

adattípus-kezelés text fájl importálásakor

Az [adattípus-kezelés I.](#) című cikkben az Excel-táblázat importálásakor alkalmazott adattípus-kezelő automatizmusokat ismertettem, most pedig a formázást nem tartalmazó szövegfájl importálásához előforduló adattípus-kezelő műveleteket mutatom be.

A táblázatos adatkezelő programok leggyakoribb adatcsere-formátuma a formázást nem tartalmazó szövegfájl. Kiterjesztése „txt”. A táblázatban egy sorban álló adatok a szövegfájlban is egy sort képeznek. A szövegfájl minden sora azonos számú adatot tartalmaz, meghatározott karakterrel szeparálva, vagy szóközökkel, valóságos oszlopokat képezve. A szövegfájl, első sorában, tartalmazhatja az oszlopneveket is. A bővítmény csak az elkülönítő-karakteres állományok importálására képes, a szóközökkel oszlopokba rendezett típust nem tudja kezelni.

A táblakészítést megelőző eljárás azonos a már megismerttel: a bővítmény a forrás-oszlop vizsgálata alapján megállapítja a létrehozandó mező adattípusát, majd az ettől a típustól eltérő adatokat a beolvasás előtt, konvertálással, csonkolással, helyettesítéssel és szelektálással egységesíti. Az elkészült tábla minden mezőjében már azonos adattípusú bejegyzések állnak.

Az adattípus megállapításánál a bővítmény az oszlop első huszonöt sorát elemzi. Megkülönböztet szöveget, egész és tört számot valamint dátumot. A mező adattípusa az első huszonöt sorban leggyakrabban előforduló típus lesz. Ha két adattípus egyenlő számban fordul elő, akkor az adattípusok rangsora dönt: szám > dátum > szöveg. Az egyes sorokban hiányzó adatok az adattípus-meghatározást nem befolyásolják. Az első huszonöt helyen adatot nem tartalmazó oszlop adattípusát a bővítmény szöveg típusúra állítja be. Ha egy oszlop első huszonöt sorában többségben számbejegyzések állnak és ezek közül akár csak egy is tört szám, akkor a mező adattípusa tizedes tört szám lesz. Ha az első huszonöt sor számai mind tört rész nélküliek, akkor a bővítmény a létrehozandó mezőnek egész szám adattípust határoz meg.

Fájl	Szerkesztés	Formátum	Nézet	Újgó		
sor száma	oszlop1	oszlop2	oszlop3	oszlop4	oszlop5	
01 sor	144			-145	-145	
02 sor	165			-123	-123	
03 sor	186			-101	-101	
04 sor	207			-79	-79	
05 sor	228			-57	-57	
06 sor				-35	-35	
07 sor				-13	-13	
08 sor				9	9,1	
09 sor				31	31	
10 sor				53	53	
11 sor				75	75	
12 sor				97	97	
13 sor				119	119	
14 sor				141	141	
15 sor				163	163	
16 sor				185	185	
17 sor				207	207	
18 sor				229	229	
19 sor				251	251	
20 sor	meggy			273	273	
21 sor	füge			295	295	
22 sor	alma			317	317	
23 sor	szilva			339	339	
24 sor	körte	144		361	361	
25 sor	eper	eper		383	383	

A képen egy tabulátorokkal tagolt szövegfájl, első huszonhat sorát látjuk. Az első sor tartalmazza az oszlopok neveit. Milyen adattípust határoz meg a bővítmény a leendő mezőknek? Az oszlop1 mező adattípusa szöveg lesz, mert a karakterláncok száma nagyobb, mint a számok darabszáma. A hiányzó adatok az adattípus-meghatározást nem befolyásolják. Az oszlop2-ben azonos darabszámú szám és karakterlánc áll, ezért az adattípusok rangsora dönt. Tehát az oszlop2 mező adattípusa egész szám lesz. Az oszlop3 első huszonöt sora nem tartalmaz adatot, ezért a belőle képzett mező adattípusa szöveg lesz. Az oszlop4 mezőnek a PowerPivot egész szám adattípust fog meghatározni, mivel az első huszonöt szám tört rész nélküli. Az oszlop5 nyolcadik sorában egy tizedes tört áll és ez az egyetlen érték már tizedes tört szám adattípusú mezőt fog eredményezni.

Miután számba vettük az adattípus-meghatározás legfontosabb szabályait, vegyük sorra, milyen adattípusúnak látja a bővítmény a text fájlban álló adatokat. A PowerPivot szöveggént detektálja [1] a karakterláncokat, [2] a logikai értékeket, [3] a szóközzel ezres csoportokba rendezett számokat, [4] a pontot és a vesszőt (ezres- és tizedes elválasztó) egyaránt tartalmazó számokat, [5] a százalékokat, [6] a fel nem ismert pénzeket, [7] a fel nem ismert időpontokat, [8] és a fel nem ismert dátumokat. A szöveg adattípusú bejegyzések a mezőben balra igazítva jelennek majd meg. Lássunk néhány példát!

Fájl	Szerkesztés	Formátum	Nézet	Úgó		
oszlop1	oszlop2	oszlop3	oszlop4	oszlop5	oszlop6	oszlop7
IGAZ	TRUE	123 456	123,156.7	12%	1 HUF	1ó 23p
HAMIS	FALSE	1 234 567	1.234.567,89	12,3 %	2 EUR	2018. 06. 07.
HAMIS	TRUE	12 345	12,345.678	45.6%	3 USD	2018. június 7., csütörtök

Néhány elgondolkodtató részletet is megfigyelhetünk a képen. A bővítmény text fájlban nem ismeri fel a program magyar pénznem formátumát (Excel 2016), sőt a vezérlőpult szerinti (Windows 10), magyar rövid- és hosszú dátum formátumát sem.

A bővítmény egész- vagy tizedes tört számként detektálja [1] a csak számokat tartalmazó adatokat, illetve a kötőjellel (negatív előjel) kezdődő számokat, [2] az egy pontot vagy egy vesszőt (tizedes elválasztó) tartalmazó számokat, [3] a normál alakú számokat és [4] a felismert pénzeket.

Fájl	Szerkesztés	Formátum	Nézet	Úgó
oszlop1	oszlop2	oszlop3	oszlop4	
-123	45.123	1,23E+02	23 \$	
33	-45.123	-3.4e+02	-45Ft	
-45	45,123	2.345e-02	6,78 €	

Ezen a képen is találunk érdekességet. A negyedik oszlopban látjuk, hogy a számunkra fontos, pénznem jelölőket (Ft, €, \$) felismeri a PowerPivot, akkor is, ha közvetlenül a szám után állnak és akkor is, ha egy szóköz választja el őket a számtól. Sőt, a bővítmény még a szám előtt állva is, szóközzel vagy anélkül, felismeri őket és tizedes tört szám adattípust állít be számukra.

A PowerPivot dátumként detektálja: [1] a felismert időpontokat és [2] a felismert dátumokat. A pozitív tizedes törtéket, amelyeknek egész része kisebb huszonnégynél, törtrésze kisebb hatvannál és tizedes elválasztójuk vessző a bővítmény időpontnak tekinti, amelyben a szám egész része az órák számát-, tört része a percek számát adja.

Fájl	Szerkesztés	Formátum
oszlop1	oszlop2	
12,34	2018.06.08	
12:34	2018.8.6	
12,34,56	18-6-8	
12:34:56	8.6.2018	

A képen megfigyelhető, hogy a bővítmény egyaránt felismeri a dátum egységek magyaros (évek, hónapok, napok) és angolos (napok, hónapok, évek) sorrendjét is.

A mező adattípusának meghatározása után, mi történik azokkal az adatokkal, melyek típusa nem azonos a beállítással? Erre a kérdésre ad választ a következő kép.

index	Egész szám	Tizedes tört szám	Dátum	Szöveg
1	5555	55,55	2005.05.05	ötösök
2	5555	55,55	2005.05.05	ötösök
3	5555	55,55	2005.05.05	ötösök
24	5555	55,55	2005.05.05	ötösök
25	5555	55,55	2005.05.05	ötösök
26	55,55	5555	5555	5555
27	2005.05.05	2005.05.05	55,55	55,55
28	ötös	ötös	ötös	2005.05.05

index	egész szám	tizedes tört szám	dátum	szöveg
22	5555	55,55	2005. 05. 05. 0:00:00	ötösök
23	5555	55,55	2005. 05. 05. 0:00:00	ötösök
24	5555	55,55	2005. 05. 05. 0:00:00	ötösök
25	5555	55,55	2005. 05. 05. 0:00:00	ötösök
26	55	5555		5555
27				55,55
28				2005.05.05

Vegyük sorra a kép tanulságait! Az egész szám adattípusú mezőbe a bővítmény a tört számot egészre csonkolva olvassa be, a dátumot és a szöveget nem importálja. A tizedes tört szám adattípusú mezőbe csak egész számot tudunk beolvasztani, dátumot és szöveget nem. A dátum adattípusú mezőbe csak dátum kerülhet. A szöveg adattípusú mezőben a másik három adattípus szöveggé lesz elhelyezve.

A tábla elkészülte után, a csak logikai értékeket tartalmazó szöveg-mezőt és a szám-mezőket állíthatjuk igaz/hamis, illetve pénznem típusúra a Kezdőlap, Formátum, Adattípus utasítással.



Margitfalvi Árpád
margitfalvi.arpad@gmail.com