Index Compression (Chapter 5)

Algorithm 1 (Variable byte code)
A number n is encoded in variable byte code in the following procedure:

1. *Take a binary representation of n with padding to the length of a multiple of 7.*
2. *Split into of 7 bit blocks right-to-left.*
3. *Add 1 to the beginning of the last block and 0 to the beginning of all previous blocks.*

Example: $VB(824) = 00000110\,10111000$

Definition 1 (α code)
Unary code, also referred to as α code, is a coding type where a number n is represented by a sequence of n 1s (or 0s) and terminated with one 0 (or 1). That is, 6 in unary code is 1111110 (or 0000001). The alternative representation in parentheses is equivalent but for this course we use the default representation.

Definition 2 (γ code)
γ code is a coding type, that consists of an offset and its length: $\gamma(n) = \alpha\big($length of offset$(n)\big)$,offset(n). Offset is a binary representation of a number n without the highest bit (1). The length of this offset encoded in the unary (α) code. Then the number 60 is encoded in γ as 111110,11100.

Definition 3 (δ code)
A number n is encoded in δ code in the following way: $\delta(n) = \gamma\big($length of offset$(n)\big)$,offset(n). Analogously, 600 is encoded in δ as 1110,001,001011000.

Definition 4 (Zipf's law)
Zipf's law says that the i-th most frequent term has the frequency $\frac{1}{i}$. In this exercise we use the dependence of the Zipf's law $cf_i \propto \frac{1}{i} = ci^k$ where cf_i is the number of terms t_i in a given collection with $k = -1$.

Definition 5 (Heaps' law)
Heaps' law expresses an empiric dependency of collection size (number of all words) T and vocabulary size (number of distinct words) M by $M = kT^b$ where $30 \leq k \leq 100$ and $b \approx \frac{1}{2}$.

Exercise 5/1

Count variable byte code for the postings list $\langle 777, 17\,743, 294\,068, 31\,251\,336\rangle$. Bear in mind that the gaps are encoded. Write in 8-bit blocks.

Encode the list of gaps $\langle 777, 16\,966, 276\,325, 30\,957\,268\rangle$. Variable byte code of the gaps:

- $VB(777) = \mathbf{0}0000110\,\mathbf{1}0001001$
- $VB(16\,966) = \mathbf{0}0000001\,\mathbf{0}0000100\,\mathbf{1}1000110$
- $VB(276\,325) = \mathbf{0}0010000\,\mathbf{0}1101110\,\mathbf{1}1100101$
- $VB(30\,957\,268) = \mathbf{0}0001110\,\mathbf{0}1100001\,\mathbf{0}0111101\,\mathbf{1}1010100$

Result: $VB(\langle 777, 17\,743, 294\,068, 31\,251\,336\rangle) = 00000110\,10001001\,00000001\,00000100\,11000110\,00010000\,01101110\,11100101\,00001110\,01100001\,00111101\,11010100$

Exercise 5/2

Count γ and δ codes for the numbers 63 and 1023.

According to the definition 2 it is necessary to count the offsets as binary representations without the highest bit $63_{10} = 111111_2$ and offset(63) = 11111. Offset length is encoded in α as $|11111| = 5 \rightsquigarrow \alpha(5) = 111110$. Finally, $\gamma(63) = 111110, 11111$. Analogically for 1023. $1023_{10} = 1111111111_2$, offset is 111111111, its length is $|111111111| = 9 \rightsquigarrow \alpha(9) = 1111111110$. Then $\gamma(1023) = 1111111110, 111111111$.

δ is a little more complicated. First we count the offset $63 = 11111$ and its length $|11111| = 5$. The value of 5 we encode in γ so $\gamma(5) = 110, 01$. By definition 3 we have $\delta(63) = 110, 01, 11111$. And finally, $\delta(1023) = 1110, 001, 111111111$.

Exercise 5/3

Calculate the variable byte code, γ code and δ code of the postings list $P = [32, 160, 162]$. Note that gaps are encoded. Include intermediate results (offsets, lengths).

---

offset $32 = 00000$ and $\alpha(|00000|) = 111110 \rightsquigarrow \gamma(32) = 111110, 00000$
offset $128 = 0000000$ and $\alpha(|0000000|) = 11111110 \rightsquigarrow \gamma(128) = 11111110, 0000000$
offset $2 = 0$ and $\alpha(|0|) = 10 \rightsquigarrow \gamma(2) = 10, 0$
$\gamma(P) = 11111000000111111100000000100$

offset $32 = 00000$ and $\gamma(|00000|) = 110, 01 \rightsquigarrow \delta(32) = 110, 01, 00000$
offset $128 = 0000000$ and $\gamma(|0000000|) = 110, 11 \rightsquigarrow \delta(128) = 110, 11, 0000000$
offset $2 = 0$ and $\gamma(|0|) = 0 \rightsquigarrow \delta(2) = 0, , 0$
$\delta(P) = 110010000011011000000000$

Exercise 5/4

Consider a posting list with the following list of gaps

$$G = [4, 6, 1, 2048, 64, 248, 2, 130].$$

Using variable byte encoding,

- What is the largest gap in G that you can encode in 1 byte?
- What is the largest gap in G that you can encode in 2 bytes?
- How many bytes will G occupy after encoding?

- 64
- 2048
- 11

Exercise 5/5

From the following sequence of γ-encoded gaps, reconstruct first the gaps list and then the original postings list. Recall that the α code encodes a number n with n 1s followed by one 0.

$$1110001110101011111101101111011$$

$[1110001, 11010, 101, 11111011011, 11011] \rightsquigarrow [1001, 110, 11, 111011, 111] \rightsquigarrow [9, 6, 3, 59, 7] \rightsquigarrow [9, 15, 18, 77, 84]$

Exercise 5/6

What does the Zipf's law say?

Answers can vary. For official definition refer to the Manning book.

Exercise 5/7

What does the Heaps' law say?

Answers can vary. For official definition refer to the Manning book.

Exercise 5/8

A collection of documents contains 4 words: *one*, *two*, *three*, *four* of decreasing word frequencies f_1, f_2, f_3 and f_4. The total number of tokens in the collection is 5000. Assume that the Zipf's law holds for this collection perfectly. What are the word frequencies?

We use the Zipf's law in Definition 4. The least frequent term is *four*, then *three*, *two* and the most frequent is *one*. Applying the Zipf's law we get

$$\begin{aligned} cf_1 + cf_2 + cf_3 + cf_4 &= 5000 \\ c \cdot 1^{-1} + c \cdot 2^{-1} + c \cdot 3^{-1} + c \cdot 4^{-1} &= 5000 \\ c + \tfrac{1}{2}c + \tfrac{1}{3}c + \tfrac{1}{4}c &= 5000 \\ 12c + 6c + 4c + 3c &= 60000 \\ 25c &= 60000 \\ c &= 2400 \end{aligned}$$

and, plugging in to the formula $cf_i = ci^{-1}$, we obtain the term frequency values:

$$cf_1 = 2400\tfrac{1}{1} = 2400$$

$$cf_2 = 2400\tfrac{1}{2} = 1200$$

$$cf_3 = 2400\tfrac{1}{3} = 800$$

$$cf_4 = 2400\tfrac{1}{4} = 600$$

Exercise 5/9

How many distinct terms are expected in a document of 1,000,000 tokens? Use the Heaps' law with parameters $k = 44$ and $b = 0.5$

By Definition 5,

$$44 \times 1{,}000{,}000^{0.5} = 44{,}000.$$

Scoring, Term Weighting, and the Vector Space Model (Chapter 6)

Definition 6 (Inverse document frequency)
Inverse document frequency of a term t is defined as

$$idf_t = \log\left(\frac{N}{df_t}\right)$$

where N is the number of all documents and df_t (the document frequency of t) is the number of documents that contain t.

Definition 7 (tf-idf weighting scheme)
In the tf-idf *weighting scheme, a term t in a document d has weight*

$$tf\text{-}idf_{t,d} = tf_{t,d} \cdot idf_t$$

where $tf_{t,d}$ is the number of tokens t (the term frequency of t) in a document d.

Definition 8 (ℓ^2 (cosine) normalization)
A vector v is cosine-normalized by

$$v_j \leftarrow \frac{v_j}{||v||} = \frac{v_j}{\sqrt{\sum_{k=1}^{|v|} {v_k}^2}}$$

where v_j is the element at the j-th position in v.

Exercise 6/1

Consider the frequency table of the words of three documents doc_1, doc_2, doc_3 below. Calculate the *tf-idf* weight of the terms *car, auto, insurance,* and *best* for each document. *idf* values of terms are in the table.

	doc_1	doc_2	doc_3	*idf*
car	27	4	24	1.65
auto	3	33	0	2.08
insurance	0	33	29	1.62
best	14	0	17	1.5

Table 1: Exercise.

After counting *tf-idf* weights by Definition 7 individually for each term we get the following table

	tf-idf		
	doc_1	doc_2	doc_3
car	44.55	6.6	39.6
auto	6.24	68.64	0
insurance	0	53.46	46.98
best	21	0	25.5

Table 2: Solution.

Exercise 6/2

Count document representations as normalized Euclidean weight vectors for each document from the previous exercise. Each vector has four components, one for each term.

Normalized Euclidean weight vectors are counted by Definition 8. Denominators m_{doc_n} for the individual documents are

$$m_{\text{doc}_1} = \sqrt{44.55^2 + 6.24^2 + 21^2} = 49.6451$$

$$m_{\text{doc}_2} = \sqrt{6.6^2 + 68.64^2 + 53.46^2} = 87.2524$$

$$m_{\text{doc}_3} = \sqrt{39.6^2 + 46.98^2 + 25.5^2} = 66.5247$$

and the document representations are

$$d_1 = \left(\frac{44.55}{49.6451}; \frac{6.24}{49.6451}; \frac{0}{49.6451}; \frac{21}{49.6451}\right) = (0.8974; 0.1257; 0; 0.423)$$

$$d_2 = \left(\frac{6.6}{87.2524}; \frac{68.64}{87.2524}; \frac{53.46}{87.2524}; \frac{0}{87.2524}\right) = (0.0756; 0.7876; 0.6127; 0)$$

$$d_3 = \left(\frac{39.6}{66.5247}; \frac{0}{66.5247}; \frac{46.98}{66.5247}; \frac{25.5}{66.5247}\right) = (0.5953; 0; 0.7062; 0.3833)$$

Exercise 6/3

Based on the weights from the last exercise, compute the similarity scores (scalar products) of the three documents for the query *car insurance*. Use each of the two weighting schemes:

a) Term weight is 1 if the query contains the word and 0 otherwise.

b) Euclidean normalized *tf-idf*.

Please note that a document and a representation of this document are different things. Document is always fixed but the representations may vary under different settings and conditions. In this exercise we fix document representations from the last exercises and will count similarity scores for query and documents under two different representations of the query. It might be helpful to view on a query as on another document, as it is a sequence of words.

We count the similarity scores for **a)** as the dot products of the representation of the query $q = (1, 0, 1, 0)$ with representations of the documents d_n from the last exercise:

$$q \cdot d_1 = 1 \cdot 0.8974 + 0 \cdot 0.1257 + 1 \cdot 0 + 0 \cdot 0.423 = 0.8974$$

$$q \cdot d_2 = 1 \cdot 0.0756 + 0 \cdot 0.7876 + 1 \cdot 0.6127 + 0 \cdot 0 = 0.6883$$

$$q \cdot d_3 = 1 \cdot 0.5953 + 0 \cdot 0 + 1 \cdot 0.7062 + 0 \cdot 0.3833 = 1.3015$$

For **b)** we first need the normalized *tf-idf* vector q, which is obtained by dividing each component of the query by the length of *idf* vector $\sqrt{1.65^2 + 0^2 + 1.62^2 + 0^2} = 2.3123$

	tf	*idf*	*tf-idf*	q
car	1	1.65	1.65	0.7136
auto	0	2.08	0	0
insurance	1	1.62	1.62	0.7006
best	0	1.5	0	0

Table 3: Process of finding the Euclidean normalized *tf-idf*.

Now we multiply q with the document vectors and we obtain the similarity scores:

$$q \cdot d_1 = 0.7136 \cdot 0.8974 + 0 \cdot 0.1257 + 0.7006 \cdot 0 + 0 \cdot 0.423 = 0.6404$$

$$q \cdot d_2 = 0.7136 \cdot 0.0756 + 0 \cdot 0.7876 + 0.7006 \cdot 0.6127 + 0 \cdot 0 = 0.4832$$

$$q \cdot d_3 = 0.7136 \cdot 0.5953 + 0 \cdot 0 + 0.7006 \cdot 0.7062 + 0 \cdot 0.3833 = 0.9196$$