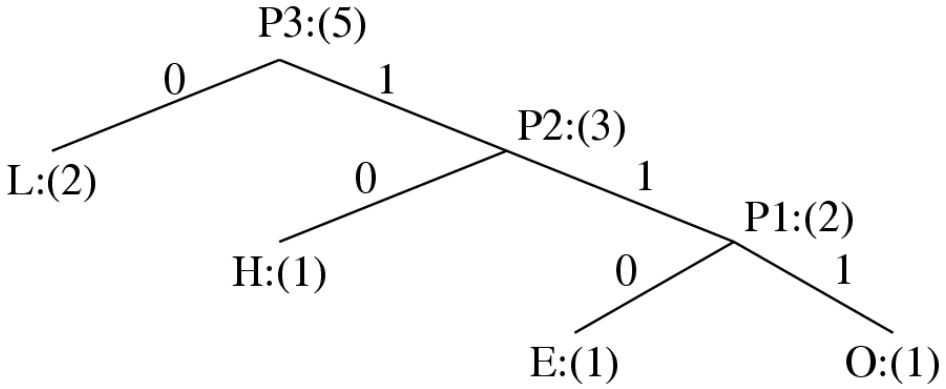
VLC



(a)



(b)



(c)

VLC

- نمادهاي جديد p1 ، p2 ، p3 براي اشاره به نودهاي پدر در درخت كدينگ هافمن توليد شده‌اند. محتويات موجود در ليست، در زير مشخص شده است:

	L(2)	H(1)	E(1)	O(1)
شروع:				
بعد از تكرار :(a)	L	H	P1	
بعد از تكرار :(b)		L	P2	
بعد از تكرار :(c)		P3		

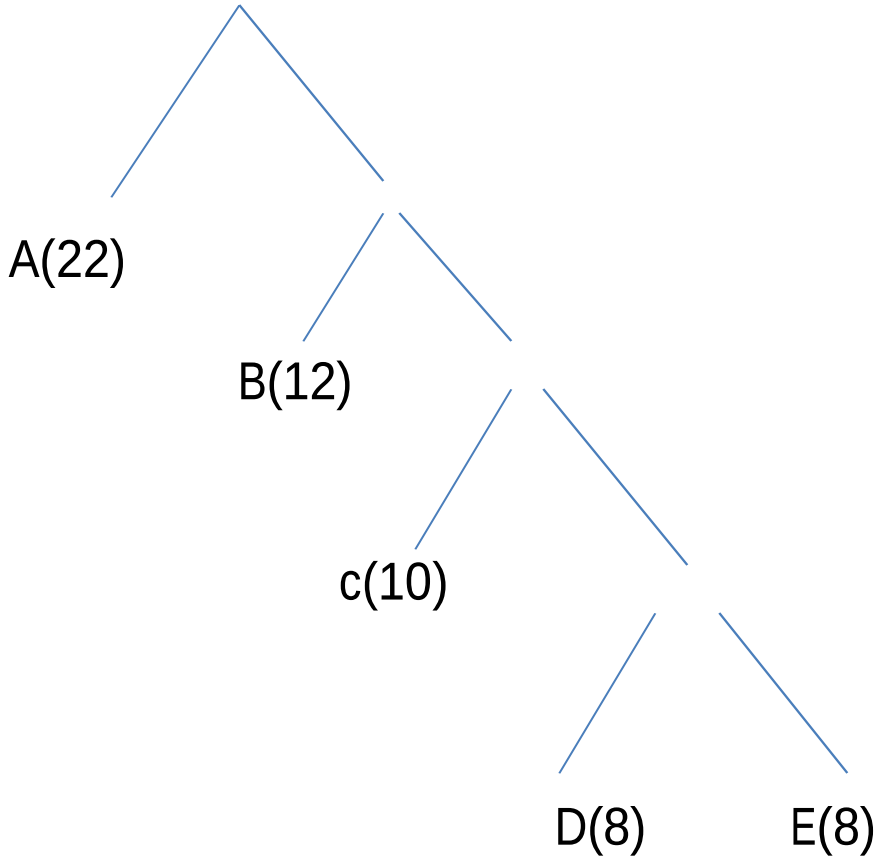
سوال

- مجموعه سمبل‌هاي موجود در سيستم شامل پنج عنصر متفاوت A، B، C، D و E است. تعداد تكرار اين سمبل‌ها در توالي ورودي به ترتيب 22، 12، 10، 8 و 8 مرتبه است.

- الف) با استفاده از روش Shannon-Fano كد هريک از سمبل‌ها را مشخص نماييد. و طول ميانگين كد براي اين روش را حساب كنيد.

- ب) با استفاده از روش Huffman كد هريک از سمبل‌ها را مشخص نماييد. و طول ميانگين كد براي اين روش را حساب كنيد.

- ج) انتروپي را براي اين سيستم و مجموعه سمبل‌ها محاسبه كنيد.


A(22)
B(12)
c(10)
D(8)
E(8)

- Average # of bit: 22*1 +12*2+10*3+ 4*8+4*8 =140 so:

  140/60=2.33

Entropy(S)= .36 *1.47 + .2*2.32+.17*2.55+ .13* 2.94 +.13*2.94= 2.19

Avg. # of bits ≈ Entropy(S)

HW 1

- بخش ب و ج را محاسبه کنید.

خصوصيات كدينگ هافمن

- **1- خصوصيت منحصر به‌فرد پيشوندي:** هيچ كد هافمني پيشوند كد هافمن ديگر نيست. اين نكته مانع ايجاد هر گونه ابهامي در مرحله رمزگشايي مي‌شود.

- **2- بهينگي:** حداقل افزونگي كد.

- دو نماد با كمترين فراواني، داراي طول مساوي براي كد هافمنشان هستند و تنها در آخرين بيت تفاوت دارند.
- نمادهايي با فراواني بيشتر داراي كد هافمن كوتاه‌تري نسبت به نمادهاي با فراواني كمتر هستند. $\bar{l} < \eta + 1$
- ميانگين طول كد، براي يک منبع اطلاعات S حتما از $\eta+1$ كمتر است.