Pályázati dokumentációk automatizált és objektív kiértékelése szakértők bevonása nélkül

(Automated and objective evaluation of tender documentations – without any human experts)

Pitlik László, Pitlik Marcell, Szani Ferenc, MY-X team

Kivonat: A cégek/intézmények napi élete kitermel számos pályázatot, melyek keretében számtalan dokumentum értelmezése/értékelése lenne a feladat. Ezt a kihívást automatizáltan és objektivizáltan illik kezelni, ami egyben a hatékonyság garantálását is jelenti. A hatékonyság nem más, mint minél hamarabb és emberi közreműködés nélkül tudni eljutni a pályázati győztes kihirdetéséhez. A szóban forgó dokumentációk lehetnek strukturáltak, de éppenséggel teljesen kaotikusak is. A (z online) nyomtatványok alkalmasak arra, hogy a strukturálatlanságot kordában tartsák, hiszen minden relevás információegység önálló pozícióként (mezőként) értelmezhető. Az emberi szakértők szerepe az emberi szövegek rugalmas értelmezésében rejlene. Így az emberi tényező kizárása azt is jelenti, hogy képessé kell tenni az automatizált rendszer komplex/absztrakt emberi gondolatok elemi log-adatokból való felépítésére. A pályázók személyisége által megtestesülő kockázatokat a pályázók ismert szabványokhoz (pl. helyesírás, kánon) való viszonyain keresztül le lehet képezni. Az objektum-attribútum mátrix objektumai a pályázók, attribútumai pedig az elemi log-adatok – vagyis az OAM automatikusan összeállítható. Ennek alapján pedig hasonlóságelemzésekkel a végső döntés (előkészítés) levezethető.

Kulcsszavak: OAM, hasonlóságelemzés, abszolút és relatív attribútumok

Abstract: There are situations in the daily life of institutions/enterprises where complex documentations from a lot of applicants should be evaluated in an efficient way. Efficiency means - as soon as possible AND without any human influence. The documentations can be structured and/or prepared in different/arbitrary ways. The entire chaos can be managed through (online) forms where each relevant data has a unique position (field). The role of the human experts is to interpret the human texts in the documents as flexible as possible. The exclusion of the human experts let expect that we are capable of concluding in case of abstract decision situations based on elementary log-positions (see unique fields). Risks of the applicants can be derived based on their human behaviours described through their relations to well-known standards/rules (e.g., grammatic, canon). An object-attribute matrix (OAM) where the object are the applicants, and the attributes are the logged fields can be derived in an automated way. The objective evaluation (decision) of an OAM can be done based on anti-discriminative similarity analyses.

Keywords: OAM, similarity analysis, absolute and relative attributes

# Bevezetés

Az esettanulmány egy teljes mértékben anonimizált és automatizált lépéssort mutat be valós (de tartalmában visszafejthetetlen, konkrét szervezethez, időponthoz, döntési helyzethez visszafelé már nem köthető) adatvagyon alapján – mégis reprodukálhatóságra törekedve.

A cél 39 átlagosan 5 dokumentumból álló pályázat, azaz közel 200 (mindösszesen strukturálatlan) dokumentum alapján annak feltárása, mely pályázóktól várható el a leginkább racionális teljesítése a pályázatkiírói szándékoknak. Az esettanulmány keretében a pályázatok emberi agy általi megértése soha nem volt cél és nem is volt lehetséges, vagyis az esettanulmány egy nagy mértékben context-free megközelítést mutat be.

A kontextushoz kötés utólagos, hiszen a robotizált kiértékelésre való felkérés a közel 200 nyers dokumentum keletkezése után fogalmazódott meg. Ideális esetben ennek fordítva kellene történnie. De az esettanulmány éppen az által is értékes, hogy a nem ideális esetben is hatásos működést mutat be.

## Szakirodalmi háttér

A szerzők (mint általában és így oly sokszor) jelen esetben sem vállalják át az Olvasótól annak a tudáshiánynak a gyors áthidalását jelentő szakirodalom feldolgozási feladatokat, mely keretében pl. bemutatásra kerülhetne:

* Mit jelent a hasonlóságelemzés? (vö. <https://miau.my-x.hu>)
  + Ezen belül is mit értünk objektivitáson és
  + anti-diszkriminatív elemzésen?
  + Hogyan lehet tetszőleges elemi log-okból MI segítségével emberi absztrakciókat utánozni?
* Mit jelent a szövegbányászat?
  + Ezen belül milyen elemi statisztikák alapján lehet milyen emberi absztrakciókra következtetni?
  + Mennyi programozási igény merül fel adott tartalmi részletek kaotikus kereteket közötti azonosítására?
* …

Az indoklás egyszerű: ezeket már mások máshol számos módon és esetben megtették. Ez az esettanulmány azoknak szól, akik vagy ismerik az alapvetéseket, vagy hajlandók felmerülő (személyesen releváns) kérdéseikre az internetes keresés keretében válaszokat feltárni.

Az esettanulmány azonban törekszik a reprodukálhatóság támogatására, vagyis kiadja az anonimizált teljes adatvagyont és ennek feldolgozási lépéseit, részeredményeit, mert ez valódi felelőssége minden szerzőnek, szemben az ad hoc, kiszámíthatatlan Olvasói tudáshiány reménytelen pótlására tett kísérletekkel.

A szerzők természetesen tisztában vannak azzal, hogy azzal az Olvasóval, akinek éppenséggel adekvát a szerzők tudáshiánypótló közléssora úm. nagy jót tesznek, de az esettanulmány nem akar tankönyvként szolgálni – kiszámíthatatlan arányban többé-kevésbé eleget téve az Olvasói nem-tudási problémák kezelésének.

Egy esettanulmányban azt kell leírni, ami az eset kapcsán itt és most releváns, s minden mást, ami általában véve (vagyis elismerten) fontos, azt bárhonnan máshonnan kell/lehet pótolni. Egy jó esettanulmány tehát maximálisan reprodukálható annak, aki érti az eset specifikumait/paramétereit. A miértek, MI-filozófiák nem keverendők ezen paraméterek közé, mert ezek meghaladják az esettanulmány terjedelmi korlátait és eleve ideális esetben több esettanulmány kapcsán jobban megérthetők, mint egy esettanulmányba beleerőszakolva.

# Adatvagyon

Objektumok (pályázatok, pályázók) száma: 39 – azért ennyi, mert véletlenszerűen az adott anonim pályázatra ennyi pályamű került befogadásra (hiánypótlással vagy anélkül).

Az attribútumok száma (tetszőleges lehet): itt és most

* 18 (az abszolút teljesítmények OAM-jában)
* 18 (a abszolút és relatív teljesítmény-mutatókat is tartalmazó OAM-ban)

## Abszolút attribútumok és ezek értelmezése

Az alábbi lista azért nem számozott, mert egyes tételei (pl. kulcsszavak) önmagukban is több elemet jelentenek, így a 18-as szám nem állna elő végül, ami nagyobb zavart jelentene, mint ezen mondattal jelezni, hogy az alábbi lista kényszerűen rövidebb, mint a célérték (vö. 1. ábra):

* **Hiánypótlás** volt-e (0), vagy sem (1) – ahol a hiánypótlás elmaradása kedvezőbb, mert aki képes pályázóként beleélni magát a pályázatkiíró gondolatvilágába hiba/hiány nélkül, ill. képes tételesen/figyelmesen dolgozni, az a pályázati célt is vélhetően ilyen minőségben fogja kezelni… (automatikusan kezelhető adat, ha a hiánypótlás log-ja eleve a kiíró által keletkezik)
* **Pályázati űrlap elnevezésének szabványossága** (0 = nem szabványos, 1 szabványos) - ahol a szabványos állománynév léte kedvezőbb, mert aki képes pályázóként beleélni magát a pályázatkiíró gondolatvilágába hiba/hiány/anomália nélkül, ill. képes tételesen/figyelmesen dolgozni, az a pályázati célt is vélhetően ilyen minőségben fogja kezelni… (automatikusan kezelhető adat, ha a névkonvenció szabálya eleve a kiíró által keletkezik)
* **Pályázati űrlap hossza** (byte) - ahol a pályázó hajlandó részletgazdagabban, azaz egyre több arcát megmutatva pályázni, az vélhetően képes lesz a pályázati dokumentációt is jobb minőségben elkészíteni… (automatikusan kezelhető adat, ha a digitális állományok kezelésének log-ja eleve a kiíró által keletkezik)
* **A pályázati kiírás kulcsszavai** (melyek mibenléte és számossága a pályázatkiíró döntésén alapul, de automatikusan a pályázat címe alapján is feltárható: pl. **xxx** technológia alkalmazása **zzz** környezetben **qqq** céllal)… (az egyes kulcsszavak pályázati dokumentációban való előfordulása kereséssel automatikusan megállapítható)
  + Minél több egy kulcsszó előfordulása, annál triviálisabb előzetesen ennek elvárása nélkül!, hogy a pályázó számára a pályázatkiírói preferenciák szervesen léteznek.
  + A pályázatban 3 kulcsszó kerül kiválasztásra, az xxx és az yyy félreérthetetlen és megkerülhetetlen fontossággal, míg a qqq annak demonstrálására, hogy további (preferált?!/preferálandó/preferálható) szakterületek/alkalmazási területek nevesítése nem ördögtől való (vö. <https://miau.my-x.hu/miau/140/la140.doc>)
* **A pályázó angol nyelvtudásának szintje** (1 = középfok) / 2 = felsőfok) – ahol minél magasabb a szint, annál jobb a pályázó, hiszen annál mélyebbre volt képes beásni magát egy probléma kezelésébe, így elvárható, hogy a pályázatkiíró számára is minőségibb munkát tud végezni (nem beszélve a direkt többletről, amit a nyelvtudás maga jelenthet)… (a középfok és felsőfok szavak szövegbányászatilag automatizáltan feltárhatók, ill. űrlapjelleggel eleve specifikus log-ként keletkeztethetők)
* **A pályázó által ismert nyelvek száma** (db) – a nyelvtanulás eleve egy fajta IQ-teszt, minél több tapasztalata van ebben valakinek, annál kreatívabbnak vélhető általában (nem beszélve a több nyelvű szakirodalmi feldolgozás és/vagy kommunikációs képesség direkt előnyeiről)… (a nyelvek száma egy előre megadott kulcsszólista alapján a szöveg adott részére vonatkozóan automatizálhatóan feltárhatók, , ill. űrlapjelleggel eleve specifikus log-ként keletkeztethetők – a nyelvek száma kapcsán nem lett figyelembe véve ezek szintje, de ez további attribútumokkal kezelhető lesz a jövőben)
* **A pályázó által kapott dicséretek száma** (db) – a pályázati űrlap önálló kérdésére adott tételek száma minél nagyobb, annál nagyobb referenciával rendelkezik a pályázó személyisége a megfelelni akarás kapcsán… (a tételek száma pl. a feltárható dátumok alapján automatizálható akár utólag is)
* **A pályázó publikációnak száma** (db) - a pályázati űrlap önálló kérdésére adott tételek száma minél nagyobb, annál nagyobb aktivitás-referenciával rendelkezik a pályázó a dolgozni akarás kapcsán… (a tételek száma pl. a feltárható dátumok alapján automatizálható akár utólag is)
* **A pályázó konferencia-aktivitásainak száma** (db) – a pályázati űrlap önálló kérdésére adott tételek száma minél nagyobb, annál nagyobb aktivitás-referenciával rendelkezik a pályázó a dolgozni akarás kapcsán… (a tételek száma pl. a feltárható dátumok alapján automatizálható akár utólag is)
* **A pályázó egyéb publikációs aktivitásainak száma** (db) – a pályázati űrlap önálló kérdésére adott tételek száma minél nagyobb, annál nagyobb aktivitás-referenciával rendelkezik a pályázó a dolgozni akarás kapcsán… (a tételek száma pl. a feltárható dátumok alapján automatizálható akár utólag is)
* **MTMT-utalás léte** (1/0) – ha van, az előnyösebb, mint ha nincs, hiszen a kánonhoz való alkalmazkodást demonstrálja az MTMT-profil léte… (az MTMT keresési kifejezésre kereséssel a bináris adat automatikusan feltárható)
* **MTMT-n a szerzői link alapján látható publikációk száma** (db) – minél nagyobb, annál jobb, hiszen az MTMT profil teljessége releváns kánon-konformitási mutatószám… (a szerzői MTMT-link alapján megnyíló oldalról az összes publikáció száma automatizáltan leolvasható – vö. pl. web-scraping)
* **Adott kulcsszóhoz tételesen kifejtett kompetenciákat bemutató szöveg karakterszáma** (db) – minél nagyobb, annál jobb, hiszen ideális esetben az a pályázó, akinek számos említésre érdemes múltbeli aktivitása létezik, az erről ösztönösen részletesen számol be… (automatizálni a karakterek számát triviálisan meg lehet határozni, de azon pályázók kapcsán, akik a szabadszöveges mezővel visszaélésszerűen bántak, sajnos emberi szakértői döntést kellett bevonni: vö. voltak olyan pályázók, akik arról értekeztek tények közlése helyett: mennyire fontos lenne az adott jelenség és mit terveznek a jövőre vonatkozóan 🡨 ezen pályázók ellenmondás-jelzést kaptak az emberi adatgyűjtés-felügyeleti munka keretében)
* **A pályázó születési dátuma** (évszám) – minél kisebb, annál jobb, mert az élettapasztalatot preferálni akarta a pályázatkiíró az adott helyzetben, de ez fordítottan is priorizálható, vagy elhagyható… (a dátum triviálisan kinyerhető automatizáltan)
* **A pályázó diplomaszerzésének dátuma** (évszám) – minél kisebb, annál jobb, mert az élettapasztalatot preferálni akarta a pályázatkiíró az adott helyzetben, de ez fordítottan is priorizálható, vagy elhagyható… (a dátum triviálisan kinyerhető automatizáltan)
* **A pályázó doktori iskolai tagságának kezdete** (évszám) - minél nagyobb, annál jobb, mert a a doktori tevékenység még le nem zárt szakaszának minél nagyobb tartamát preferálni akarta a pályázatkiíró az adott helyzetben, de ez fordítottan is priorizálható, vagy elhagyható… (a dátum triviálisan kinyerhető automatizáltan – itt kell utalni a második ellenmondás-jelre, mely az elsővel szemben automatikusan vezethető le: nem illik, hogy valaki később szerezzen diplomát, minthogy felvételt nyerjen egy doktori programba)

A képen szöveg, térkép, beltéri, világos látható

Automatikusan generált leírás

1. Ábra: Az abszolút teljesítmények OAM-ja (forrás: saját ábrázolás)

## Relatív attribútumok és ezek értelmezése

Továbbra is abszolútként kezelendő attribútumok:

* **Hiánypótlás** volt-e (0), vagy sem (1)
* **Pályázati űrlap elnevezésének szabványossága** (0 = nem szabványos, 1 szabványos)
* **Pályázati űrlap hossza** (byte)
* **A pályázati kiírás kulcsszavai**
* **MTMT-utalás léte** (1/0)
* **A pályázó doktori iskolai tagságának kezdete** (évszám)

A pályázó életkorával elosztott abszolút adatok annak kifejezésére, hogy egy életév alatt milyen teljesítmények jöttek létre (vö. 2. ábra):

* **A pályázó angol nyelvtudásának szintje** (középfok=1, felsőfok=2)
* **A pályázó által ismert nyelvek száma** (db)
* **A pályázó által kapott dicséretek száma** (db)
* **A pályázó publikációnak száma** (db)
* **A pályázó konferencia-aktivitásainak száma** (db)
* **A pályázó egyéb publikációs aktivitásainak száma** (db)
* **MTMT-n a szerzői link alapján látható publikációk száma** (db)
* **Adott kulcsszóhoz tételesen kifejtett kompetenciákat bemutató szöveg karakterszáma** (db)
* **A pályázó diplomaszerzésének dátuma** (évszám) – minél kisebb, annál jobb
* **A pályázó születési dátuma** (évszám)🡨kiesik, mint vetítési alap
* **Az MTMT-ben tárolt objektumszám és a (publikáció + konferenciaaktivitás) abszolút távolságának aránya a pályázati aktivitások (publikáció + konferenciaaktivitás) számához képest** (abs(db-db)/(db)) 🡨 vagyis minél inkább teljeskörű az MTMT profil a pályázati adatokhoz képest, annál hitelesebb a pályázó személyisége, minőségbiztosításhoz való viszonya… (ez a származtatás és általában véve a fentebb jelzett relativálás értelemszerűen minden esetben automatizálható)

A képen asztal látható

Automatikusan generált leírás

1. Ábra: A vegyes (abszolút és relatív) OAM (forrás: saját ábrázolás)

# Elemzési lépések

Az alábbi ábrasor bemutatja, miként juthatunk el a nyers adatoktól az eredményekig:

Az anti-diszkriminatív hasonlóságelemzés (Y0) inputként (vö. <https://miau.my-x.hu/myx-free/>, <https://miau.my-x.hu/myx-free/coco/index.html>) mindenkor rangsorszámokat és egy konstanst (1000000) vár el (vö. 3. és 4. ábra):

A képen szöveg, világos látható

Automatikusan generált leírás

1. Ábra: Az abszolút OAM rangsornézete (forrás: saját ábrázolás)

A képen asztal látható

Automatikusan generált leírás

1. Ábra: A vegyes (relatív) OAM rangsornézete (forrás: saját ábrázolás)

Az anti-diszkriminatív optimalizálás keresi azt a lépcsős függvényt, mely a legjobban engedi közelíteni pályázónként az 1000000 pontos normaértéket. Ha ez sikerül(ne), akkor minden pályázat egyformán értékes a 18 dimenziós értékelési térben. Ha nem sikerül, akkor lesznek norma feletti (>1000000) és norma alatti (<1000000), ill. normaszerű (1000000 pontos) pályázatok. A norma értéke minél nagyobb, annál jobb – solver-technikai okból.

A 3. és 4. ábra inverze (pl. 39 objektum esetén 40-nyersadat cellánként) inverz idealitás-becslésre vezet, így minden olyan pályázat validan értelmezhető, ahol a direkt becslés normafelettiségéhez az inverz elemzésben normaalattiság kapcsolódik és fordítva.

# Eredmények

A képen asztal látható

Automatikusan generált leírás A képen asztal látható

Automatikusan generált leírás

1. Ábra: Abszolút és vegyes/relatív OAM-ok objektív eredményei (forrás: saját ábrázolás)

Jelmagyarázat:

* Zöld-sárga-piros színkódok: minél zöldebb, annál ideálisabb a pályázó/pályázat
* Kiszürkített karakterek: ellentmondásos háttérrel rendelkező pályázatok(kizárásuk javasolt)
* Vastag karakterek: normafelettiséget jelző objektumok
* Becslés: optimalizált norma körüli idealitás
* Sorrend\_opt: A becslések sorrendje, minél nagyobb, annál jobb (zöldebb) alapon
* Naiv: OAM-sorszámok átlagának egész értéke (minél kisebb, annál jobb)
* Sorrend\_naiv: a naiv rangsorátlagok sorszámozása (minél kisebb, annál jobb)
* Eltérés: eltérés a naiv és az optimalizált sorrend sorszámai között
* Validitás: direkt és inverz modellek alapján racionális inputok jele (1=valid, 0=invalid)
* Id: pályázati azonosító (anonimitás alapja)
* Ellentmondás: a korábban jelzett értelmezési zavarok státuszváltozója
* Narancssárga mezők: a relatív modell győztes tartománya
* Abszolút-helyezés: a relatív helyezés előkelő objektumai által elért helyezés az abszolút modellben
* Vastag keret: az abszolút teljesítmény értékelésben norma alatt, de a relatív értékelésben norma feletti objektum jelzése (nagy potenciállal rendelkező, de még kezdő pályázó jele)

Az 5. ábra validált elemzési eredményekkel automatizáltan (az attribútumok és irányaik kijelölésén túl, mely inputparaméter-halmaz nem pályázat függő, vagyis már a kiíráskor meghatározható, szubjektív elemek/lépések nélkül) levezetve mutatja be a leginkább szerződésre alkalmas személyiségeket.

# Vita

Mint az eddig esetleírásból látható:

* Az attribútumok definiálása a pályázatkiíró feladat.
* Annál egyszerűbb az OAM-ok előállítása, minél inkább előre tudjuk, milyen attribútumokra lesz szükség.
* Az attribútumok iránya nem pályázófüggő = objektív preferenciaként értelezhető.
* Ha változnak az attribútumok, ezek irányai, akkor változhatnak a becslések.
* Invalid objektumról nem szabad nyilatkoznia a robot-evaluátornak.
* Az ellentmondást (irracionalitást, visszaélést) felmutató pályázatok akár már előzetesen kizárhatók az értékelésből.
* A bemutatott módszertan akár teljes mértékben is automatizálható, amennyiben a szabadszövegek keletkezését legördülő űrlaptartalmakkal minimalizáljuk.
* A fenti lépéssor sok, triviálisan létező log-adat alapján tesz kísérletet arra, hogy a pályázó hitelességét modellezve el lehessen tekinteni a munkatervek mindegyikének tételes emberi értelmezésétől.
* A mindenkori pályázatkiíró szakmai, etikai felelőssége, hogy az objektíven kevésbé kockázatos személyiségek munkaterveit emberi szemmel is ellenőrizze az esetleges visszaélésszerű szómágiák feltárása érdekében…

# Konklúziók

A log-alapú, objektivitásra törekvő pályázatértékelés automatizált működtetése lehetséges, s ha nem tudná valaki, mit is tesz a robot-evaluátor, akkor a Turing-teszten a fenti lépéssor racionális pályázók esetén (akik szintén nem tudnak a robotról), meg kell, hogy feleljen lényegében tetszőleges kontextusú pályázat esetén.

# Referenciák

… szövegközben …

Valamint:

* <https://miau.my-x.hu/miau/275/review_anomalies.docx>
* <https://miau.my-x.hu/miau/273/Naiv_optimalizalt_verziok2.docx>
* <https://miau.my-x.hu/miau/277/mirror_ball_based_multi_layered_evaluation.docx>
* <https://miau.my-x.hu/digeco/2020/2020osz/digeco_tdk_publication.docx>
* <https://miau.my-x.hu/miau/273/otdk_2021_abstract_standards.docx>
* <http://miau.my-x.hu/miau/200/otdk_v1.doc>
* <http://miau.my-x.hu/miau/200/otdk_v2.doc>
* <https://miau.my-x.hu/miau2009/index.php3?x=e0&string=letra>
* (inkl. a fenti dokumentumokban jelzett hivatkozások)