

4. Szógyakoriság

Ebben a feladatban a magyar nyelv szavainak webes előfordulását vizsgáljuk adatbázis-kezelő segítségével. A *szo10000.txt* szöveges állományban megtalálható egy-egy szó szótöve, szófaja és a vizsgált weboldalakon való előfordulásának száma. A szótár elkészítéséhez 2004-ben gyűjtötték össze az interneten található magyar szövegeket. Ezt a több mint ötszázmillió szóelőfordulást tartalmazó adathalmazt, használták fel a szótár összeállításához. A feladatban szereplő forrás nem teljes, mert abban csak azok a főnevek, melléknévek, határozószók és igék szerepelnek, amelyek előfordulása legalább 10000.

Az állományban azért szerepelnek szótövek, mert egy-egy szó ragozott alakja valójában ugyanazt a szót jelenti. Például a „kell” szótó előfordulása a „kell”, „kellett”, „kellene” stb. szavak előfordulásának összege, így csak egyszer szerepel az állományban. Vannak olyan szótövek, amelyek több szófajhoz tartoznak, például a „fog” szó főnév is és ige is. Ezek természetesen többször fordulnak elő az állományban, például az előbb említett „fog” szó főnévként is és igeiként is.

1. Készítsen új adatbázist *szogyak* néven! A forrásként kapott *szo10000.txt* – tabulátorokkal tagolt, UTF-8 kódolású – szöveges állományt importálja a **szavak** nevű táblába! Az állomány első sora tartalmazza a mezőneveket. A létrehozás során állítsa be a megfelelő típusokat!

Tábla:

szavak (azon, szoto, szofaj, gyakori)

<i>azon</i>	A szó azonosítója (szám), ez a kulcs
<i>szoto</i>	A szó szótöve (szöveg)
<i>szofaj</i>	A szó szófaja (szöveg), lehetséges értékei: fn, mn, ige, hsz (azaz: főnév, melléknév, ige, határozószó)
<i>gyakori</i>	A szótó előfordulásának gyakorisága (szám)

A következő feladatok megoldásánál a lekérdezéseket a zárójelben olvasható néven mentse! Ügyeljen arra, hogy a megoldásban pontosan a kívánt mezők szerepeljenek!

2. Készítsen lekérdezést, amely megadja azoknak az igéknek a szótövét, amelyeknek az előfordulása legalább 500 000! (**2ige500**)
3. Lekérdezéssel adja meg azokat az adatbázisban megtalálható melléknéveket („**mn**”), amelyek szótöve a „**br**” szórészlettel kezdődik! A melléknév szótövét és gyakoriságát jelenítse meg! (**3brmellek**)
4. Készítsen lekérdezést, amely megadja a 10 leggyakoribb szótövet a határozószó („**hsz**”) szófajú szavak közül! (**4hatar10**)
5. Lekérdezés segítségével listázza ki, hogy mely szófajban hány szótó szerepel az adatsorozatban! A szófajok jelölését és a szótövek számát jelenítse meg! (**5szofajok**)
6. Bizonyos szótövek többször is előfordulhatnak az adatbázisban. Ennek az az oka, hogy egy a szótónek különböző jelentései is lehetnek, és ezért eltérő szófajokhoz is tartozhat. Készítsen lekérdezést, amely megadja azokat a szótöveket, amelyek legalább háromszor szerepelnek az adatbázisban! (**6tobb**)

15 pont