Idegen szavak arányának változása az MTI cikkeiben az elmúlt harminc évben (1992-2022)

## Hipotézis

Az elmúlt harminc évben (1992-2022) az MTI híreiben az egyes nyelvekből származó idegen szavak aránya jelentősen változott, az angol eredetű szavak használatának aránya nőtt, a latin, német és görög szavaké csökkent, egyéb nyelveké pl. arab viszonylag változatlan maradt.

## Szükséges adatok

* Idegen szavak listája
  + A <https://idegen-szavak.hu/szavak/abc_sorrendben/-ról>, pythonban írt programmal, web scrapinggel, a requests és a beautifulsoup4 könyvtárakat használva.
    - Kattintások automatizálása, releváns html kinyerése
  + A program egy csv fájlba menti a szavakat, azok eredetét.
* Excelben megtisztítom az idegen szavak listáját

## MTI cikkek feldolgozása

* Adataok kinyerése, feldolgozása
  + A <https://archiv1988-2005.mti.hu-ról> pythonban írt programmal, web scrapinggel, a requests, time, selenium és a beautifulsoup4 könyvtárakat használva.
    - Űrlapkitöltés, lapozás, kattintások automatizásása, releváns html kinyerése
  + Kinyert adatok:
    - Szavak száma egy adott cikkben
    - Összes idegen szó száma az adott cikkben
    - Egyes nyelvekből származó szavak száma a cikkben
* Megoldandó problémák az adatok kinyerésénél
  + Több nyelvből származik a szó egyszerre (pl. ógörög és latin) -> mindkettőhöz beleszámítja az adott szót
  + Nyelvek elnevezésének standardizálása (pl. óskandinávot és óészakit ugyanannak vegye („sicariót”)
  + Ragozott szavak figyelembevétele, akkor is ha a rag megnyújtja az utolsó magánhangzót
  + Több szóból álló idegen nyelvi kifejezések felismerése (pl. „ab ovo”, „ad urbe condita”)

## Adatok megtisztítása és ábrázolása

* Adatok megtisztítása és ábrázolása chartokon excelben