アルゴリズムとデータ構造III
13回目：1月17日（木）

暗号：符号化：テキスト圧縮

授業資料　http://ir.cs.yamanashi.ac.jp/~ysuzuki/algorithm3/index.html

授業の予定（中間試験まで）

1	10/11	スタック（後置記法で書かれた式の計算）
2	10/18	文脈自由文法
3	10/25	構文解析　CKY法
4	11/01	構文解析　CKY法，チャート法
5	11/08	構文解析　CKY法，チャート法
6	11/15	構文解析　チャート法
7	11/29	グラフ（動的計画法，ダイクストラ法，DPマッチング）
8	12/06	グラフ（DPマッチング，ビームサーチ，A＊アルゴリズム）
9	12/13	中間試験

授業の予定（中間試験以降）

10	12/17	全文検索アルゴリズム（simple search, KMP）
11	12/20	全文検索アルゴリズム（BM, Aho-Corasick）
12	01/10	全文検索アルゴリズム（Aho-Corasick），データ圧縮
13	01/17	暗号（黄金虫，踊る人形） 符号化（モールス信号，Zipfの法則，ハフマン符号）テキスト圧縮
14	01/28（月）	テキスト圧縮（zip） 音声圧縮　ADPCM，MP3 音声圧縮（CELP），画像圧縮（JPEG）
15	01/31	期末試験

補講の実施日時

- 1月28日（月）5時限　B2-11

期末試験の日時変更の提案

- 2月7日（木）2時限　→　1/31日（木）2時限？
- 2月7日（木）2時限は鈴木が入試作業のため試験監督が出来ないため

- 1月31日（木）2時限　B2-11に決定

本日のメニュー

- モールス信号
- 暗号
  - 黄金虫（The gold bug）
  - 踊る人形（The Adventure of the Dancing Men）
- 符号化
- テキスト圧縮

データ圧縮

- 対象データ
  - テキスト
  - 音声
    - 音楽
    - 話し声
  - 画像
  - 動画
- 圧縮方式
  - 可逆圧縮
  - 不可逆圧縮

モールス信号の符号

- ・(短点)と−(長点)を用いてアルファベットを表現する
- 情報を早く送るための工夫
  - よく使われる文字(例えばe,t)は短い
    - e: ・ (短点1文字)
    - t: − (長点1文字)
  - あまり使われない文字(例えばqは4文字)は長い
    - q: −−・−

モールス信号の符号

- ・(短点)と−(長点:短点3つ分の長さ)を用いてアルファベットを表現する
- 区切り記号
  - 文字の切れ目:短点3つ分の間隔
  - 単語の切れ目:短点7つ分の間隔
- L: ・−・・ (LifeカードのCMに使われていた)
- SOS: ・・・ −−− ・・・

携帯電話の文字キー

- 覚えやすい，わかりやすい並べ方だが，
- 文字毎の出現確率を調べることで，キーを押す回数を減らすことが出来る．

小説の中で暗号解読の解説

- 黄金虫(The gold bug)
  - エドガー・アラン・ポー
- 踊る人形(The Adventure of the Dancing Men)
  - アーサー・コナン・ドイル

黄金虫(エドガー・アラン・ポー)に出てくる暗号(換字式)

- 暗号解読の解説
- 暗号は多分英語
- 英語は文字によって出現確率が違う
  - 出現確率の高い方から並べると
    - eaoidhnrstuycfglmwbkpqxz
  - eは頻出，eeも頻出
  - theも頻出
- 対応がとれた文字は置き換え，前後の文字を推理する

踊る人形
(The Adventure of the Dancing Men)
アーサー・コナン・ドイル

- 人形の形をアルファベットに置き換える
- 暗号の元の文は英語だろう
- 旗は語の区切りを意味するのでは？
- 頻出する形がある → eだろう
- 対応がとれた形を文字に置き換える

自然言語の統計的性質

- 文字の使用頻度（英語） _はスペース

順位	文字	%	2	%	3	%	4	%
1	_	17.4	e_	3.0	_th	1.6	_the	1.2
2	e	9.7	_t	2.4	the	1.3	the_	1.0
3	t	7.0	th	2.0	he_	1.3	_of_	0.6
4	a	6.1	he	1.9	_of	0.6	and_	0.4
5	o	5.9	_a	1.7	of_	0.6	_and	0.4
6	i	5.5	s_	1.7	ed_	0.5	_to_	0.4
7	n	5.5	d_	1.5	_an	0.5	ing_	0.3

単語の使用頻度

順位	単語	%	2	%	3	%
1	the	6.1	of the	0.9	one of the	0.03
2	of	3.5	in the	0.5	as well as	0.02
3	and	2.7	to the	0.3	the United States	0.02
4	to	2.5	on the	0.2	out of the	0.02
5	a	2.1	and the	0.2	some of the	0.01
6	in	1.9	for the	0.1	the end of	0.01
7	that	0.9	to be	0.1	the fact that	0.01

ジップの法則(Zipf's law)

経験則：「あるタイプの現象が生起する確率はその現象の生起する順位に反比例する」

- ジップの法則が当てはまる事象
  - 文字毎の出現頻度
  - コンピュータにおけるコマンドの使用頻度
  - Webページのアクセス頻度
  - 都市の人口
  - 文献の参照回数
  - 会社でのランク（役職）と給料など

会社でのランク（役職）と給料

ランク	役職	給料	割合(確率)
1	社長	2,000,000	0.54
2	副社長	1,000,000	0.27
3	部長	700,000	0.19
.....	.....		.....

- 「あるタイプの現象が生起する確率はその現象の生起する順位に反比例する」

単語の出現頻度分布

- ジップの法則(Zipf's law)を単語の出現頻度に適用すると

単語の出現頻度分布

- ジップの法則(Zipf's law): 単語の出現順位(r)と出現頻度(f)は反比例の関係にある

$$r = \frac{C}{f} \qquad f = \frac{C}{r}$$

n番目の単語の出現確率 P_n

$$P_n = \frac{C}{n}$$

Cは定数
低頻度の語には当てはまらない

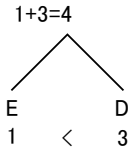
ハフマン符号

- 2分木を使って文字の出現頻度順に並べる
- 葉＝文字
- 浅い:符号長が短い，深い:符号長が長い
- 平均符号長が最小になることが保証されている

ハフマン符号の作り方　1/5

順位	文字	頻度
1	A	9
2	B	7
3	C	5
4	D	3
5	E	1

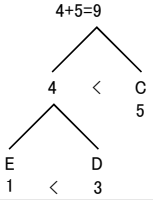
- 頻度の低い文字を2文字(D,E)選び，頻度の低い方を左の葉，頻度の高い方を右の葉に置き，2分木をつくる．
- ルートノードには2つの葉の頻度の和を書き込む



ハフマン符号の作り方　2/5

順位	文字	頻度
1	A	9
2	B	7
3	C	5
4	D+E	4

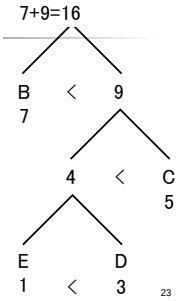
- 次に頻度の低い文字(C)を選び，DE連合と頻度を比較し，Cの頻度が高ければ右の葉にする
- ルートノードには頻度の和を書き込む



ハフマン符号の作り方　3/5

順位	文字	頻度
1	A	9
2	B	7
3	C+D+E	9

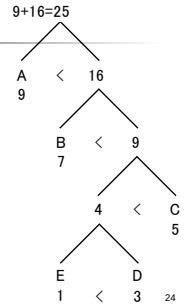
- 次に頻度の低い文字(B)を選び，CDE連合と頻度を比較し，Bの頻度が低ければ左の葉にする
- ルートノードには頻度の和を書き込む



ハフマン符号の作り方　4/5

順位	文字	頻度
1	A	9
2	B+C+D+E	16

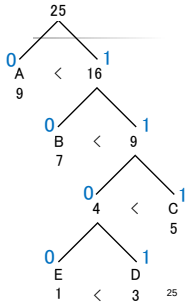
- 次に頻度の低い文字(A)を選び，BCDE連合と頻度を比較し，Aの頻度が低ければ左の葉にする
- ルートノードには頻度の和を書き込む



ハフマン符号の作り方 5/5

- 左のノードに0,
  右のノードに1を付与する

順位	文字	頻度	符号
1	A	9	0
2	B	7	10
3	C	5	111
4	D	3	1101
5	E	1	1100



文字⇔ハフマン符号の変換

順位	文字	頻度	符号
1	A	9	0
2	B	7	10
3	C	5	111
4	D	3	1101
5	E	1	1100

- 01011111011100
  - 0|10|111|1101|1100
  - A|B|C|D|E
- 11000110110111
  - 1100|0|1101|10|111
  - E|A|D|B|C
- BBCEDA
  - 1010111110011010

ASCII文字コード(8bit)からハフマン符号へ

順位	文字	頻度	符号
1	A	9	0
2	B	7	10
3	C	5	111
4	D	3	1101
5	E	1	1100

- A
- ASCII: 01000001 (0x41) 8bit
- Huffman: 0 : 1bit

- E
- ACSII: 01000101 (0x45) 8bit
- Huffman: 1100 : 4bit

ハフマン符号の特徴

- 各記号がリーフノード(葉)に対応している
  - ハフマン符号列を左からトレースすることで，記号の区切りが分かる
  - 区切り記号を入れる必要がない