DUM č. 5 v sadě

34. Inf-10 Praktická typografie s LO Writer/MS Word

Autor: Lukáš Rýdlo

Datum: 30.01.2014

Ročník: 4AV, 4AF

Anotace DUMu: Nahrazování textu a opravy pomocí regulárních výrazů v LibreOfficec Writer (s odkazy i pro MS Word).

evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Nahrazování

Teorie

V delších nebo kopírovaných textech se často objevují stejné chyby ve větší míře. Kupříkladu jsme zkopírovali text ve kterém chybí pevné mezery u předložek nebo autor používá chybné „by jsme“, špatné uvozovky, mezery po otevírací a před uzavírací závorkou.

Je velmi nepohodlné v dlouhém textu vyhledávat a ručně opravovat tytéž chyby, tím spíš, když přeci pracujeme s počítačem, který nám má práci ulehčit a ne přidávat. Řešením je hromadné nahrazování chyb.

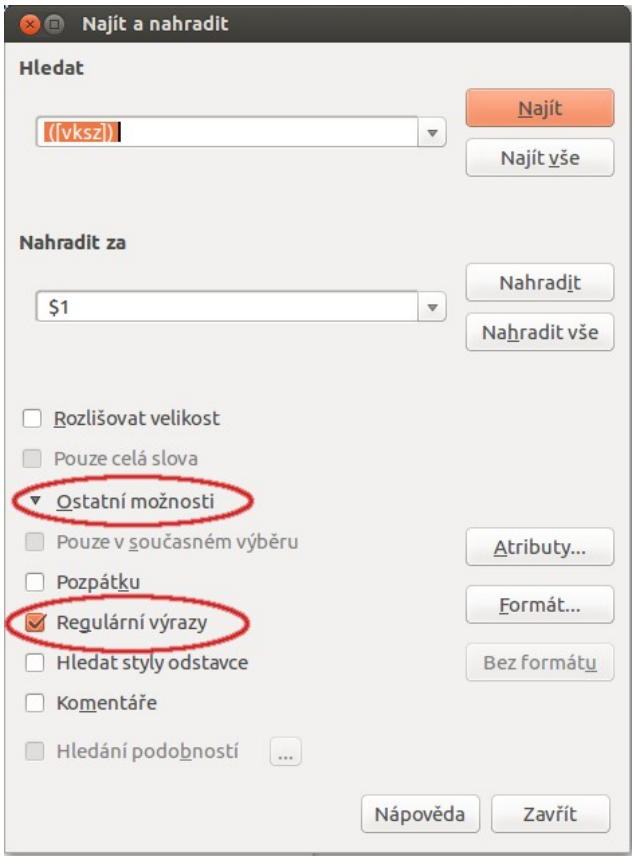
Nahradit konkrétní chybné tvary (zmíněné „aby jsme“ za „abychom“) je jednoduché, ale jak postupovat v případě, že chceme nahrazovat chybu, která se vyskytuje u různých znaků (mezera za předložkami v, k, s, z)? Řešením jsou *regulární výrazy*.

Regulární výraz je výraz složený z běžných (hledaných) znaků a znaků nebo skupin znaků, které popisují nějakou větší skupinu znaků nebo požadovanou vlastnost. Umožňuje tedy pomocí jediného výrazu popsat celou skupinu slov nebo vět.

Je nutné znát význam speciálních symbolů a sekvencí, pomocí kterých popisujeme, co má být na jejich místě v konstruovaném výrazu. Pokud například víme, že znak tečka nahrazuje jeden libovolný znak, pak regulární výraz „`psal.`“ popisuje slova: psala, psali, psaly, psalo, ale i jakékoliv jiné slovo, kde za „psal“ následuje ještě nějaký jeden znak – např. psalX, psal3 nebo i samotné psal. či psal – ale s mezerou za slovem. Pokud by za slovem psal už nenásledovala mezera, pak regulárnímu výrazu neodpovídá. Pomocí regulárních výrazů lze vyhledávat, ale i nahrazovat, což je nejužitečnější využití.

LibreOffice Writer

Jako ve většině programů vyvoláme okno nahrazování (a hledání) zkratkou Ctrl+H (taktéž v menu Úpravy→Najít a nahradit…). Aby fungovalo nahrazování regulárních výrazů, je nutné nejprve kliknout na „Ostatní možnosti“ a pak zkontrolovat, že jsou zatržené „Regulární výrazy“.



Do políčka „Hledat“ napíšeme regulární výraz, který popisuje slova, která chceme najít. Do políčka „Nahradit za“ pak znaky, které mají nalezený text nahradit. Výhodou je, že pokud jsme v regulárním výrazu použili závorky (obyčejné kulaté), pak znaky, které jsou uvnitř (i pokud byly zastoupeny nějakým speciálním symbolem), si program pamatuje a v poli „Nahradit za“ je doplníme pomocí sekvence „$1“, kde číslo (zde jednička) označuje, o kolikáté závorky se jedná.

Chceme-li nahrazovat za nějaký symbol (např. pevnou mezeru), je nejvýhodnější tento symbol vepsat do textu a pak pomocí Ctrl+C a Ctrl+V vložit do políčka „Hledat“ nebo „Nahradit“.

Možnosti nahrazování jsou poměrně široké, detailní seznam všech sekvencí je k dispozici v nápovědě na adrese https://help.libreoffice.org/Common/List_of_Regular_Expressions/cs. My se nyní omezíme na několik příkladů, které snad vysvětlí to nejdůležitější. V příkladech namísto mezery, která psali psaly by nebyla vidět, používám symbol ␣ – místo něj se do políček píše normální mezera.

Hledat	Nahradit za	Význam
␣+	␣	Nahradit libovolný počet mezer za jednu jedinou. (Plus říká, že předchozí znak se může opakovat libovolně-krát, ale aspoň jednou.)
psal([iy])	spal$1	Nahradí všechna slova „psali“ za „spali“ a „psaly“ za „spaly“. Hranat https://help.libreoffice.org/Common/List_of_Regular_Expressions/cs á závorka označuje jeden libovolný symbol ze seznamu (takže [iy] znamená jeden znak – buď i nebo y) a kulaté závorky tento jeden nalezený symbol uloží jako $1 pro nahrazení.
12,[1-5]*	12,0	Nahradí za číslo 12,0 všechny výskyty čísel 12,1; 12,2; 12,3; 12,4; 12,5, ale i čísel, která začínají „12,“ a pokračují libovolně dlouhou řadou z číslic 1 až 5 – např. 12,11225 (hranaté závorky slouží pro výběr jednoho znaku z nabízených, ale spojovník mezi znaky v závorce udává rozsah, takže třeba a-d je libovolné malé písmeno od a po d včetně). Hvězdička udává, že předchozí znak se může opakovat libovolně-krát, ale také vůbec. Proto bude nahrazená i sekvence „12,“ za 12,0. Také je potřeba si pamatovat, že hledání neprobíhá po slovech, takže číslo 5432**12,11**83 bude nahrazené za 5432**12,0**83 a číslo 32**12,**882 za 32**12,0**882, jelikož tučně zvýrazněné části si odpovídají.
␣10,[0-9]+␣	␣deset␣	Místo čísla 10 s libovolně dlouhým desetinným rozvojem za desetinnou čárkou se napíše slovy „deset“. Tentokrát se vzor neaplikuje třeba na číslo 110,2, protože v hledaném vzoru musí být před 10 mezera a za desetinným rozvojem také. Proto se například neaplikuje ani na „10,12.“, protože tečka za dvojkou není mezera, kterou vzor požaduje.
(pod|vý|zá)chod		Odstraní z textu slova „podchod“, „východ“ a „záchod“. Svislítko slouží k oddělení různých variant. Opět je ale možné, že za „chod“ pokračují další písmena, takže třeba slovo „podchody“ se nahradí za „y“, jelikož „podchod“ se nahradí za nic a „y“ zůstane.
^([1-9]).␣	($1)\t	Pokud je na začátku nějakého řádku (symbol ^ značí začátek řádku) číslice 1 až 9 následovaná tečkou a mezerou (např. „1. “), pak se nahradí stejným číslem v závorce a tabulátorem (takže místo „1. “ bude na začátku řádku „(1)→“, kde šipka označuje tabulátor). Je dobré si všimnout, že v poli „Nahradit“ mají závorky význam opravdu jen závorek a ne speciálního znaku, ale sekvence „\t“ si význam „tabulátor“ podržela. Podobně i tečka v poli „Nahradit za“ nebude znamenat libovolný znak, ale tečku.
\([A-Z]\.[a-zA-Z]\)	–kód–	Sekvence závorka (jelikož je před závorkou zpětné lomítko,

		považuje se za obyčejný znak a nemá speciální význam), jedno velké písmeno, tečka (i před tečkou je zpětné lomítko a proto zůstává obyčejným znakem a nemá speciální význam), jedno malé nebo velké písmeno a zavírací závorka se nahradí sekvencí „–kód–“. Nahrazovat se tedy bude například „(A.b)“ nebo „(Z.W)“.

MS Word

V MS Word lze také vyhledávat pomocí regulárních výrazů. Nabídka speciálních znaků a sekvencí se od LO Writeru značně liší a navíc (až na pár výjimek) ani vzdáleně neodpovídá zvyklostem, které se používají v regulárních výrazech programovacího jazyka Perl, což je dnes de facto standard pro regulární jazyky. Je tedy na zvážení každého jednotlivce, zda se chce jejich význam a syntaxi učit, jelikož ji (na rozdíl od standardizovaných sekvencí, které používá i LO Writer) nikde jinde než v aplikaci Word nevyužije. Naopak jejich znalost bude značně matoucí v systémech založených na regulárních jazycích Perlu.

Přehled symbolů a sekvencí lze nalézt v nápovědě programu nebo na stránce http://office.microsoft.com/cs-cz/word-help/vyhledani-a-nahrazeni-textu-nebo-dalsich-polozek-HA001230392.aspx.

Praxe – úkoly (pro LO Writer)

1. Odstraňte z textu všechny prázdné odstavce jediným nahrazením.
2. Odstraňte z textu všechny odstavce obsahující pouze mezery (libovolný počet) jediným nahrazením.
3. Odstraňte ze začátků řádků všechny mezery a tabulátory v libovolném počtu.
4. Mějme soubor s citáty osobností. Za každým citátem je sekvence „ autor: Jméno Příjmení“. Nahraďte tento zápis v celém souboru sekvencí „ (J. P.)“. Předpokládejte, že jméno může obsahovat znaky s diakritikou, proto místo [a-zA-Z] použijte raději [:alpha:].
5. Nahraďte všechny výskyty neslabičných předložek (v, k, s, z) následovaných obyčejnou mezerou za tytéž předložky a pevnou mezeru.
6. Vložte jako neformátovaný text obsah některé stránky Wikipedie. V textu se nyní objevují hranaté závorky a v nich čísla, která ve Wikipedii odkazují na zdroje. Odstraňte z celého souboru všechny tyto závorky i s čísly uvnitř.
7. Nahraďte v souboru všechen text v uvozovkách za text v závorkách.
8. Váš kamarád napsal v podkladech k referátu několikrát slovo „moc“ a „hodně“ s několika „o“ a pokaždé s jiným počtem (hooodně, mooooooc, mooc, hooooooodně, …). Jediným nahrazením nahraďte tato slova za slovo „mnoho“.

Praxe – řešení (pro LO Writer)

1. Nahrazujeme ^$ za nic.
2. Nahrazujeme ^ +$ (pak v odstavci musí bý alespoň jedna mezera) nebo ^ *$ (odstraní i odstavce prázdné – bez mezer) za nic.
3. Nahradíme ^[\t]+ za nic.
4. Nahradíme autor: ([:alpha:])[:alpha:]* ([:alpha:])[:alpha:]* za ($1. $2.) – případně můžeme místo hvězdiček použít plus a za vzor doplnit $, pokud má být jméno na konci řádku.
5. Nahradíme ([.])([vkszVKSZ]) za $1$2 , kde mezera v „Nahradit za“ je pevná (vykopírovaná z textu). Jednodušší varianta nádhrady ([vksz]) za $1 , bude nahrazovat jen předložky uvnitř vět, nikoli na začátku vět.
6. Čísla v hranatých závorkách odstraníme nahrazením \[[0-9]+\] nebo \[[:digit:]+\] za nic. Zpětná lomítka u okrajových závorek jsou nezbytně nutná, aby se považovaly za obyčejné závorky a ne speciální symbol.
7. Nahradíme „(([:alnum:]|[,.:?;])+)“ za ($1).
8. Nahradíme (m|h)o+(c|dně) za mnoho. To není spolehlivé, protože tomu odpovídá i slovo „hooc“ nebo „moooodně“, která se ale asi vyskytovat nebudou. Také bychom mohli provádět dvě nahrazení: nějprve „ho+dně“ za „hodně“ a pak „mo+c“ za „moc“.

Zdroje

Veškeré texty i obrázky jsou původní prací autora. Jako podklady byly využity stránky https://help.libreoffice.org/Common/List_of_Regular_Expressions/cs, http://office.microsoft.com/cs-cz/word-help/vyhledani-a-nahrazeni-textu-nebo-dalsich-polozek-HA001230392.aspx a http://cs.wikipedia.org/wiki/Regul%C3%A1rn%C3%AD_v%C3%BDraz.

Pro doplnění doporučuji přečíst http://www.openoffice.cz/navody/jak-vyhledavat-a-nahrazovat-text a pro rozšíření http://www.regularnivyrazy.info/regularni-vyrazy-zaklady.html.