Textová informácia

Úloha: Napíšte text „Práca s textom“ v Poznámkovom bloku a v MS Worde. Súbory uložte a porovnajte ich veľkosť. Prečo sú také výrazné rozdiely vo veľkosti súborov?

Riešenie: Existuje veľa programov od rôznych softvérových firiem na spracovanie textu, ktoré sa líšia výkonnosťou, možnosťami spracovania textu atď. Tie možno rozdeliť podľa toho, čo všetko dokážu, aké obsahujú možnosti na prácu s textom.

Prvú skupinu tvoria **jednoduché textové editory**, ktoré sú súčasťou programovacích jazykov, operačných systémov. Slúžia na zápis zdrojových textov programov, systémových príkazov, poznámok, jednoduchých textov bez nároku na úpravu vloženého textu. Umožňujú napísať text štandardným typom písma, opravovať, kopírovať, presúvať a vymazávať text, vyhľadávať v texte znaky, slová, prípadne ich nahrádzať inými znakmi. Text sa dá jednoducho vytlačiť.

Do druhej skupiny patria **textové editory na tvorbu kancelárskej dokumentácie**, korešpondencie,

ktorá nemusí mať tlačovú kvalitu. Na rozdiel od prvej skupiny editorov poskytujú rôzne druhy písma, písmo rôznej veľkosti, zarovnanie textu... Poskytujú väčší komfort ako elektronický písací stroj. Patria sem **Poznámkový blok, WordPad**, ktoré sú súčasťou príslušenstva OS Windows.

V tretej skupine sú **textové procesory**, ktoré už tvoria prechod k programom pre malú publikačnú činnosť (angl. desktop publishing DTP – tvorba časopisov, novín), samy dokážu upraviť text do formátu, ktorý poznáme z časopisov a kníh. Procesory dokážu pracovať s množstvom typov písma ľubovoľnej veľkosti, vkladať obrázky, tabuľky, grafy a iné objekty, automaticky číslovať kapitoly, vytvárať obsah, register. Umožňujú zautomatizované činnosti zapísať v tvare jedného príkazu – makroinštrukcie, dopĺňať dokument textom z pripojenej databázy a podobne. Do tejto skupiny patrí napr. Microsoft Word, alebo OO Writer.

A toto je dôvod, prečo tie súbory majú veľmi rozdielnu veľkosť. Word pri ukladaní, ukladá aj spôsob formátovania – akým je to písmom napísané, akou veľkosťou, ako je odsadený odsek...

Poznámkový blok ukladá len písmená, formátovanie neukladá.

Reprezentácia znakov v počítači (dvui Počítačové systémy 1**)**

Všetky bežné počítače reprezentujú údaje, programy, alebo postupnosti inštrukcií v binárnej forme. Nech je to obrázok, text, video, zvuk, alebo nejaký spustiteľný program, všetko je na tej najnižšej úrovni uložené a spracovávané ako vhodná kombinácia jednotiek a núl.

Reprezentácia písmen a iných znakov prešla, a ešte stále prechádza, zložitým vývojom. Prvé kódovanie znakov vzniklo už v 40. Rokoch 19. Storočia s príchodom telegrafov. Bola to známa Morseova abeceda, obsahujúca 26 znakov anglickej abecedy a 10 číslic. Medzery medzi znakmi a medzery medzi slovami sa kódovali ako vhodne dlhé odmlky. Morseova abeceda teda používa na zakódovanie textu krátke pípnutie (bodka), dlhé pípnutie (čiarka) a odmlku. Odmlka, ktorá je dlhá ako dlhé pípnutie predstavuje oddeľovač znakov, a odmlka, ktorá je dlhá ako dve dlhé pípnutia a jedno krátke pípnutie, predstavuje medzeru medzi slovami. Práve odmlky sú dôvodom, prečo Morseova abeceda nie je považovaná za binárnu, ale ternárnu (používa až tri značky). Pre použitie na kódovanie znakov v počítači je teda nevhodná.

Až v ére prvých počítačov, v roku 1963, sa štandardizovalo binárne kódovanie znakov ASCII (American Standard Code for Information Interchange) pôvodne určené pre ďalekopisy. Hranice medzi znakmi sa jednoducho vyriešili tak, že každý znak má rovnako dlhý 7 miestny binárny kód. Z medzery medzi slovami sa stal obyčajný znak s binárnym kódom 1000000.

Dec	Hex	Zkratka	Význam
0	00	NUL	Null character
1	01	SOH	Start of Header
2	02	STX	Start of Text
3	03	ETX	End of Text
4	04	EOT	End of Transmission
5	05	ENQ	Enquiry
6	06	ACK	Acknowledge
7	07	BEL	Bell
8	08	BS	Backspace
9	09	HT	Horizontal Tab
10	0a	LF	Line Feed
11	0b	VT	Vertical Tab
12	0c	FF	Form Feed
13	0d	CR	Carriage Return
14	0e	SO	Shift Out
15	0f	SI	Shift In
16	10	DLE	Data Link Escape
17	11	DC1	(XOn)
18	12	DC2	
19	13	DC3	(XOff)
20	14	DC4	
21	15	NAK	Negative Acknowledge
22	16	SYN	Synchronous Idle
23	17	ETB	End of Transmission Block
24	18	CAN	Cancel
25	19	EM	End of Medium
26	1a	SUB	Substitute
27	1b	ESC	Escape
28	1c	FS	File Separator
29	1d	GS	Group Separator
30	1e	RS	Record Separator

Dec	Hex	Znak
32	20	SP (mezera)
33	21	!
34	22	"
35	23	#
36	24	$
37	25	%
38	26	&
39	27	'
40	28	(
41	29	)
42	2a	*
43	2b	+
44	2c	,
45	2d	-
46	2e	.
47	2f	/
48	30	0
49	31	1
50	32	2
51	33	3
52	34	4
53	35	5
54	36	6
55	37	7
56	38	8
57	39	9
58	3a	:
59	3b	;
60	3c	<
61	3d	=
62	3e	>

Dec	Hex	Znak
64	40	@
65	41	A
66	42	B
67	43	C
68	44	D
69	45	E
70	46	F
71	47	G
72	48	H
73	49	I
74	4a	J
75	4b	K
76	4c	L
77	4d	M
78	4e	N
79	4f	O
80	50	P
81	51	Q
82	52	R
83	53	S
84	54	T
85	55	U
86	56	V
87	57	W
88	58	X
89	59	Y
90	5a	Z
91	5b	[
92	5c	\
93	5d	]
94	5e	^

Dec	Hex	Znak
96	60	`
97	61	a
98	62	b
99	63	c
100	64	d
101	65	e
102	66	f
103	67	g
104	68	h
105	69	i
106	6a	j
107	6b	k
108	6c	l
109	6d	m
110	6e	n
111	6f	o
112	70	p
113	71	q
114	72	r
115	73	s
116	74	t
117	75	u
118	76	v
119	77	w
120	78	x
121	79	y
122	7a	z
123	7b	{
124	7c	|
125	7d	}
126	7e	~

31	1f	US	Unit Separator	63	3f	?	95	5f	_	127	7f	DEL (delete)

Kódovanie ASCII obsahuje okrem znakov, ktoré sa zobrazujú na výstupných zariadeniach aj 33 riadiacich znakov (znaky s kódmi $(0000000)_2$ až $(0011111)_2 = (31)_{10}$ a znak $(1111111)_2 = (127)_{10}$). Tie boli určené pre ovládanie tlačiarní ďalekopisov pripomínajúcich mechanický písací stroj, alebo ako rôzne riadiace dáta určené na ovládanie periférnych zariadení počítača. Mnoho y týchto riadiacich znakov sa už v súčasnosti nepoužíva.

Kódovanie ASCII sa stalo základom pre rôzne národné kódovania. Na dodefinovanie národných znakov sa využíval ďalší ôsmy bit, čím sa využil celý bajt, ktorý je najmenšou adresovateľnou jednotkou v pamäti počítača. Prvých 128 znakov sa zachovalo z kódovania ASCII a ďalších 128 sa využilo na ďalšie znaky.

V našej oblasti sa využívalo kódovanie Kamenických, kódovanie ISO-8859-2 a neskôr kódovanie Windows -1250 (označované niekedy ako CP1250 code page 1250 – kódová stránka 1250). Každé z týchto kódovaní obsahovalo všetky slovenské znaky, ale nie vždy s rovnakými kódmi. Nasledujúca tabuľka je kódová stránka 1250.

	0	16	32	48	64	80	96	112	128	144	160	176	192	208	224	240
0				**0**	**@**	**P**	**`**	**p**	**€**			**°**	**Ŕ**	**Đ**	**ŕ**	**đ**
1			**!**	**1**	**A**	**Q**	**a**	**q**		**’**	**ˇ**	**±**	**Á**	**Ń**	**á**	**ń**
2			**"**	**2**	**B**	**R**	**b**	**r**	**‚**	**“**	**˘**	**˛**	**Â**	**Ň**	**â**	**ň**
3			**#**	**3**	**C**	**S**	**c**	**s**	**ƒ**	**”**	**Ł**	**ł**	**Ă**	**Ó**	**ă**	**ó**
4			**$**	**4**	**D**	**T**	**d**	**t**	**„**	**•**	**¤**	**´**	**Ä**	**Ô**	**ä**	**ô**
5			**%**	**5**	**E**	**U**	**e**	**u**	**…**	**–**	**Ą**	**µ**	**Ĺ**	**Ő**	**Í**	**ő**
6			**&**	**6**	**F**	**V**	**f**	**v**	**†**	**—**	**¦**	**¶**	**Ć**	**Ö**	**ć**	**ö**
7			**'**	**7**	**G**	**W**	**g**	**w**	**‡**	**—**	**§**	**•**	**Ç**	**×**	**ç**	**÷**
8			**(**	**8**	**H**	**X**	**h**	**x**	**ˆ**	**˜**	**¨**	**¸**	**Č**	**Ř**	**č**	**ř**
9			**)**	**9**	**I**	**Y**	**i**	**y**	**‰**	**™**	**©**	**ą**	**É**	**Ů**	**é**	**ů**
10			*****	**:**	**J**	**Z**	**j**	**z**	**Š**	**š**	**Ş**	**ş**	**Ę**	**Ú**	**ę**	**ú**
11			**+**	**;**	**K**	**[**	**k**	**{**	**‹**	**›**	**«**	**»**	**Ë**	**Ű**	**ë**	**ű**
12			**,**	**<**	**L**	****	**l**	**|**	**Ś**	**ś**	**¬**	**Ľ**	**Ě**	**Ü**	**ě**	**ü**
13			**-**	**=**	**M**	**]**	**m**	**}**	**Ť**	**ť**	**-**	**˝**	**Í**	**Ý**	**í**	**ý**
14			**.**	**>**	**N**	**^**	**n**	**~**	**Ž**	**ž**	**®**	**ľ**	**Î**	**Ţ**	**î**	**ţ**
15			**/**	**?**	**O**	**_**	**o**		**Ź**	**ź**	**Ż**	**ż**	**Ď**	**ß**	**ď**	**˙**

V tejto tabuľke čítame takto: Predstavme si, že hľadáme napr. kód pre znak Š. Tento znak nájdeme v stĺpci s nadpisom 128, v riadku 10. Keď tieto dve čísla spočítame , dostaneme kód znaku Š, čiže číslo 128, ktoré prevedieme do dvojkovej sústavy, čo je 10001010.

Opačne by sme postupovali takto: máme napr. znak, ktorého kód je 01001101, rozdelíme ho na dve štvorbitové kúsky teda na 0100 a 1101. Prvú štvoricu prevedieme do desiatkovej sústavy a vynásobíme číslom 16 čo je 4x16=64 – číslo stĺpca, druhé číslo len prevedieme do desiatkovej sústavy čo je 13 a to je číslo riadku, takže je to písmeno M.

Úloha: Nasledujúci kód je zapísaný v šestnástkovej sústave 4A 61 72 6F 20 46 69 6C 69 70. Pomocou kódovej tabuľky nájdite text, ktorému zodpovedá.

Rôznorodosť národných kódovaní dodnes spôsobuje mnohé problémy pri výmene textových informácií napr. pri rôznom nastavení kódovania mailových klientov, webových stránok... Ďalšou nevýhodou týchto osembitových kódovaní je nemožnosť ich použitia pre východoázijské krajiny s veľkým počtom znakov v abecedách.

Nahradiť pluralitu kódovaní jediným kódovaním má za cieľ univerzálne kódovanie **UTF-8** (Unicode Transformation Format). Pomocou tohto kódovania je možné reprezentovať ľubovoľný znak ľubovoľnej abecedy. Kódovanie UTF-8 má premenlivú dĺžku kódu, čo znamená, že jednotlivé znaky sú reprezentované 8 až 32 bitmi (1-4 bajty).

Prvých 128 znakov kódovania UTF-8 sa zhoduje s kódmi kódovania ASCII (podobne ako všetky spomínané národné kódovania). Ak je prvý bit 1, znamená to, že znak má kód dlhší ako jeden bajt. Všetky znaky našej abecedy majú kód dĺžky 1 alebo 2 bajty. Do kódov dvojbajtovej dĺžky sa dostali aj všetky znaky azbuky, gréckej abecedy či arabskej abecedy. Trojbajtové kódy majú hlavne východoázijské znaky. Štvorbajtové kódy sú určené na rôzne historické znakové sady (klinové písmo, hieroglyfy ...)

Keďže slovenské texty obsahujú relatívne málo národných znakov (tie sú reprezentované 2 bajtmi), priemerná dĺžka kódu znaku v bežnom slovenskom texte s kódovaním UTF-8 je asi 1,1bajtu.

Okrem kódovania UTF-8 sa používa najmä vo východoázijských krajinách kódovanie UTF-16, ktoré kóduje znaky buď 2 alebo 4 bajtmi.(kódovanie UTF -16 má dokonca dva varianty Little Endian, Big Endian, ktoré určujú ktorý bajt z dvojajtového resp. štvorbajtového kódu znaku bude prvý. Táto nekompatibilita pramení z rôznej reprezentácie viacbajtových čísel v rôznych typoch procesorov. Bežné počítače používajú Little Endian – ktoré sa niekedy nazýva aj Unicode).

Pri kódovaní UTF-8 súbory môžu začínať špeciálnou nepovinnou trojbajtovou BOM (Byte Order Mark) sekvenciou. BOM sekvenciu pridáva na začiatok súboru pri konverzii do UTF-8 napríklad program NotePad.

Úloha: Napíšte text „PREŠOV“ v Poznámkovom bloku a uložte ho v rôznych kódovaniach (ANSI, Unicode, UTF-8). Porovnajte veľkosť súborov a zdôvodnite rozdiely.

Riešenie: Každé kódovanie používa na zakódovanie znaku iný počet bitov (ANSI: 8 bitov, Unicode: 16 bitov, UTF-8: 8, 16, 24, 32 bitov). Kódovania Unicode a UTF-8 používajú na začiatku súboru špeciálny znak (označenie poradia bajtov), na základe ktorého aplikácia vie v akom kódovaní je dokument uložený.

V tabuľke sú uvedené šestnáskové kódy znakov, ktoré sú uložené v súboroch.

	začiatok súboru	P	R	E	Š	O	V
ANSI		50	52	45	8A	4F	56
Unicode	FF FE	50 00	52 00	45 00	60 01	4F 00	56 00
UTF-8	EF BB BF	50	52	45	C5 A0	4F	56

Spracovanie textu v počítači

Na spracovanie textu pomocou počítača slúžia špeciálne aplikácie – textové procesory. Tieto aplikácie umožňujú vytvárať textové dokumenty, ukladať ich na disk, vkladať do nich rôzne objekty, tlačiť na tlačiarni a mnoho ďalších vecí. Patrí k nim napr. MS Word.

Formátovať text znamená meniť vlastnosti písma a vlastnosti odseku. Okrem nastavovania vlastnosti písma a odseku môžeme dokument formátovať vkladaním nových riadkov a strán a vkladaním špeciálnych znakov, ktoré majú vplyv na vytváranie rozumných medzier v texte. Skôr ako začneme vytvárať dokument je rozumné nastaviť formát stany, čo má vplyv na celkový vzhľad dokumentu.

Písanie textov – všeobecné pravidlá

- Pred interpunkčným znamienkom (. , ! ?) nedávame medzeru.
- Za interpunkčným znamienkom dávame medzeru.
- Okrem špeciálnych prípadov sa vystríhame použitiu viacerých medzier, nových riadkov (Enterov) za sebou.
- Rozlišujeme medzi spojovníkom (–) a pomlčkou (—)

Pravidlá pre písmo

- V jednom dokumente používame maximálne dva druhy písma, ktoré spolu ladia.
- Ak použijeme v bežnom texte pätkové písmo, text je lepšie čitateľný.
- Bezpätkové písmo je vhodnejšie použiť pri tvorbe plagátov a prezentácií, keď chceme upútať pozornosť.
- Na zvýraznenie v texte používame kurzívu alebo tučné písmo. Podčiarknutie nie je vhodné, pretože čiara narúša písmená s dolnými ťahmi (<u>yjp</u>)

Pravidlá pre odsek

- Ak použijeme zarovnanie do bloku, treba mať zapnuté automatické delenie slov, aby rôzna veľkosť medzier nenarúšala čitateľnosť textu.
- V nadpisoch slová nedelíme
- Na konci riadka nenechávame predložky, spojky sa tolerujú.
- Odseky by mali byť oddelené odsadením prvého riadka alebo medzerou.
- Posledný riadok odseku by nemal byť prvým riadkom strany.
- Prvý riadok odseku by nemal byť posledným riadkom strany.

Pravidlá pre formátovanie strany úradného dokumentu

- Na veľkosť okrajov papiera A4 existuje v obchodnom styku norma: horný—2,7 cm, dolný—2,5 cm, ľavý—2,5 cm, pravý—2,5 cm.

Špeciálne znaky pri formátovaní

Textový procesor má k dispozícii špeciálne znaky na označenie časti textu, ktoré je potrebné formátovať upraveným spôsobom, napr. nerozdeliť skupinu slov na konci riadku (slovo s predložkou), vložiť pomlčku namiesto spojovníka (ak neslúži na spájanie viacerých slov) apod.

Ak treba použiť takéto označenie, možno tak urobiť prostredníctvom dialógového okna **Symbol**, ktoré je prístupné z ponuky **Vložiť → Symbol**.

Na záložke **Špeciálne znaky** dialógového okna **Symbol** je zoznam špeciálnych znakov, ktoré možno v texte použiť. Medzi najpoužívanejšie patria:

- **Pevná medzera** – zabezpečí, aby sa nerozdeľovala dvojica slov na konci riadu (napr. jednopísmenková predložka a nasledujúce slovo).
- **Pomlčka** – je dlhšia ako spojovník. V texte ju väčšinou používame namiesto menných slovies (staroba — choroba), pri oddeľovaní vsuviek (MS Word — textový editor), na označenie časového rozpätia (13. 6. — 4. 7. 2009). Pomlčku oddeľujeme od slov medzerami.
- **Pevný spojovník** — zabezpečí, aby sa nerozdeľovalo na konci riadka slovo obsahujúce spojovník (napr. Rakúsko-Uhorsko).
- **Voliteľný spojovník** — umožní vyznačiť v slove miesta, kde je možné slovo roz–de–liť. Je vhodné ho používať, ak odsek zarovnávame podľa okraja.

Písmo

Názvy kapitol, podkapitol sa líšia od bežného textu **veľkosťou** a **typom —** konkrétne písmo. Každý dokument začneme písať nejakým predvoleným typom písma Calibri s veľkosťou 11 pt (point — bod).

Rez písma — modifikácia písma (normálne, šikmé, tučné, podčiarknuté, kapitálky,... a ich kombinácie).

Druh písma — trieda písma so spoločnou vlastnosťou ((bez)pätkové, (ne)proporcionálne, ozdobné, písané, obrázkové...) Medzi nainštalovanými písmami sa nachádzajú pätkové a bezpätkové druhy písma. **Pätkové druhy** (napr. Times New Roman, Courier New,...) majú písmená ukončené kolmými čiarkami — pätkami. Sú dobre čitateľné, preto sa používajú na písanie bežného textu. **Bezpätkové** písma (napr. Calibri, Arial,...) nemajú pätky sú vhodné na písanie nadpisov a podnadpisov. Písma môžeme rozdeliť aj podľa proporcionality. **Neproporcionálne** písma (napr. Courier New) boli doménou písacích strojov, kde mal každý znak rovnakú šírku — aj najširšie písmeno m dostalo ten istý priestor ako najužšie písmeno i . **Proporcionálne** písmo (napr. Arial, Times New Roman, Calibri) naopak zachováva prirodzenú šírku písmen a znakov.

Efekty písma — spôsob zobrazenia písma (horný index, dolný index, prečiarknutie,...), výber efektov závisí od aplikácie.

Font

- Dátový súbor definujúci kompletnú sadu znakov jednotnej veľkosti a štýlu.
- Písmo konkrétneho typu, rezu, veľkosti.
- Ucelená znaková sada.

Odsek

Odsek je základným stavebným kameňom textových dokumentov. Odsek môžeme definovať aj ako text ukončený klávesom Enter. Na koniec odseku sa vloží netlačiteľný znak, nazývaný tiež „tvrdý koniec“ riadku, ktorý nie je pri normálnom zobrazení textu viditeľný. Pre lepšiu orientáciu a kontrolu formátovania môžeme zapnúť zobrazovanie netlačiteľných znakov. Okrem znaku pre koniec odseku sa v spomínanom režime zobrazia medzery, pevné medzery, tabulátory, zlomy strán, stĺpcov a iné.

Zoznamy

Zoznamy sa používajú na vizuálne zoskupenie položiek, ktoré logicky súvisia.

Zoznamy môžu byť:

- **Odrážkové** — používajú sa na nečíslované zoznamy údajov. Celý text sa posunie doprava a pred začiatok každého odseku sa pridá grafická značka.
- **Číslované** — podobné ako odrážkové zoznamy, ale jednotlivé odseky sú číslované (arabskými alebo rímskymi číslami, prípadne písmenami). Práca so zoznamom je rovnaká ako pri odrážkových zoznamoch, len pri zrušení/pridaní niektorého odseku sa čísla nasledujúcich odsekov automaticky upravia.

Jednotlivé zoznamy možno podľa potreby kombinovať do viacúrovňových zoznamov.

Používanie štýlov

Štýl je pomenovaná množina atribútov formátovania, ktorú môžeme spoločne aplikovať na úseky textu v dokumente. Aplikácia štýlov má viaceré výhody:

- Priradenia celého súboru atribútov formátu v štýle je rýchlejšie ako ich postupné nastavovanie.
- Štýly napomáhajú dodržovať jednotný formát dokumentu – priradiť napr. všetkým nadpisom v dokumente zabudovaný štýl **Nadpis 1** je podstatne jednoduchšie ako si pamätať všetky parametre nadpisu a definovať ich zakaždým zvlášť.
- Ak použijeme štýly pre nadpisy na všetky nadpisy v dokumente, textový procesor potom dokáže automaticky vytvoriť obsah dokumentu.

Hlavička a päta

Množstvo dokumentov má na každom liste papiera hore alebo dole rovnaký (podobný) text — napr. údaje o autorovi, dokumente, čísla strán a podobne. Na tento účel nám slúži hlavička a päta dokumentu. Hlavička a päta sa zadáva tak, že sa prepneme do režimu zobrazenia hlavičky a päty.

Hromadná korešpondencia

MS Word umožňuje praktické spojenie **zdroja údajov** (napr. adries vytvorených v textovom procesore alebo nejakom inom tabuľkovom alebo databázovom programe) a **hlavného dokumentu,** v ktorom sú nahrádzané premenlivé položky konkrétnymi údajmi zo zdroja. Takto môžeme tlačiť obálky, menovky, posielať to isté oznámenie mnohým osobám s iným oslovením a pod.