

Adatelemzés a közigazgatásban. Jogi hálózatkutatás.

Dr. Kovács Tamás

NAV Békés Vármegyei Adó-és Vámigazgatósága
igazgató



Nemzeti Adó-
és Vámhivatal

Jogalkalmazási tevékenység átalakulása

1999

- Pp. (1952. évi III. tv.)
elektronikus szó nem szerepel
- Be. (1973. évi I. tv.)
elektronikus szó nem szerepel
- az államigazgatási eljárás általános szabályairól (1957. évi IV. tv.)
elektronikus szó nem szerepel
- a cégnyilvántartásról,
a cégnyilvánosságról és a bírósági
cégeljárásról (1997. évi CXLV. tv.)
elektronikus szó nem szerepel

Az adóhivatalra vonatkozó 3 hatályos eljárási törvény:

- az adóigazgatási rendtartásról (2017. évi CLI. tv.)
- az adózás rendjéről (2017. évi CL. tv.)
- az adóhatóság által fogatosítandó végrehajtási eljárásokról (2017. évi CLIII. tv.)

Több mint 210x szerepel az elektronikus szó!

2024

- Pp. (2016. évi CXXX. tv)
több mint 120x szerepel
- Be. (2017. évi XC. tv.)
több mint 200x szerepel
- az általános közigazgatási
rendtartásról (2016. évi CL. tv.)
több mint 9x szerepel
- a cégnyilvánosságról, a bírósági
cégeljárásról és a végelszámolásról
(2006. évi V. tv.)
több mint 140x szerepel

**NAV digitális ügyfélforgalma 2020. évben 86,14 % volt,
ez 2024. évre 97,61 %-ra nőtt!**

Miben segítheti a közigazgatást az adattudomány és új technológiák?

1. **Hatékonyabb döntéshozatal** – Adataalapú elemzések révén a vezetők pontosabb, megalapozottabb döntéseket hozhatnak (pl. költségvetés-tervezés, erőforrás-elosztás).
2. **Közszolgáltatások optimalizálása** – Az állampolgári igények és használati minták elemzése javítja a szolgáltatások minőségét és célzottságát (pl. egészségügy, közlekedés).
3. **Költségcsökkentés és hatékonyságnövelés** – Felesleges folyamatok azonosítása, automatizálás és jobb folyamatmenedzsment révén csökkennek az állami kiadások.
4. **Átláthatóság és korrupció megelőzése** – Anomáliák, szabálytalanságok felismerése segíti a visszaélések feltárását és növeli a közbizalmat.
5. **Előrejelzés és stratégiai tervezés** – Prediktív modellek alkalmazásával jobban tervezhető a jövő (pl. demográfiai változások, válságkezelés, infrastruktúra-fejlesztés).

Legal technology megjelenése

(elkerülhetetlen a
közigazgatásban az alkalmazás)

- Adatok rendszerezése
- Változások, tendenciák elemzése, következtetés levonása
- Jogi esetek elemzése

A legal tech előnye, hogy olcsó, hatékony, pontosabb eredményt hoz mint egy ember, azonban a kreativitás, pertaktika, empátia, asszociációk, súlyozás, mérlegelés nem jellemzőek, ahogyan az embereknél. (MÉG!)

Jogalkalmazás

A hatósági munka legnehezebb része a mérlegelés - bizonyítás

legnagyobb veszély az **ún. kognitív torzítások**

elfogult véleményből származó
gondolkodási, logikai hibák

- Minden vagy semmi típusú gondolkodás
- Egyik érvrendszer figyelmen kívül hagyása
- Elhamarkodott következtetés
- Felnagyítás vagy lekicsinyítés
- **Megerősítési torzítás** (confirmation bias) - bizonyítékhalmozás egyik oldalra

Veszélyes továbbá

Concorde effektus és a Dunning-Krueger hatás

Az intuíciótól a szenzoros, azaz érzékelő, technológiai úton beszerzett és elemzett objektív adatokra épülő helyzetfelismerés felé...

Pontos objektív tényállás

A gyakorló jogászoknál pl. hosszú ideig jelentős szerepe volt az **intuíciónak, ösztönös megérzésnek**.

Ez azt eredményezte, hogy *fontos ügyekben* téves tényállásból indultak ki, melynek okai:

- Valótlan tartalmú emberi nyilatkozatok
- Torzított tartalmú iratok, bizonyítékok
- Szimpátia/antipátia – megerősítési torzítás
- Időbeli nyomás
- Berögzült magatartásformák...vagy helytelen korábbi döntések, melyek mintául szolgáltak

A mai technológiai eszközökkel sokkal szélesebb körben van lehetőség a **valós tényállás megállapítására és jogértelmezésre**, mely elengedhetetlen a hatékony képviselőhez/hatósági/bíróági munkához.

Az ember befolyásolható, sokszor téves következtetéseket von le az észleltekből!

- A könnyen elérhető vagy sokszor ismételt információkat hajlamosak vagyunk könnyebben elhinni.
- Megerősítési torzítás (a véleményünkhöz, álláspontunkhoz, hiedelmeinkhez közelebb álló információkra figyelünk, a többit figyelmen kívül hagyjuk.)
- Concorde effektus (ragaszkodunk káros módon valamihez)

Hajlamosak vagyunk a tényeket tévesen értelmezni és ezen tényekhez ragaszkodni a ténylegesen megtörtént, észlelt tényekkel szemben.

Hajlamosak vagyunk tévedésben/illúzióban élni, mert **JÓ A JÓT VAGY ÁLTALUNK VALÓSNAK VÉLT DOLGOT HINNI!**

Kutatási terv alapja

- Kutatás címe: A bírósági döntések hivatkozásainak hálózata adóperekben, a hangsúlyos, csomóponti ítéletek kiválasztása.
- Kutatás célja:
 - Jogalkotóknak, jogalkalmazóknak segítségnyújtás a nagy mennyiségű szövegben való eligazodáshoz
 - Legal tech bevezetése adózásban a jogszabályok értelmezésénél
 - Emberi agy korlátainak legyőzése, objektív-rendszerszintű és logikus jogértelmezés fejlesztése
- Várható eredmény:

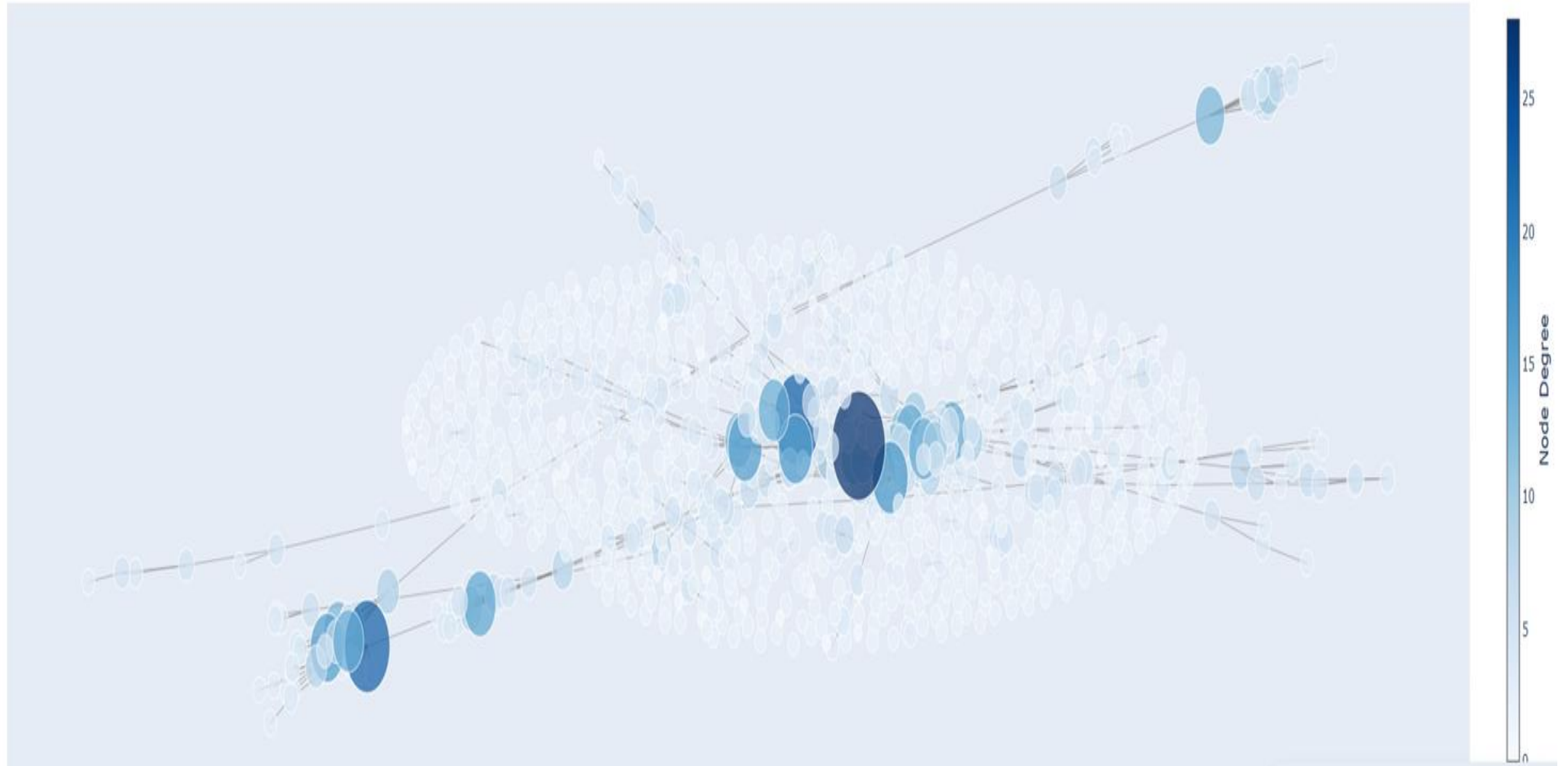
Ezen kutatással gyorsan lehet(ne) eredményt kapni a bírósági ítéletek rendszeréről, jobban átlátjuk az ítéletekben más döntésekre hivatkozások, ítéleti-indokolási súlypontok rendszerét.

Kutatási szakmai tartalma

- Bírósági ítéletek egy jól vizsgálható szöveganyagot, a szöveganyag tartalma pedig hálózatot alkotnak.
- A hálózat kutatás módszerével feltérképezhető az ítéletekben rejlő hivatkozási, kapcsolódási háló.
- A hálózatban az ítéletek a gráfok csúcspontjai, a hivatkozások pedig az élek.
- Korlátozott precedensrendszer súlyponti és komplex ítéleteinek kiválasztása:
 - Súlyponti (amire sok ítélet hivatkozik) ítéletek
 - Komplex (ami sok ítéletre hivatkozik) ítéletek
 - A jogász munka kvalitatív (minőségi- emberek gondolatai, viselkedése, motivációja, nyelvezete), ezen új módszerekkel kvantitatív (mennyiségi – számszerű adatok) módszerekkel vizsgálhatunk.
- A módszer lehetővé teszi az adat alapú keresést, csökkentheti a kognitív torzításokat.

1042 db ítélet (Kúria/Legfelsőbb Bíróság)

Kúriai döntések hivatkozási hálózata



Módszerek

| Fogalom | Mi ez? | Kategória |  |
|---------------------------------|--|------------|---|
| Colab | Böngészőben futó Python környezet, főleg gépi tanuláshoz | Eszköz | |
| PageRank | Csúcsfontosságot mérő algoritmus (Google találta ki) | Gráfmutató | |
| Total Degree | Egy csúcs összes kapcsolata (be- + kimenő élek) | Gráfmutató | |
| In Degree / Out Degree eloszlás | Bejövő és kimenő kapcsolatok eloszlása a hálózatban | Gráfmutató | |
| Betweenness eloszlás | Mutatja, mennyire „átjáró” egy csúcs a hálózatban | Gráfmutató | |
| Closeness eloszlás | Mutatja, mennyire közel van egy csúcs átlagosan a többiekhez | Gráfmutató | |
| Plotly-dropdown | Interaktív menü adatvizualizációkhoz | Eszköz | |
| NLP | Natural Language Processing – természetes nyelv feldolgozása | Terület | |

Page rank eredményei

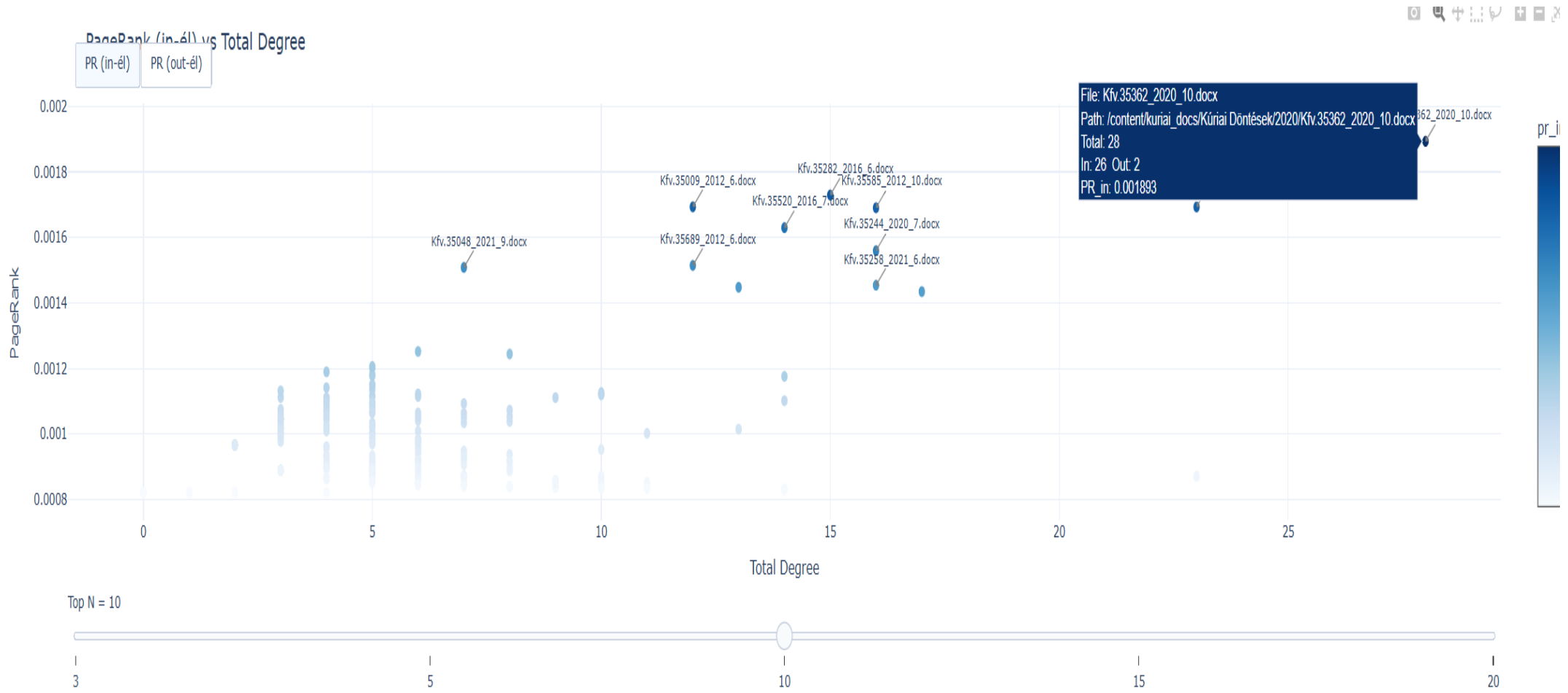
- A Kúria által hozott ítéletek feltérképezése megtörtént, teljes pontossággal meg lehet állapítani, hogy az ítéleteknek hány db bemenő és kimenő fokszáma van. (számszerű hivatkozással)
- A hivatkozási hálók pontosan megmutatják, hogy mik a legtöbbet hivatkozott és mik a legtöbb ítéletre hivatkozó ítéletek. Ezzel a leghangsúlyosabb és legkomplexebb ítéletek beazonosíthatóak.
- Megállapítható, hogy az ítéletekben vannak kisebb csoportok, klaszterek, melyek egy egy részterület hivatkozási hálóját jelentik.
- Egy egy klaszter egy egy problémához vagy időszakhoz köthető (pl: NAV 2010 vagy 2015 évi változásai megnövelik a klaszterekben lévő ítéletek)

- Regex alapú Kfv-kinyerés. Két mintát használunk, római előtagos és egyszerű. A minták szövegszerűen így néznek ki:

Kfv\\.s*[IVXL]+\\.s*\\d+(?:\\.\\d+)?\\s*\\/s*(19|20)\\d{2}\\s*\\/s*\\d+

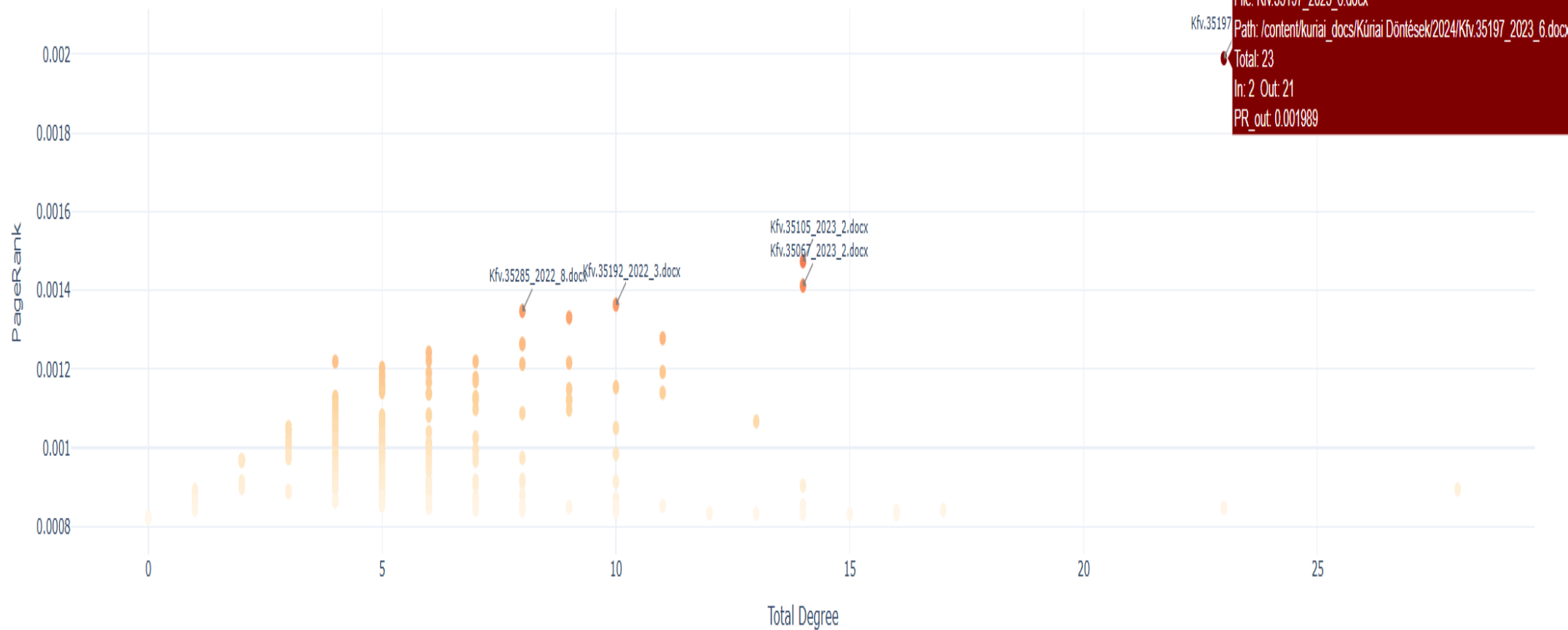
Kfv\\.s*\\d+\\s*\\/s*(19|20)\\d{2}\\s*\\/s*\\d+

- Page Rank csúcsfontosság
- kontra Total Degree összfokszám

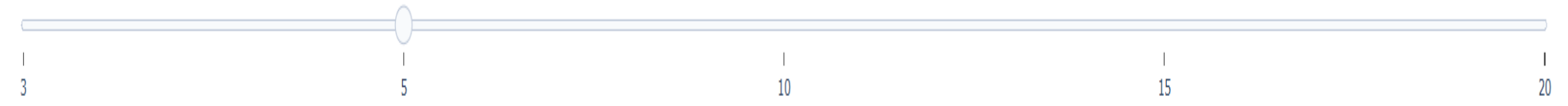


PageRank (out-él) vs Total Degree

PR (in-él) PR (out-él)

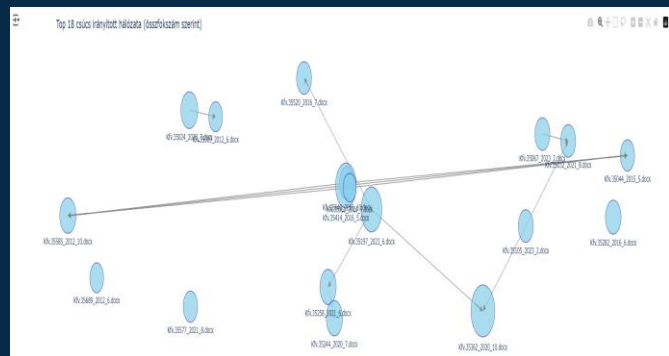
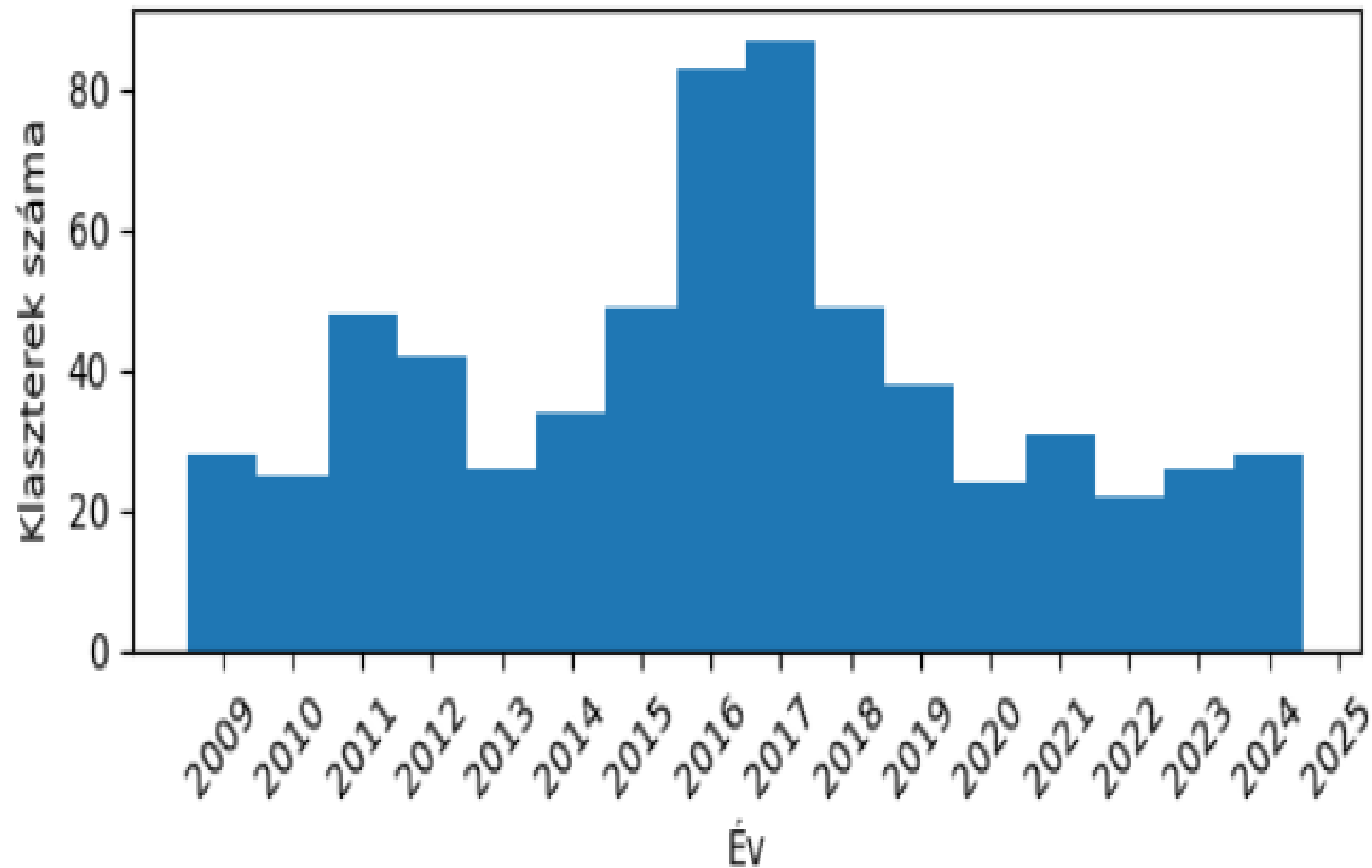


Top N = 5



Klaszterek keletkezése (összefüggő adathalmazok)

Mely évben jöttek létre a kis klaszterek?



Példa egy számítási módra

4.3. TF-IDF hasonlóság (dokumentszint)

Minden dokumentumot vektorizálunk. Szó-alap: 1-2 n -gram; karakter-alap: pl. 5-gram. A tipikus (sublinear) TF-IDF:

$$\text{tf}(t, d) = 1 + \log c(t, d), \quad \text{idf}(t) = \log \frac{N + 1}{\text{df}(t) + 1} + 1, \quad x_{t,d} = \frac{\text{tf}(t, d) \cdot \text{idf}(t)}{\|x_d\|_2}.$$

Két dokumentum pontszáma a koszinusz:

$$\text{sim}(u, v) = \cos(x_u, x_v) = \frac{x_u \cdot x_v}{\|x_u\|_2 \|x_v\|_2}.$$

Kiválasztás: i -hez a $\mathcal{C}(i)$ -beli top $k(i)$ legnagyobb koszinuszértékű j kerül élként. *Megjegyzés:* karakter-TF-IDF jól kezeli a ragozást és az idézet-variánsokat; szó-TF-IDF gyorsabb és tematikus hasonlóságot mér.

1. Mi a cél?

Dokumentumokat (szövegeket) szeretnénk **összehasonlítani**, hogy mennyire hasonlítanak egymásra tartalmilag.

2. Hogyan működik a TF–IDF?

- **TF (term frequency)**: Megnézi, hogy egy szó (vagy karakterdarab, n-gram) hányszor fordul elő az adott dokumentumban. (De nem simán a darabszámot nézi, hanem logaritmussal csökkenti a nagyon gyakori szavak súlyát.)
- **IDF (inverse document frequency)**: Megnézi, hogy egy szó mennyire ritka az egész szövegtörzsben. Ha minden dokumentumban előfordul (pl. „és”, „van”), akkor keveset ér; ha ritka, akkor nagy súlyt kap.
- **TF × IDF**: A kettőt összeszorozzuk → így egy szó nagyobb súlyt kap, ha **gyakori az adott dokumentumban, de ritka az egész gyűjteményben**.

3. Dokumentum mint vektor

- Minden dokumentumot egy **számvektorral** írunk le, ahol minden dimenzió egy szóhoz (vagy n-gramhoz) tartozik.
- Így a szövegeket át tudjuk alakítani számokká → amiket össze lehet hasonlítani.

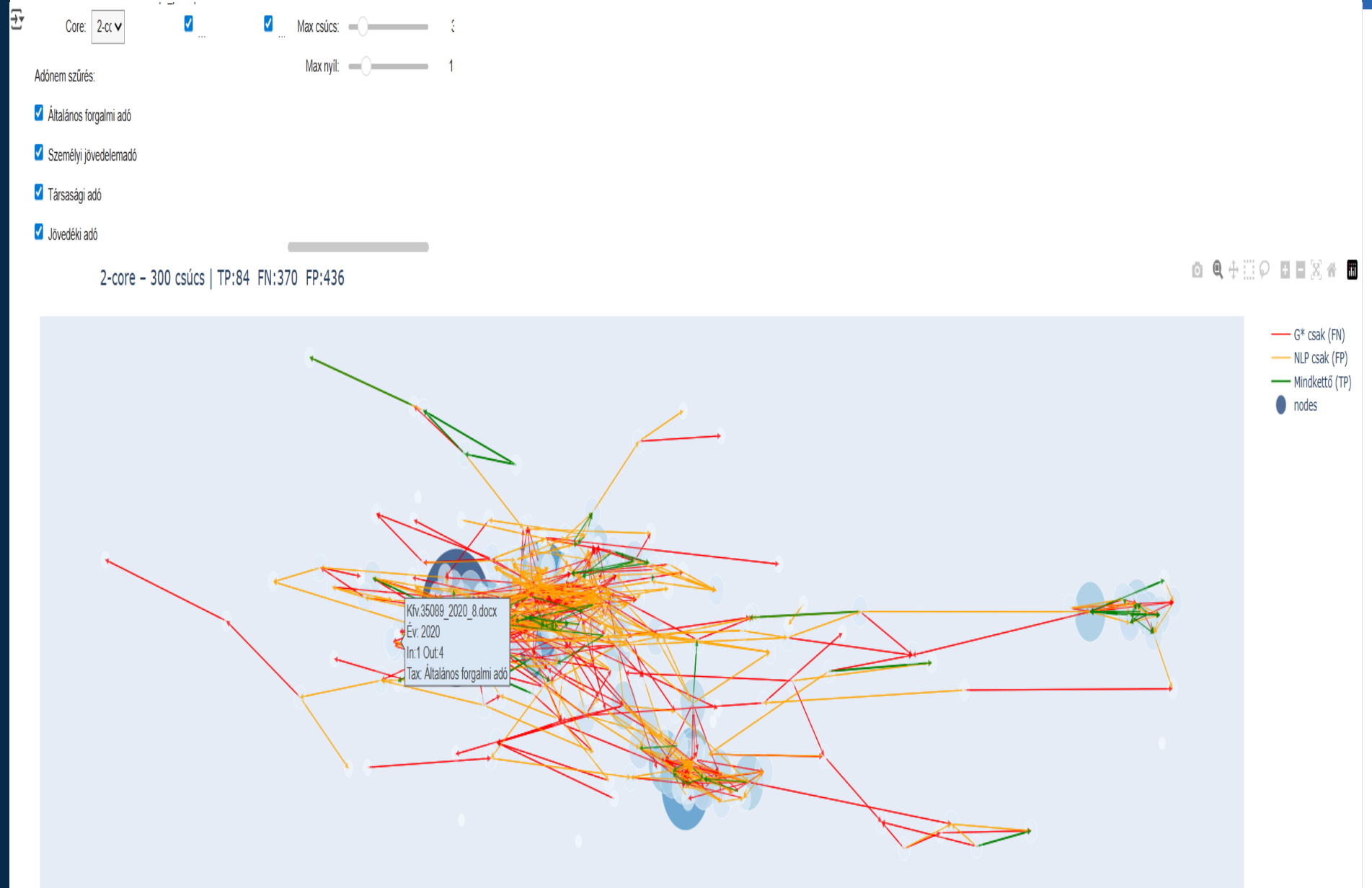
4. Hasonlóság mérése

- Két dokumentum hasonlóságát a **koszinusz hasonlósággal** mérjük:
- $\cos(u,v) = \frac{u \cdot v}{\|u\| \|v\|}$
- Ez azt méri, hogy két vektor (két szöveg) mennyire „egy irányba mutat” → 1-hez közeli érték = hasonló, 0-hoz közeli = teljesen más.

5. Lényeg

- Átalakítjuk a szöveget számmá (vektorokká) a TF–IDF segítségével.
- Összehasonlítjuk őket koszinusz-hasonlósággal.
- Így objektíven meg tudjuk mondani, melyik szöveg mennyire hasonlít a másikra, függetlenül attól, hogy mennyire hosszúak, és figyelembe véve, hogy egyes szavak fontosabbak, mint mások.

Page Rank és NLP módszereinek összevetése



A G* regex gráf és NLP (Natural Language Processing) módszerek együttes eredményei:

True Positive (TP) - valós egyező találat

False Negative (FN) – hibásan nem lett kiválasztva (NLP nem/G* igen) – konkrét hivatkozás (valószínű csak szám szerinti a hivatkozás, tartalmi vizsgálat ezt nem igazolja)

False Positive (FP) – hibásan lett kiválasztva (NLP igen/G* nem) – nincs számszerű hivatkozás, de tartalmilag szorosan összefügg a két ítélet

- A kutatás egyik célja hogy, hogyan lehet számítógéppel előre jelezni, hogy egy új bírósági ítélet mely régebbi ítéletekre fog hivatkozni.
- Különböző módszerekkel vizsgáltuk a kapcsolódásokat: szöveges hasonlóság alapján, a hálózat szerkezete alapján, vagy ezek kombinációjával.
- A gépi tanulósos megközelítés tanul a múltbeli példákból, és a legokosabb rangsort adja a lehetséges hivatkozásokról.
- Az eredmények azt mutatják, hogy a szöveges és hálózati információk együttes használata adja a legjobb találatokat.

Eddigi eredmények

- 1042 db Kúria által hozott ítélet hivatkozási hálóját sikerült feltérképezni.
- Azonosíthatók a hangsúlyos és komplex ítéletek, amelyek erős jogi döntésként/precedensként szolgálnak.
- 660 klasztert sikerült elkülöníteni, amelyek egy-egy részproblémát fednek le.
- Az adóhivatal szervezeti változásai (pl. 2010 és 2015-2016) jelentősen befolyásolták a klaszterek számát.
- Különösen fontos szerepük van a híd-szerepet betöltő ítéleteknek (betweenness centralitás).

A bírósági ügyszámra való hivatkozás mellett az NLP vizsgálati módszere sokkal elmélyültebb információkat ad, az együttes alkalmazás ad komplex képet.

Amiben segíthetnek az új eszközök:

- Nagy mennyiségű (esetleg) strukturálatlan szöveg feldolgozása.
- Számszerű adatok kezelése (pl: vagyonszám, fizetendő áfa, keletkezett vagyoni hátrány...)
- Bizonyítékok összefüggései, bizonyítási eszközök objektív súlya.
- Döntések közötti kapcsolati hálók (súlyponti, komplex döntések...), logikai összefüggések, közvetlen-közvetett utalások.
- Döntések számának alakulása, időszakokhoz kötése – jogalkotás számára.
- Rejtett mintázatok, egy egy ítélet/döntés kötőereje a jogrendszerben, a következmények, az ügyfelek viselkedésére/döntéseire való hatása.
- Döntések költségvetési bevételekre történő hatása.



Nemzeti Adó-
és Vámhivatal

Köszönöm a figyelmet!