GAZDASÁGINFORMATIKA DOKTORI ISKOLA

TÉZISGYŰJTEMÉNY

Neusch Gábor Loránd

Munkaerőpiaci Adattárház Tervezése Kompetenciatrendek Elemzésére

című Ph.D. értekezéshez

Témavezető:

Borbásné Dr. Szabó Ildikó

egyetemi docens

BUDAPEST, 2021

Informatikai Intézet

TÉZISGYŰJTEMÉNY

Neusch Gábor Loránd

Munkaerőpiaci Adattárház Tervezése Kompetenciatrendek Elemzésére
című Ph.D. értekezéshez

Témavezető:

Borbásné Dr. Szabó Ildikó

egyetemi docens

Tartalomjegyzék

1. Kutatási előzmények és a téma indoklása

Grundke és szerzőtársai szerint a jövőben a munkaerőpiacon jelentős változások fognak történni a digitális technológiák térnyerésének folytatódásával, ami azt fogja eredményezni, hogy egyes munkákat teljesen automatizálnak, míg mások esetében a feladatok és a munka természete fog jelentős változáson keresztülmenni (Grundke *et al.*, 2018). Azok a szolgáltató szektorban dolgozók, akiknek a feladatai elsősorban kognitív készségeket, képességeket és tudást igényelnek, valamint könnyen algoritmizálhatók, elveszíthetik az állásukat, de legalábbis számolniuk kell az automatizáció által generált nagyívű változások hatásaival. Nagyon valószínű, hogy bárhogy is lesz a jövőben, a 21. század emberének pár évente meg kell újítania magát, folyamatosan új tudásra és készségekre kell szert tennie az egyre gyorsuló technológiai fejlődés közepette, olyan ütemben, ahogy "új munkakörök alakulnak ki a digitális forradalomnak köszönhetően", és "az egyes munkakörök tartalma, természete és a szükséges készségek köre folyamatosan változik" (Grundke *et al.*, 2018, p. 6).

A munkaerőpiaci kereslet nagyon volatilis, rendkívüli ütemben változik. A kompetenciák és a tudás a kereslet szempontjából gyorsan avulnak, egyre újabb és újabb ismeretekre van szükség. Az automatizáció is egyre inkább teret nyer mind több területen, a gépi tanulás és a mesterséges intelligencia alkalmazásához kötődő gépesítés és robotizáció sok, kognitív készségeket igénylő feladat esetében is kiváltja az emberi munkaerőt. Azonban szemben a munkaerőpiac rapid keresletváltozásával, a kínálati oldalon nagyon nehezen változó, nehezen alkalmazkodó intézmények igyekeznek ennek a keresletváltozásnak reaktív módon „megfelelni". A képzőintézmények többségére a merev, hierarchikus felépítés, a bürokrácia, a belső és külső politikai környezetnek való kitettség jellemző. A tanterveket általában kizárólag az oktatók saját, szükségszerűen korlátozott tapasztalata, esetlegesen szűkebb kutatási területe, elfogultságai stb. alapján állítják össze. Objektív információk hiányában a szak- és tantervek a legjobb szándék mellett sem tudják a piaci igényeket[1] megfelelő mértékben tükrözni. A képzési tervek

[1] Sok esetben természetesen nem (csak) ez a cél, például az alapozó tárgyak, vagy általánosságban az alapképzés esetében, amikor nem a szakmaspecifikus ismereteket, hanem az azok megalapozására szolgáló általános kompetenciákat kívánja az oktatás erősíteni. Emellett természetesen számos egyéb, például pedagógiai, társadalompolitikai stb. szempontot is figyelembe kell venni a tantervfejlesztés során.

akkreditálási eljáráson mennek át, ami szintén hatalmas bürokratikus szervezetekben történik. Ez szintén megnehezíti az igényváltozásra való gyors reagálást.

Az előzőeken felül, a felsőoktatási intézmények esetében további nehézséget jelent, hogy tulajdonképpen nem is a jelenlegi piaci igényeket kellene figyelembe venniük, hanem a jövőbelieket, hiszen az a diák, akinek a képzési tervét egy adott évben összeállítják, csak évekkel később fog kilépni a munkaerőpiacra. Viszont a jövőbeli igényekre való felkészülés, pontosan a munkaerőpiaci igények rendkívül gyors változása miatt, rengeteg bizonytalansággal terhelt.

A rendkívül változékony környezetben zajló extrém gyors változások közepette tehát, mikor a technológia exponenciális ütemben fejlődik – habár a Moore által előrevetített ütem mára lassulni látszik (Waldrop, 2016) – , nagyon nehéz a következő 10-20 évre, hosszú távra előrejelezni, és olyan javaslatokat adni a felsőoktatási intézményeknek, melyek alapján a munkaerőpiaci kereslethez alkalmazkodó tanterveket tudnak kidolgozni. Akik mégis megpróbálkoznak ilyen hosszú távú javaslatokkal, azt hangsúlyozzák, hogy az érzelmi intelligenciához és a szervezőkészséghez kapcsolódó tanulói kompetenciákat kell elsősorban fejleszteni (Beck and Libert, 2017), és egy olyan erős alapot kell biztosítani a hallgatóknak, amelyre könnyedén építhetnek, amikor az aktuális munkaerőpiaci keresletnek megfelelően új készségeket és tudást sajátítanak el.

A legfontosabb alapkészségek megalapozásán túl, melyeket a hallgatók hosszú távon használni tudnak, a rövid távú versenyképesség biztosítása érdekében az oktatási intézményeknek folyamatosan összhangban kellene tartania a tanterveket a munkaerőpiaci kereslettel, amit számos tényező befolyásol és előrejelzése egyáltalán nem triviális. Az érintettek, a tárgy- és szakfelelősök legnagyobb nehézsége ebből a szempontból az, hogy nem látják előre az igények változását. Ha objektív képet kaphatnának arról, hogy a kompetenciák iránti kereslet hogyan alakul a munkaerőpiacon az időben, akkor következtetéseket tudnának levonni a jövőbeli trendekre vonatkozóan. A legfontosabb információ, amire egy ilyen előrejelzéshez szükség van, az a munkaerőpiaci kereslet reprezentációja az egyes pillanatokban, idősorosan rögzítve. Ezt tulajdonképpen „pillanatképek” összességének is felfoghatjuk az egyes időpillanatokban igényelt kompetenciahalmazokról.

Disszertációmban erre, a kompetenciakereslet és a felsőoktatási kínálat rövid távú összehangolására igyekszem fókuszálni egy olyan keretrendszer felvázolásával, melynek

segítségével a munkaerőpiaci trendek és a keresletet leíró adatok valós időben elemezhetőek. Egy munkaerőpiaci „adattárház” koncepciójának kidolgozására teszek kísérletet, melyben a munkaerőpiaci keresletre vonatkozó, álláshirdetésekben megjelenő információk összegyűjthetőek. Megvizsgálok továbbá olyan módszereket, melyek segítségével az állásajánlatokban explicit és implicit módon megjelenő, kompetenciakeresletet tükröző információk kinyerhetőek. Egy ilyen jellegű információforrás alapján – esetlegesen kiegészítve egyéb forrásokból származó adatokkal, piaci és iparági elemzésekkel, riportokkal, mint például a *Gartner Hype görbéje* stb. – a felsőoktatás döntéshozói folyamatosan hozzá tudnák igazítani a kurzusok tematikáját a munkaerőpiaci kereslethez, és olyan kompetenciák elsajátítására tudnak lehetőséget kínálni a hallgatóknak, amit azok karrierjük első éveiben sikeresen értékesíteni tudnak a piacon. A jelenleg zajló, folyamatos diszkontinuitást okozó technológiai fejlődés közepette, mikor az *MI* és *GT* megbontó hatása folyamatosan átalakítja a munkaerőpiacot, ezek az információk jelentős versenyelőnyt jelenthetnek egy oktatási intézmény, illetve magasabb szinten akár egy nemzetgazdaság számára.

A tágabb kutatásom tehát ahhoz a problémához kapcsolódik, hogy milyen módszerekkel, hogyan lehet elemezni és előrejelezni a munkaerőpiaci keresletet, vizsgálni a kínálat kereslethez való illeszkedését. Továbbá összehangolni azokat, elérni, hogy a képzési kimeneti követelmények egyre jobban tükrözzék az álláshirdetésekben megjelenő kompetenciaigényeket. Ahogy az előzőekben részleteztem, kínálat alatt itt nem az egyének (munkavállalók) összességét értem, hanem valamiféle képzés által nyújtott kimeneti minőséget, míg a keresleti oldalon kompetencia (mint erőforrás) szükségletről beszélek. Mivel a teljes kutatás kidolgozása túlnyúlna egy doktori értekezés keretein, ezért az értekezésben a tágabb kutatást megalapozó keretrendszer alkalmazhatóságát mutatom be. Azaz a disszertáció és annak kutatási kérdései a teljes problématérnek csak egy részletére kívánnak megoldásokat találni, a munkaerőpiac keresleti oldalára fókuszálva, arra, hogy miként lehet a kompetenciakeresletet jelző információkat hatékonyan feltárni és tárolni, ezzel megalapozva olyan elemzéseket, melyek később a kínálat jobb illesztését lehetővé teszik.

A bemutatott keretrendszer – implementációja után – a felsőoktatás döntéshozói, elsősorban az oktatók, illetve a szakfelelősök számára nyújthat majd segítséget olyan tantervek létrehozásában, melyek által a hallgatók számára kínált kompetenciák még akkor is érvényesek és eladhatóak lesznek, amikor az, aki ma kezdi egyetemi karrierjét,

kilép a munkaerőpiacra. Ez elsősorban azt igényli, hogy a tantervek olyan elemekből álljanak, melyek biztosítják a naprakész és a munkaerőpiaci kereslet alapján megfogalmazott kompetenciák átadását a hallgatók számára.

Az információgyűjtés során azt az alapfeltevést használom, hogy a kutatásom keretei között elérhető módon, a kompetenciák iránti kereslet legjobban az álláshirdetésekben képződik le. Mivel az emberierőforrás-szükségletet, azaz azt, hogy milyen munkaerőre van szüksége egy adott cégnek, az elvégzendő feladat, illetve a betöltésre szoruló pozíció alapján lehet legjobban megfogalmazni, és kommunikálni a potenciális jelöltek felé, így az álláshirdetésekben valamilyen mértékben szükségszerűen meg kell jelennie a vállalatok kompetencia-szükségletének.

Ezt a feltevést megerősíti a szakirodalomban például Pitukhin és szerzőtársai, akik azt írják, hogy a legtöbb szakmai követelmény, amit a munkáltató a jelentkezővel szemben állít, megjelenik az állásajánlatokban (2016, p. 2028). Wowczko szintén kiemeli, hogy az online toborzás megjelenésével hatalmas mennyiségű, potenciálisan hasznos információ áll a kutatók rendelkezésére a keresett kompetenciákról (Wowczko, 2015, p. 34). Zhao és szerzőtársai (2015, p. 4013), a CareerBuilder.com kutatói, fél milliárd angol nyelvű álláshirdetésből vett minta vizsgálata alapján azt találták, hogy azok kilencven százalékában megjelennek a jelen dolgozat szempontjából kompetenciaként elfogadható kifejezések. Az előzőekkel ellentétben Nasir és szerzőtársai (2020) tapasztalataik között megemlítik, hogy az általuk feldolgozott hirdetések inkább a pozíciók leírását (*vacancy information*) emelik ki, és kevésbé részletezik az elvárt készségeket.

1.1. Kutatási kérdések

Disszertációmban tehát egy olyan rendszer tervét kívántam felvázolni, és legfontosabb, hogy az egész megoldás alapját biztosító moduljainak megvalósíthatóságát alátámasztani, melynek fő céljául tűztem ki, hogy olyan információkat szolgáltasson, melyek az álláshirdetésekben megjelenő foglalkozásokról, az igényelt kompetenciákról, végzettségekről és képzettségekről, illetve mindezek időbeli alakulásáról adnak képet riportok, elemzések stb. formájában. Egy ilyen rendszer implementációja természetesen nem egyemberes feladat. Az értekezés keretei arra adtak lehetőséget, hogy elméletileg megalapozott javaslatot tudjak tenni az adatok tárolásának módjára, elkezdjem az adatgyűjtést, és erre építve megvizsgáljak olyan módszereket, melyek segítségével az

álláshirdetések leírásaiban a kompetenciaelemek, illetve a kapcsolódó foglalkozások beazonosíthatóak. Ennek megfelelően határoztam meg a kutatási kérdéseimet is.

A rendszer elsődleges input adatai internetes álláskereső portálokon közzétett hirdetések, melyek tárolására több probléma miatt is szükség van. Egyrészt mivel azok internetes forrásból származnak, nem garantálható, hogy elérhetőek lesznek a teljes elemzési ciklus alatt. Például egy álláshirdetés esetében, mikor az többé már nem releváns a hirdető számára – azaz a pozíciót betöltötték – lekerül az állásportálról, de számunkra továbbra is fontos a benne tárolt információ, hiszen a tervezett elemzésekhez a múltbeli adatokra is szükségünk van. Másrészt az álláshirdetések zajossága miatt elő kell azokat készíteni az elemzéshez. Az előkészítésre azért van szükség, mert azok az információk, melyek az elemzéseink alapjául szolgálnak, nem mindig állnak rendelkezésre strukturált formában, vagy akár explicit módon az álláshirdetésekben, így azokat annak szövegezéséből kell valamilyen módon kinyerni, és amennyiben ez az előfeldolgozás megtörtént, érdemes a feltárt információkat letárolni. Az első kutatási kérdésem tehát azt vizsgálja, hogy hogyan érdemes kidolgozni az adatok tárolására szolgáló megoldást, illetve milyen adatköröket érdemes gyűjteni.

1. Kutatási kérdés: Mi a legmegfelelőbb eszköz (adattárolási platform) az internetes forrásokból legyűjtött – rendkívül heterogén, nagy mennyiségű, strukturálatlan adat – álláshirdetések, illetve kapcsolódó gazdasági és statisztikai tények tárolására, oly módon, hogy az így rögzített információ alapján az érintettek számára értékes elemzéseket lehessen adni, melyek a munkaerőpiacon igényelt kompetenciák időbeli, illetve egyéb dimenziók mentén történő alakulását mutatják be.

1.1. Milyen adatköröket érdemes gyűjteni és mi lehet az adatok forrása? A dolgozatban megvizsgáltam az adatgyűjtéshez használható technológiákat, és bemutattam annak a „keresőrobotnak” az implementációját, mellyel a rendszer megalapozásához és elvégzett kísérletekhez szükséges adatokat összegyűjtöttem.

1.2. Milyen szempontok szerint érdemes kiválasztani az adattárolási platformot? A dolgozat elméleti részében összegyűjtöttem azokat a megfontolásokat, amelyek aztán a jelen probléma szempontjából legmegfelelőbb rendszer kiválasztásának alapját képezhetik. Megvizsgáltam, hogy melyik az az adattárolási struktúra, ami legjobban szolgálja a felépíteni kívánt rendszer céljait. Azaz hogy melyik az a tárolási technológia, amely a célnak legjobban megfelel, például egy hagyományos adattárház megoldás, vagy

egy *big data* környezetben divatos adattó (*data lake*)? Továbbá a felvázolt szempontok alapján a dolgozat elméleti részében összehasonlítottam konkrét adattárolási architektúrákat is, azzal a céllal, hogy a felvázolt koncepció helyességét alátámasszam az egyes megoldásoknak a választás alapjául szolgáló szempontok szerinti összehasonlításával.

1.3 A disszertációban kitérek továbbá arra is, hogy hogyan érdemes a legyűjtött adatokat betölteni és sémába rendezni. Szükséges-e egyáltalán sémákat definiálni a tároláshoz, mint ahogy egyes elterjedt adattárház architektúrák esetében, ahol általában egy *ETL* (*extract, transform, load*) folyamat során előre meghatározott relációs adatmodellbe töltik be a megtisztított és rendszerezett adatot. Ha igen, hogyan nézzen ki ez a séma? Vagy hatékonyabb az adatokat – mint egy adattó esetében – abban a formában tárolni, ahogy legyűjtöttük a forrásból – például egy elosztott fájlrendszeren – míg a logikát az elemzés, feldolgozás során alkalmazzuk, ami történhet egy *MapReduce* algoritmus vagy egy *Spark* alkalmazás segítségével?

Összegezve tehát fontos megvizsgálandó kérdés, hogy melyik megközelítés a legmegfelelőbb a célra, egy adattárház, egy adattó vagy éppen egy hibrid megoldás? Kapcsolódó kérdés továbbá, hogy érdemes-e az állásajánlatok adatait előre definiált sémákban tárolni és egy *ETL* folyamatot építeni az adattisztítás és betöltés elvégzésére, avagy az adatokat a legyűjtött formában érdemesebb-e inkább tárolni, majd a logikát igény esetén az adatok elemzése során alkalmazni?

A szükséges adatkörök és a számukra legmegfelelőbb tárolási megoldás feltárása után, az összegyűjtött adatok feldolgozását végeztem el annak érdekében, hogy megalapozzam a kutatás későbbi szakaszait, ahol a cél a kapcsolat kiépítése lesz a munkaerőpiaci kereslet és az oktatási kínálat között, és az alap megteremtése azok későbbi harmonizációjához. Jelen tézis feltevése, hogy ez a kapcsolat kompetenciaalapon építhető fel, ennek megfelelően a dolgozat egyik célja, a keresleti oldalon az álláshirdetésekben megjelenő kompetenciaelemek beazonosítása.

2. Kutatási kérdés: Milyen információtechnológiai megoldásokkal lehet az egyes szabadszöveges leírásokban, például jelen tézis esetében álláshirdetésekben – de hasonló módon akár folyamatmodellek feladatleírásaiban (*task description*) – explicit módon megjelenő kompetencia-elemeket automatikusan beazonosítani? Illetve az egyes megoldásokkal milyen pontosság és felidézési arány érhető el?

Az álláshirdetések, vagy a folyamatmodellek feladatleírásai általában tartalmazzák a kompetenciáknak azt a legfontosabb listáját, melyekre egy munkavállalónak szüksége van az adott álláshirdetésben leírt pozíció betöltéséhez vagy adott feladat elvégzéséhez. Azonban a szabadszöveges korpuszban ezen kompetenciákat reprezentáló szóennesek beazonosítása nem triviális. A kutatási kérdés vizsgálata során arra keresek megoldást, hogy hogyan, illetve milyen eszközökkel lehetséges az álláshirdetések korpuszában ezeket a releváns tudáselemeket, kompetenciákat reprezentáló, explicit megjelenő n-gramokat beazonosítani.

A kompetenciák beazonosítása történhet például szövegbányászati eszközökkel. Ezt a folyamatot külső források felhasználásával is támogathatjuk, mely források mintegy kompetenciaszótárként segíthetik az elemek beazonosítását az álláshirdetésekben. Ilyen külső források lehetnek például az ESCO ontológia készség (*skill*) pillére, vagy a STUDIO ontológia, melyeket az értekezésben részletesen ismertetek.

A disszertációban továbbá részletesen megvizsgálom, hogy miként lehet az állásajánlatokban látensen, implicit, rejtett módon „jelen levő" kompetenciaelemeket feltárni. Ezt a kérdéskört járja körül a harmadik kutatási kérdésem.

3. Kutatási kérdés: Milyen módszerekkel, illetve milyen technológiák segítségével lehet az implicit (látens), az álláshirdetésekben közvetlenül nem megjelenő, de az azok által meghatározott kontextusban releváns kompetenciaelemeket feltárni? Milyen adatforrásokra lehet és érdemes támaszkodni ezen rejtett objektumok beazonosításához?

Az álláshirdetések explicit kijelölnek olyan kompetenciákat, melyek meglétére szükség van az adott pozíció ellátásához, ugyanez igaz a folyamatmodellek esetében a munkakör vonatkozásában. Ezek az expliciten megjelenő kifejezések – szemantikusan (például jelentésük vagy a köztük lévő kapcsolatok alapján) vagy egyszerű statisztikai alapon (például együttes előfordulás, az egyes elemek távolsága stb.) – kijelölhetnek olyan, expliciten nem megjelenő, látens kompetenciákat is, melyek szintén relevánsak lehetnek az adott pozíció, vagy munkakör kontextusában.

1.2. A kutatás keretei

Az adatok közül a tézis célja szempontjából legfontosabbak a kompetenciaigények, így a kutatás fontos mérföldkövei közé tartozik azok beazonosítása, gyűjtése és szemantikus gazdagítása a szöveges bemeneti adatok alapján, egy adott objektumtípushoz – a

dolgozatban felvázolt felhasználási eset (*use case*) esetében egy pozícióhoz – kapcsolódóan (ami tulajdonképpen itt a granularitást, a feldolgozás legkisebb egységét jelöli). Disszertációm jelentős része ennek megfelelően tehát az álláshirdetésekben explicit megjelenő, illetve az azokhoz implicit kapcsolódó kompetenciák beazonosításának lehetőségeit tárgyalja.

A munkaerőpiaci kínálatot sok különböző csoport alkotja; például a kezdő munkavállalók – köztük a frissen végzettek –, a már tapasztalt és éppen munkahelyet váltók, vagy éppen azok, akik ilyen-olyan okokból éppen újraintegrálódnak, visszakapcsolódnak a rendszerbe stb. Jelen kutatás fókuszában ebből a szempontból a felsőoktatási szektor áll, ezen belül a friss diplomások, a dolgozat végső céljával összhangban, azaz hogy olyan információkat nyújtson az oktatók számára, ami alapján a képzési kimeneti követelmények olyan tartalommal tölthetőek meg, melyet a végzett hallgatók el tudnak adni a piacon.

A keretrendszer kidolgozása során a problémát egyszerűsítendő, a vizsgált állásajánlatok körét az informatikai szektorra szűkítettem, ennek megfelelően szűrtem meg a külső ontológiák tartalmát is, amit a kompetenciák beazonosításához, mintegy szótárként használok. Természetesen amennyiben a felvázolt koncepció működőképessége bizonyítást nyer (*proof of concept, PoC*), úgy a megoldás könnyen adaptálható más területekre is.

Az értekezésben a teljes kutatási térnek csak a munkaerőpiaci kereslet oldalát érintő feladatok kerültek kidolgozásra, tehát a munkaerőpiaci „adattárház” koncepciója, és az, hogy miként lehet kinyerni a kapcsolódó kompetenciákat az álláshirdetésekből, illetve hogy egyes szemantikus forrásokat milyen módon lehet a feldolgozásba bekapcsolni a látens információ kinyerése érdekében.

A tézisben elsősorban a felvázolt koncepció helyességét kívánom vizsgálni, és a feldolgozott források (álláshirdetések) nyelvéül az angolt választottam. A választás egyrészt azért esett az angol nyelvre, mert az az „informatika nyelve”, és számtalan technológia neve vagy éppen szakkifejezés magyar fordításban nincs elterjedten használatban. Továbbá a magyar nyelv feldolgozása (a nyelv sajátosságai miatt) bonyolultabb és több erőforrást igényel, mint az angol nyelvé, így egy *PoC* kidolgozása során indokoltabb utóbbi használata. Végül pedig angol nyelven jelentősen több releváns álláshirdetés érhető el az interneten, így az input adatok számossága növelhető, a magyar

nyelvű feldolgozás esetéhez képest. Az előző okokból kifolyólag tehát a koncepció helyességének vizsgálatához jobban megfelelnek az angol nyelvű források, míg amennyiben a tézisben felvázolt modell működőképessége bizonyítást nyer, és igény mutatkozik annak eredményeire, akkor a kutatás későbbi fázisaiban a rendszert fel lehet készíteni a magyar nyelv kezelésére is. A dolgozat feltételezése továbbá, hogy az élvonalbeli technológiákhoz szükséges kompetenciákra mutatkozó igény az angol anyanyelvű piacokon jelenik meg először, mely feltevésről részletesebben a disszertációnak a kutatás kiterjesztési lehetőségeit tárgyaló alfejezetében írok.

1.3. A kutatás jelentősége és lehetőségei

Az elmúlt években számos kutatás indult a kompetenciák témakörében több tudományterületen is. A pszichológia, a neurológia és a közgazdaságtan is vizsgálja a témát, persze más-más megközelítésből. Közgazdaságtani szempontból, új kompetenciák kiépítése a humán erőforrásban befektetés, és mint olyan, természetes módon szeretnénk, ha ez a befektetés legalább megtérülne, de még inkább, hogy profitot termeljen. Munkáltatói szempontból ez a profit a vállalat versenyképességében, míg oktatáspolitikai szempontból az adott ország gazdasági teljesítőképességében nyilvánul meg. Az előzőekből adódik, hogy szeretnénk kontrollt gyakorolni a kompetenciákba fektetett erőforrásaink felett, és ezt a kontrollt legjobban szigorú és következetes tervezéssel tudjuk elérni.

Az oktatáspolitikusoknak döntéseiknél kifejezetten fontos figyelembe venni, hogy milyen tudásra és készségekre lesz szükség rövid- és középtávon a munkaerőpiacon, hogy miként hidalják át a „forradalmian átalakuló gazdaságban szükséges készségek és a munkakínálat között egyre növekvő” szakadékot (Fazekas, 2017, p. 6). Ezért a politikai, illetve oktatásügyben érintett döntéshozók számára felbecsülhetetlen információértékkel bírhat az, ha tudják, hogy hogyan változnak a kompetenciaigények. Ennek megfelelően a kompetenciák absztrakciós szintjétől függően az értekezésben javasolt keretrendszer segítségével különböző döntéshozói szintek támogathatók. A tézisben vázolt megoldás kifejezetten a tárgy- és szakfelelősök támogatását célozza, de egy magasabb szinten az eredmények a politikai döntéshozók számára is hasznosak lehetnek.

A felvázolni kívánt döntéstámogató infrastruktúra azonban nem áll rendelkezésre széleskörben elérhető módon, ami igazolja a dolgozat és a kutatás relevanciáját.

Wowczko (2015) kutatása alapján azt találta, hogy bár a munkaerőpiaci kereslet és kínálat összehangolása szempontjából elengedhetetlen a keresett készségek figyelembe vétele a tantervfejlesztés során, a korábbi kutatások mégis hajlamosak kizárólag a foglalkozások végzettségigényére koncentrálni, ami persze fontos, de nem elégséges. Jelen dolgozat tehát ennek a hiánynak a pótlására indított kutatás egyik lépése.

Bár a szak- és tárgyfelelősök a dolgozatban felvázolt megoldás elsődleges érintettjei, azonban a keretrendszer által nyújtott információknak a hallgatók közvetlenül is haszonélvezői lehetnek. A kinyert adatok alapján például tanulási utak (*learning path*) ajánlhatóak számukra. Ez többféleképpen is elképzelhető. Egyrészt amennyiben egy hallgató tudja, hogy milyen munkát szeretne végezni a jövőben, azaz egy adott pályát választ és afelé orientálódik, akkor kidolgozható számára egy olyan egyedi tanulási út, aminek követésével pontosan és célzottan azokat a készségeit fejlesztheti, illetve azt a tudást szerezheti meg, ami a vágyott karrierpályának leginkább megfelel. Másrészt, amennyiben fel tudjuk mérni egy adott szakon tanuló hallgató tudásában mutatkozó hiányosságokat, úgy tudunk számára nyújtani egy listát azokról az állásokról, amit jellemzően az adott szak elvégzése után hatékonyan be tudna tölteni, kiegészítve egy másik listával arról, hogy ahhoz, hogy egy adott pozíciót sikerrel meg tudjon pályázni, milyen kompetenciáit kell fejlesztenie, milyen tudáshiányokat kell pótolnia.

A munkavállalók számára ugyanilyen fontos lehet a kompetenciakereslet alakulásának ismerete, hiszen az élethosszig tartó tanulás zászlaja alatt, a rendkívül gyorsan változó követelmények között nekik is folyamatosan naprakészen kell tartaniuk a kompetenciakészletüket annak érdekében, hogy versenyképesek maradjanak a piacon. Mivel a konkrét anyagi kiadások mellett egy új kompetencia kiépítése jelentős haszonáldozati költséggel is jár, ezért nem mindegy, hogy az egyének mibe fektetik erőforrásaikat.

Összefoglalóan elmondható tehát, hogy a rossz vagy éppen jó befektetési döntések ezen a területen, a gazdaság összes szintjén, az egyéni, a vállalati és a nemzeti, illetve regionális versenyképességben, illetve teljesítményben is éreztetik hatásukat. A jó döntések elősegítéséhez a disszertációmban felvázolt keretrendszer felbecsülhetetlen információkat nyújthat.

2. Kutatásmódszertan és a kutatás eredményei

2.1. Az első kutatási kérdés vizsgálata során használt módszertan

Az első kutatási kérdés esetében a cél az adattárolási architektúra kiválasztását megalapozó szempontrendszer-, majd segítségével a legmegfelelőbb architektúramodell kidolgozása volt. A disszertációhoz kapcsolódó munka során elkezdtem az automatizált adatgyűjtést, kidolgoztam az adatoknak egy lehetséges modelljét, továbbá a kiválasztási szempontok részletes vizsgálatával ajánlást tettem a megvalósításra javasolt adattárolási architektúrára. Ebből a szempontból a kutatásom feltáró jellegű. Az igazoló jellegű kutatásokkal ellentétben, „a feltáró jellegű kutatások tipikusan három célból készülnek: a téma jobb megértését biztosítják, egy későbbi alaposabb kutatás megvalósíthatóságát tesztelik, és további kutatások számára fejlesztenek alkalmazható módszereket” (Varga, 2014, p. 4; Szabó, 2000).

Egy kisebb szoftverfejlesztési projekt keretében megvizsgáltam az adatgyűjtéshez elérhető eszközök kínálatát, és a *Scrapy* keresőrobot-rendszert választottam, ami egy *Python*ban íródott, nyílt forráskódú alkalmazás, amely biztosít minden infrastruktúrát a feladathoz. A felhasználónak csak az „üzleti logikát” kell leírnia *CSS Selector*ok vagy *XPATH* segítségével az úgynevezett *spider* osztályokban, azaz hogy a legyűjtött dokumentumok mely elemei mit jelentenek, és hogy milyen formában szeretné azokat tárolni.

Szintén az első kutatási kérdés vizsgálata során meghatároztam a tárolni kívánt adatok körét, melyekre a végső célok eléréséhez elengedhetetlenül szükség van, illetve felvázoltam egy lehetséges relációs sémát, amely egy relációs adattárolási megoldás választása esetén implementálható.

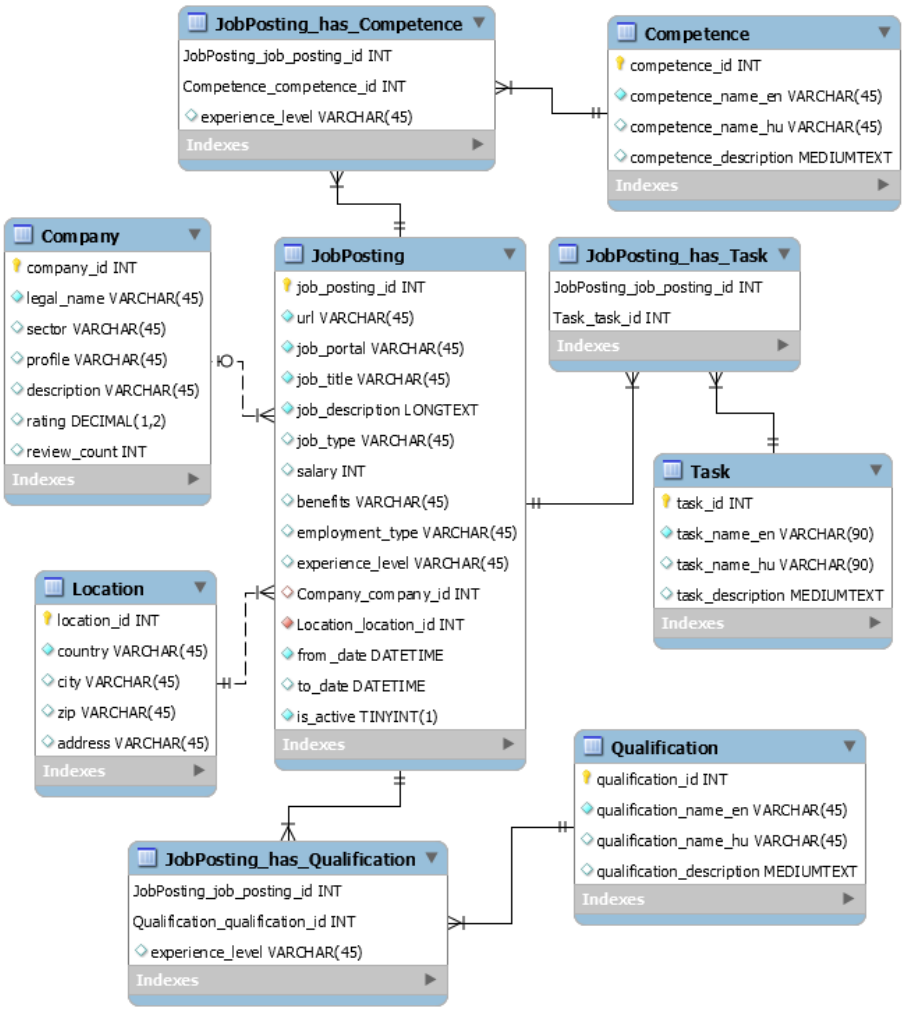
Az architektúraválasztási szempontrendszert szekunder kutatás, illetve szakirodalmi áttekintés segítségével dolgoztam ki. A legfontosabb megvizsgált szempontok között helyet kaptak klasszikus tulajdonságok, mint a sebesség, megbízhatóság, skálázhatóság, a megoldás költségei, a megvalósíthatóság és az elérhető támogatás. Ezeken kívül az megoldásokat összehasonlítottam a *CAP-tétel* (*consistency, availability, partition tolerance*) következményei, az adatok struktúrájának és rendelkezésre állásának

sajátosságaiból eredő igények alapján is. Megvizsgáltam továbbá az idősor-adatbázisok alkalmazhatóságának lehetőségeit is.

2.2. Az első kutatási kérdés eredményei

A kutatás során kidolgozott *scraper* forráskódja megtalálható a dolgozathoz kapcsolódó, https://github.com/gneusch/JobPostingScraper címen elérhető GitHub repozitóriumban. Az adatok gyűjtését az indeed.co.uk oldalról 2019 január 16-án kezdtem meg és egy év alatt közel 400 ezer hirdetést gyűjtöttem össze.

A tézis 5.2. fejezetében kimerítően listázom a tárolni kívánt adatok körét, forrását, és részletezem azt is, hogy milyen információigény kielégítésére szolgálnak. Az implementációtól függetlenül az adatok közötti kapcsolatok szemléltetésére bemutatom továbbá az adatoknak egy, a probléma szempontjából megfelelő relációs sémáját (1. ábra).

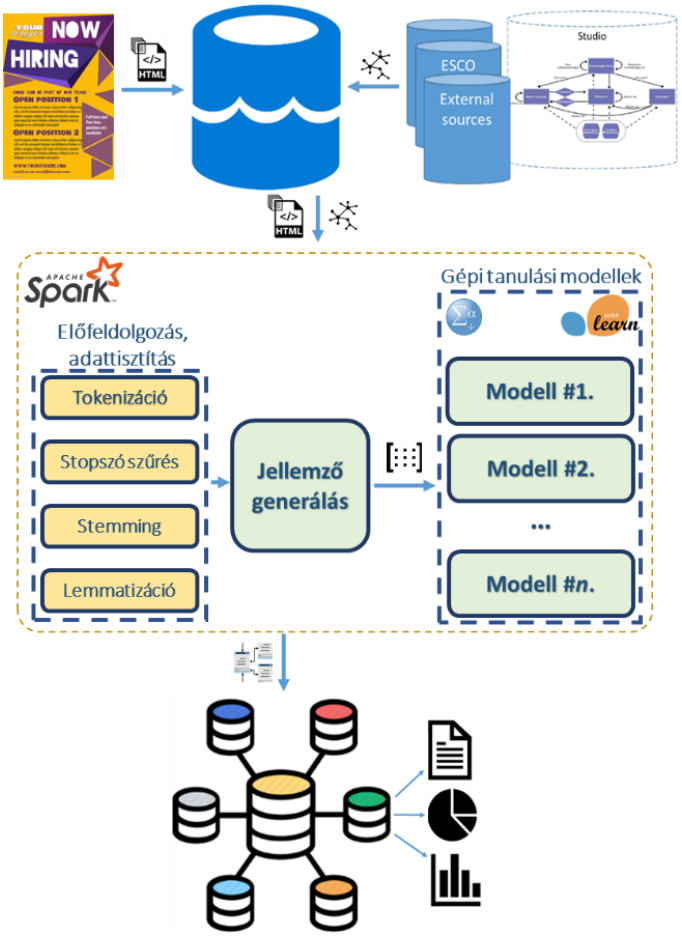


1. ábra: Az adattárház lehetséges logikai adatmodellje

Az adattárolási technológia legfontosabb kiválasztási és értékelési szempontjai a következők:

- A *scraping* folyamat eredményeként az adatok *JSON* formában állnak rendelkezésre, egyes attribútumok mentén strukturálva.
- Számos fontos, információértékkel bíró tartalmat azonban az álláshirdetések leírásaiból további feldolgozással szükséges kinyerni. Ezen feladat során mindenképpen szükséges az adatok tárolása egy előkészítő tárban.
- Az álláshirdetéseket, azok leírását feldolgozatlan formában továbbra is tárolni kívánom, hogy szükség esetén a jövőben is rendelkezésre álljanak. Továbbá az egyéb külső rendszerek felhasznált adatait is szeretném, hogy a feldolgozás egyszerűsítése és gyorsítása, illetve a folyamatosság biztosítása érdekében az adatbázisban is rendelkezésre álljanak.
- A rendszer sebességének szempontjából támasztott legfontosabb követelmény, hogy a felhasználók lekérdezései valós időben megválaszolhatóak legyenek, ezért szükséges a kompetenciák, foglalkozások stb. előzetes feltárása és tárolása, mely utóbbira jelen esetben egy relációs adatbázis alkalmasabb.
- Fontos, hogy a rendszer jól skálázható legyen, és magas rendelkezésre állást garantáljon. Az erős konzisztencia azonban nem elvárás.

Az előzőek fényében, figyelembe véve az ismertetett technológiák előnyeit és hátrányait, illetve a legfontosabb célt, a felhasználók információigényének lehető leggyorsabb kiszolgálását, egy hibrid architektúrára épülő megoldást hoznék létre, melynek fő alkotóelemeit a 2. ábra szemlélteti. Ezt a fajta hibrid megközelítést, amikor az adott probléma fényében a legmegfelelőbb, akár különböző megoldásokat egyszerre használó adatbázis-architektúrát építünk fel, Sadalage és Fowler (2012) *polyglot* perzisztencia néven hivatkozzák.



2. ábra: Hibrid tárolási megoldás sematikus architektúrája

Ebben az architektúrában az összegyűjtött álláshirdetések és a külső rendszerek adatai először egy dokumentumtárba kerülnek. Nincs szükség tehát a *scraper* által gyűjtött adatok előzetes átalakítására, azonnal, a letöltés pillanatában letárolhatók abban a részlegesen strukturált formában, amibe a *scraping* folyamat során kerülnek. Az előkészítő (*staging*) tár biztosításán kívül további előnye ennek a kialakításnak, hogy nagymértékű rugalmasság jellemzi, azaz ha például a kapott adatok struktúrája változik, vagy ha új adatköröket kívánunk bevonni a feldolgozásba stb., a keretrendszer ezen pontján nem szükséges változásokat eszközölnünk. Megőrizhetjük itt továbbá a külső rendszerekből származó- és az olyan adatokat is, melyeket jelenleg nem kívánunk használni, vagy a létrehozandó sémába nem illeszkednek, de adott esetben később hasznosak lehetnek.

A dokumentumtár azonban nem alkalmas arra, hogy a lehetséges felhasználói igényeket a lehető leggyorsabban kielégítsük, többek között azért sem, mert architekturálisan aggregátumok kezelésére, és nem az adatok különböző dimenziók mentén való szeletelésére és forgatására vannak felkészítve, illetve a szöveges formában tárolt

hirdetésleírások feldolgozása hosszadalmas és költséges, így érdemes annak eredményét eltárolni.

Ebből kifolyólag egy „integrációs” rétegben, az adattóra épülő *Spark* alkalmazásban tervezem a hirdetésleírások nyelvi feldolgozását, és gépi tanulási algoritmusok segítségével a szükséges információ kinyerését megvalósítani. Azért előnyös az architektúra ilyen kialakítása, mert további rugalmasságot biztosít. Az adatfeldolgozás így bármikor elvégezhető, illetve adott esetben újra végrehajtható, amennyiben a felhasznált adatok köre (például egy új kompetenciaszótárt vonunk be a folyamatba, vagy a kompetenciaként beazonosított kifejezéseket csatoljuk vissza), vagy a modellek változnak. Nemcsak az információk kinyeréséhez használt algoritmus változtatható így könnyűszerrel, de a felhasznált implementáció is. A disszertációban ismertetett módon, a gépi tanulási modelljeimet például két különböző eszközben, az *IBM SPSS Statistics*ben és a *Scikit-learn* programcsomagban építettem fel és teszteltem (6.1.4 és 7.1.1.2. alfejezetek). Mind a két kapott modell könnyűszerrel integrálható egy elosztott dokumentumtár adatain dolgozó párhuzamosított *Spark* alkalmazásba.

Ezután a feltárt információk immár strukturált formájukban kerülhetnek a 1. ábrán bemutatott relációs sémába egyszerű *SQL insert* és *update* utasítások segítségével. Azonban a legnagyobb méretű adattagot, a pozíció leírását (*job_description*), mivel a javasolt *NoSQL* adatbázisban hosszútávon megőrizni kívánom, a redundáns tárolást és a relációs adattárház horizontális skálázásának szükségességét elkerülendő, a felvázolt sémából kihagynám (ellentétben a 1. ábrával).

Ebben az architektúrában tehát biztosítható, hogy az egyes lekérdezések nagyon gyorsan, csak *SQL* utasításokra építve kiszolgálhatóak legyenek, ugyanakkor a nyers dokumentumok tovább tárolhatóak a dokumentumtárban. Így az is biztosított, hogy ha a jövőben változtatni kell a feldolgozási folyamaton, vagy a modelleken, vagy új információra van szükség a már archivált hirdetésleírásokból, minden adat rendelkezésre álljon.

2.3. A második kutatási kérdés vizsgálata során használt módszertan

A munkaerőpiaci adattárház legfontosabb céljaként fogalmaztam meg, hogy a hirdetések kompetenciatartalmáról, annak időbeli- és térbeli alakulásáról hasznos információkat szolgáltasson a döntéshozók számára. Azonban ezek a kompetenciák strukturált

formában nem állnak rendelkezésre az álláshirdetésekben, hanem azok leírásából kell ezt az információt a természetesnyelv-feldolgozás és a gépi tanulás témaköréhez kapcsolódó technikákkal kinyerni. Ennek megfelelően, a második kutatási kérdéssel összhangban a disszertáció 6. fejezetében olyan módszereket mutatok be, melyek alkalmasak lehetnek arra, hogy a strukturálatlan szövegben „minőségi kifejezéseket", ebben az esetben kompetenciajelölteket azonosítsunk be segítségükkel.

- A fejezet elején statisztikai alapon meghatároztam – a későbbi feldolgozás során használt – szóennesek (egymást követő szavak n hosszú szekvenciája) hosszát.
- Bemutattam azt, hogy miért nem, vagy csak nagyon korlátozottan alkalmasak a kifejezés-és dokumentumgyakoriságra (*tf, df*), illetve a tf-idf (*term frequency - inverse document frequency*) értékre épülő modellek a kompetenciajelöltek beazonosítására.
- Bemutattam továbbá egy külső ontológiák tartalma alapján felépített kompetenciaszótár használatában rejlő lehetőségeket az információt hordozó kifejezések feltárására. Továbbá részleteztem ennek a módszernek a határait is.
- A fejezet utolsó blokkjában szakirodalmi áttekintés alapján részletesen bemutattam 10 olyan metrikát (Jaccard index, Sørensen-Dice- és koszinusz távolság stb.), melyek kifejezések hasonlóságának, illetve távolságának meghatározására használhatóak. Ezek után számszerűsítettem az álláshirdetések leírásaiból képzett szóennesek és a kompetenciaszótár elemei közötti hasonlóságot.
- Az így kapott eredményeket magyarázó változóként használva egy logit modellben, arra kerestem a választ, hogy megadható-e a segítségükkel egy olyan egyenlet, amivel számszerűsíthető, hogy egy kifejezést elfogadhatunk-e kompetenciajelöltként. A modell felépítésére az *IBM SPSS Statistics* programcsomagot használtam. Mivel a modellparaméterek tesztelését a tanító halmazon végeztem el, és nem állt fent a „kukucskálás" (*data peeking*) veszélye, nem képeztem az adatokból külön validációra használt csoportot. Változószelekcióra a *Backward Wald* eljárást használtam, és azokat az eseteket fogadtam el találatnak, melyekre a becsült valószínűség 40%-nál nagyobb (*cut value = 0,4*)

2.4. A második kutatási kérdés eredményei

Az elfogadott modell globális mutatói megfelelőek. Az Omnibus teszt alapján minden szokásos szignifikancia szinten elfogadhatjuk az alternatív hipotézist, azaz biztos, hogy van olyan változó a modellben, melynek együtthatója szignifikáns. A pszeudo R^2 mutatók közepes determináltságot jeleznek. Cox és Snell mutatója alapján 45%-ban határozzák meg a magyarázó változók annak esélyét, hogy a kifejezés valós kompetenciaelemet azonosít, míg Nagelkerke R^2 mutatója alapján a determináltság 78,8%-os. De ezen mutatók közvetlen értelmezése félrevezető lehet, mert csak annyit „mondanak, hogy a csak konstanst tartalmazó modellhez tartozó log *likelihood* értéket hány százalékkal sikerült csökkenteni" (Kovács, 2014; Fliszár *et al.*, 2016, p. 46). Az irodalomban a megfelelőség vizsgálatára inkább a Hosmer-Lemeshow tesztet ajánlják. Ennek során a megfigyeléseket és a becsült valószínűségeket decilisekre osztjuk, és azt a hipotézist vizsgáljuk, hogy a ténylegesen bekövetkező események száma megegyezik-e az előrejelzettel az egyes decilisekre. Ezt a hipotézist jelen modellre elfogadhatjuk.

A disszertációban bemutatott *logit* modell a tesztadatokon elfogadhatóan teljesített; a felidézési arány 85%, míg a precizitás 71,9%. Mivel a folyamatba való manuális beavatkozás kezdetben elkerülhetetlen, azaz mielőtt elfogadhatnánk ezeket a kompetenciajelölteket valós kompetenciaként, egy szakértőnek át kell néznie az eredményeket, így a modell elfogadható, mint ami hasznos információkkal tud szolgálni, és hozzáadott értékkel bír. A modellt abból a szempontból is elfogadom, hogy megfelelő választ ad a második kutatási kérdésemre, és alkalmas az explicit megjelenő kompetenciák szignifikáns részének beazonosítására.

A fejezet összefoglalásában kiemeltem azokat az irányokat, melyekben a jövőben a kutatást folytatni szeretném. Ezek közül a legfontosabb a hirdetések szövegének előfeldolgozása által a zaj csökkentése a modell pontosságának növelése érdekében. Ebben az irányban azóta történtek is fejlesztések, melyek már felhasználásra kerültek a harmadik kutatási kérdés vizsgálata során végrehajtott kísérletekben is. (A hirdetések előfeldolgozását végző forráskód szintén megtalálható a dolgozathoz tartozó GitHub repozitóriumban.) A másik fontos irány, amerre a kutatással a jövőben haladni szeretnék, az a modell adaptálása magyar nyelvre is, hogy a keretrendszer hosszabb távon alkalmazható legyen hazai kompetenciakereslet elemzésére, annak területi összehasonlítására stb. is.

2.5. A harmadik kutatási kérdés vizsgálata során használt módszertan

A kutatás harmadik nagy blokkjában, a harmadik kutatási kérdéshez kapcsolódóan (7. fejezet) megvizsgáltam, hogy az álláshirdetésekben közvetlenül nem megjelenő, látens kompetenciák beazonosítására milyen módszereket és alkalmazásokat találok alkalmasnak. A legfontosabb irány, amit részletesen elemeztem, hogy az ESCO és az O*NET ontológiák tartalma segítségével az állásajánlatokban meghirdetett pozícióhoz kapcsolódó foglalkozás beazonosíthatósága.

- Ennek érdekében kidolgoztam egy reguláris kifejezésekre és egyszerű szabályokra alapuló módszert.
- Továbbá részletesen bemutattam egy döntési fára alapuló osztályozó modellt, melynek segítségével arra tettem kísérletet, hogy a hirdetéseket – címük alapján – a felhasznált ontológiákban leírt foglalkozásokhoz társítsam. Ehhez a feladathoz a döntési kritériumokat a hirdetések előkészítő lépések során megtisztított címének, és a felhasznált ontológiákban rögzített foglalkozás-megnevezéseknek a lexikográfiai és „kvázi-szemantikai” hasonlóságértékei szolgáltatták.
    - A modell építéséhez a *Scikit-learn* programcsomagot használtam.
    - Az adathalmaz kiegyensúlyozatlanságát – tekintve, hogy a döntési fa algoritmusok hajlamosak a domináns osztályok irányában elfogult, hibás modellek létrehozására – kísérleteim során a kisebbségi osztályba tartozó megfigyelések felül-mintavételezésével kívántam ellensúlyozni. Megvizsgáltam az egyszerű véletlenszerű helyettesítéses-, és a *SMOTE* felül-mintavételezési algoritmusok teljesítményét.
    - Szintén az adatok kiegyensúlyozatlansága miatt a tanító- és a teszthalmazokat rétegzett mintavételezéssel, keveréssel állítottam elő.
    - A megfigyelések relatíve alacsony száma miatt a modellparaméterek finomhangolására és tesztelésére, illetve az egyes modellek predikciós hibájának becslésére a keresztvalidáció technikáját választottam a különálló tanító, validációs és teszthalmazok használata helyett.
    - A túlillesztés, azaz a tanítóhalmazra, annak sajátosságaira, hibáira való túlzott rátanulás elkerülése és a lehető legpontosabb modell megtalálása érdekében a mélység, a minimális levelenkénti mintaszám és a hibaarány-

komplexitási metszés effektív α paraméterének különböző értékei mentén kerestem az elérhető legjobb modellt, azaz azt a döntési fát, mellyel maximalizálható a keresztvalidáció során elért eredmény.

- A *t-SNE* módszer segítségével igazoltam Csepregi (2020) eredményeit, miszerint álláshirdetések klaszterei (témacsoportok) nem, vagy csak nagyon korlátozottan, kis hatásfokkal beazonosíthatóak a leírásukból készült *tf-idf* mátrix alapján.
- A disszertáció 7. fejezetének lezárásaként bemutattam olyan módszereket, melyek segítségével a látens igények, az explicit megjelenő kompetenciák kapcsolatain keresztül tárhatóak fel. Ezeket az irányokat az értekezésben részletesen nem vizsgáltam, azonban további kutatásaim során mindenképpen érdemesnek tartom őket górcső alá venni.

2.6. A harmadik kutatási kérdés eredményei

A disszertáció harmadik kutatási kérdéséhez kapcsolódóan a hirdetésekben explicit nem megjelenő, implicit vagy látens kompetenciák beazonosításának lehetőségeit vizsgáltam. Koncepcionálisan három nagyobb irányt vázoltam fel. Az első a hirdetésekhez kapcsolódó foglalkozás beazonosításán keresztül, az annak kontextusában – a felhasznált külső ontológiák alapján – releváns kompetenciák elfogadása, mint amelyek implicit szükségesek az adott pozíció megfelelő ellátásához.

- A reguláris kifejezésekre és egyszerű szabályokra épülő módszerrel a 2019 októberi álláshirdetések 23,7%-át tudtam foglalkozáshoz kapcsolni, 97,5%-ban helyesen.
- A döntési fákkal végzett kísérletek során a legjobbként elfogadott modell teszthalmazon mért precizitása 58%, míg felidézési aránya 68%, ami bár nem kiemelkedő, de felveszi a versenyt az irodalomban fellelhető, hasonló feladatra megalkotott modellekkel. Például Amato és szerzőtársai (2015) több módszerrel vizsgálták foglalkozások beazonosíthatóságát hirdetések címében. Bár a módszerek használatát a szerzők mélységükben nem részletezik, de a közölt eredmények alapján elmondható, hogy a felidézés és a precizitás átlagosan az LDA (*Latent Dirichlet Allocation*) esetében 50% körül, míg a többi módszer (lineáris tartóvektor-gép, perceptron osztályozó) esetében 25-35% között alakult.

A modelleket a jövőben nagyobb elemszámú adathalmazzal tanítva, illetve más módszereket is felhasználva, a kísérletet megismételni kívánom, hiszen az így beazonosított foglalkozások alapján az ontológiákban hozzájuk kapcsolódó kompetenciák a hirdetés szempontjából implicit relevánsként elfogadhatóak. A harmadik kutatási kérdés szempontjából ezt az irányt egy lehetséges megoldásként elfogadom, azaz az implicit kompetenciák a hirdetésekben kijelölt foglalkozások beazonosításának segítségével feltárhatóak, azonban ennek a beazonosítási folyamatnak a hatékonysága javításra szorul a jövőben.

A második nagy, vizsgált irányt azonban, hogy a hirdetések tartalmuk alapján csoportokba sorolhatóak, mely csoportok alapján következtetni lehet az átfedő kompetenciatartalomra, vagy valamiféle foglalkozási csoportra, elvetettem. Ezt a döntést Csepregi (2020) és saját *t-SNE* módszerrel végzett kísérletem alapján hoztam meg.

A 7. fejezet utolsó részében bemutattam a 3. lehetséges irányt a látens igények feltárására, az explicit megjelenő kompetenciák ontológiakapcsolatainak segítségével. A mester tézisemben elért eredményeim alapján ezt az irányt szintén el tudom fogadni, mint ami alkalmas a látens igények beazonosítására, ahogy azt a harmadik kutatási kérdésben felvetettem. Így a kutatás jövőbeli szakaszaiban ezt a területet is fejleszteni kívánom.

3. Hivatkozások

Amato, F., Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M., Moscato, V., Persia, F., *et al.* (2015), "Classification of Web Job Advertisements: A Case Study", *23rd Italian Symposium on Advanced Database Systems (SEBD 2015)*, presented at the 23rd Italian Symposium on Advanced Database Systems (SEBD 2015) (Gaeta, 14/06/2015 - 17/06/2015), Curran, Gaeta, Italy, pp. 144–151, available at: http://hdl.handle.net/10863/10468.

Beck, M. and Libert, B. (2017), "The Rise of AI Makes Emotional Intelligence More Important", *Harvard Business Review*, 15 February, available at: https://hbr.org/2017/02/the-rise-of-ai-makes-emotional-intelligence-more-important (accessed 4 May 2019).

Bodon, F. (2010), "Adatbányászati algoritmusok", Dr. Bodon Ferenc, available at: http://www.cs.bme.hu/~bodon/magyar/adatbanyaszat/tanulmany/adatbanyaszat.pdf (accessed 1 October 2019).

Bodon, F. and Buza, K. (2013), "Adatbányászat", Elektronikus tananyag, available at: http://www.cs.bme.hu/~buza/pdfs/adatbanyaszat-cover.pdf (accessed 20 December 2020).

Chang, H.-C., Wang, C.-Y. and Hawamdeh, S. (2018), "Emerging trends in data analytics and knowledge management job market: extending KSA framework", *Journal of Knowledge Management*, Vol. 23 No. 4, pp. 664–686, available at: https://doi.org/10.1108/JKM-02-2018-0088.

Choi, S., Cha, S. and Tappert, C.C. (2010), "A Survey of Binary Similarity and Distance Measures", *Journal of Systemics, Cybernetics and Informatics*, Vol. 8 No. 1, pp. 43–48.

Csepregi D. (2020), *Első Lépés az Automatizált Oktatásfejlesztés Felé: Hogyan tehet szert előnyre szövegelemzési eszközökkel a Corvinus Gazdaságinformatikus képzése?*, TDK dolgozat, Budapesti Corvinus Egyetem, Gazdálkodástudományi kar, Budapest, available at: http://publikaciok.lib.uni-corvinus.hu/publikus/tdk/csepregi_d_2020a.pdf (accessed 27 December 2020).

Fajszi, B. and Cser, L. (2004), *Üzleti Tudás Az Adatok Mélyén*, Budapesti Műszaki és Gazdaságtudományi Egyetem.

Fajszi, B., Cser, L. and Fehér, T. (2010), *Üzleti Haszon Az Adatok Mélyén*, Alinea Kiadó - IQSYS Informatikai és Tanácsadó Zrt.

Fazekas, K. (2017), *Nem Kognitív Készségek Kereslete És Kínálata a Munkaerőpiacon*, Institute of Economics, Centre for Economic and Regional Studies, Hungarian Academy of Sciences, Budapest, available at: http://www.mtakti.hu/wp-content/uploads/2017/11/FK-BWP1709-jav-OE_FKjav.pdf (accessed 20 July 2019).

Fliszár, V., Kovács, E., Szepesváry, L. and Szüle, B. (2016), *Többváltozós Adatelemzési Számítások*, Budapesti Corvinus Egyetem, available at: http://unipub.lib.uni-corvinus.hu/2438/.

Gábor, A., Kő, A., Szabó, Z. and Fehér, P. (2016), "Corporate Knowledge Discovery and Organizational Learning: The Role, Importance, and Application of Semantic Business Process Management—The ProKEX Case", in Gábor, A. and Kő, A. (Eds.), *Corporate Knowledge Discovery and Organizational Learning*, Springer International Publishing, pp. 1–31, available at: https://doi.org/10.1007/978-3-319-28917-5_1.

Gajdos, S. (2019), *Adatbázisok*, A 2015. évi kiadás negyedik javított utánnyomása., BME, Budapest, Magyarország.

Gruber, T.R. (1993), “A Translation Approach to Portable Ontology Specifications”, *Knowledge Acquisition*, Vol. 5(2), pp. 199–220, available at: http://tomgruber.org/writing/ontolingua-kaj-1993.htm.

Grundke, R., Marcolin, L., Nguyen, T.L.B. and Squicciarini, M. (2018), “Which skills for the digital era?”, *OECD Science, Technology and Industry Working Papers*, available at: https://doi.org/10.1787/9a9479b5-en.

Gugnani, A. and Misra, H. (2020), “Implicit Skills Extraction Using Document Embedding and Its Use in Job Recommendation”, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34 No. 08, pp. 13286–13293, available at: https://doi.org/10.1609/aaai.v34i08.7038.

Kovács, E. (2014), *Többváltozós Adatelemzés*, Typotex Kiadó.

Manning, C.D., Raghavan, P. and Schütze, H. (2009), *Introduction to Information Retrieval*, Online edition., Cambridge University Press, New York.

Nasir, S.A.M., Yaacob, W.F.W. and Aziz, W.A.H.W. (2020), “Analysing Online Vacancy and Skills Demand using Text Mining”, *Journal of Physics: Conference Series*, IOP Publishing, Vol. 1496, p. 12, available at: https://doi.org/10.1088/1742-6596/1496/1/012011.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., *et al.* (2011), “Scikit-learn: Machine Learning in Python”, *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830, available at: https://dl.acm.org/doi/10.5555/1953048.2078195.

Pitukhin, E., Varfolomeyev, A. and Tulaeva, A. (2016), “JOB ADVERTISEMENTS ANALYSIS FOR CURRICULA MANAGEMENT: THE COMPETENCY APPROACH”, *ICERI2016 Proceedings*, presented at the 9th annual International

Conference of Education, Research and Innovation, IATED, Seville, Spain, pp. 2026–2035, available at: https://doi.org/10.21125/iceri.2016.1456.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. and Manning, C.D. (2020), "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, presented at the ACL, Association for Computational Linguistics, Online, pp. 101–108, available at: https://doi.org/10.18653/v1/2020.acl-demos.14.

Russell, S. and Norvig, P. (2005), *Mesterséges Intelligencia Modern Megközelítésben*, Panem Kft., Budapest, available at: https://mialmanach.mit.bme.hu/aima/index (accessed 22 April 2017).

Sadalage, P.J. and Fowler, M. (2012), *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*, Addison-Wesley, available at: https://dl.acm.org/doi/book/10.5555/2381014.

Szabó Z. (2000), *A szervezeti információfeldolgozás strukturális és technológiai tényezőinek összerendelése*, phd, Budapesti Corvinus Egyetem, available at: http://phd.lib.uni-corvinus.hu/212/ (accessed 24 January 2021).

Varga, K. (2014), *A Szemantikus Folyamatmenedzsment Hasznosítási Lehetősége Az Üzleti Folyamatok Tudásalapú Fejlesztésében*, PhD thesis, Budapesti Corvinus Egyetem, available at: http://phd.lib.uni-corvinus.hu/818/ (accessed 13 September 2019).

Vas, R.F. (2007), *Tudásfelmérést Támogató Oktatási Ontológia Szerepe És Alkalmazási Lehetőségei*, PhD thesis, Budapesti Corvinus Egyetem, available at: http://phd.lib.uni-corvinus.hu/258/.

Waldrop, M.M. (2016), “The chips are down for Moore’s law”, *Nature News*, Vol. 530 No. 7589, pp. 144–147, available at: https://doi.org/10.1038/530144a.

Wowczko, I.A. (2015), “Skills and Vacancy Analysis with Data Mining Techniques”, *Informatics*, Vol. 2 No. 4, pp. 31–49, available at: https://doi.org/10.3390/informatics2040031.

Zhao, M., Javed, F., Jacob, F. and McNair, M. (2015), “SKILL: a system for skill identification and normalization”, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI Press, Austin, Texas, pp. 4012–4017, available at: https://dl.acm.org/doi/10.5555/2888116.2888273.

4. Publikációk

Referált szakmai folyóirat

Neusch, G. (2014), “Domain Ontology Tailoring Based on Business Processes in the Frame of the ProKEX Project”, *SEFBIS Journal*, Vol. I No. XI, pp. 51–59.

Szabó, I. and Neusch, G. (2015), “Dynamic Skill Gap Analysis Using Ontology Matching”, in Kő, A. and Francesconi, E. (Eds.), *Electronic Government and the Information Systems Perspective*, Vol. 9265, Springer International Publishing, pp. 231–242, available at: http://dx.doi.org/10.1007/978-3-319-22389-6_17.

Beel, J., Carevic, Z., Schaible, J. and Neusch, G. (2017), “RARD: The Related-Article Recommendation Dataset”, *D-Lib Magazine*, Vol. 23 No. 7/8, available at: https://doi.org/10.1045/july2017-beel.

Szabó, I., Neusch G., Vas R. (2021, megjelenés alatt), „Design Thinking based Ontology Development for Robo-Advisors”, *Proceedings of the 20th International Conference Intelligent Systems Design and Applications (ISDA 2020)*, Advances in Intelligent Systems and Computing, Springer Verlag

Lektorált konferenciakötetben megjelent tanulmányok

Neusch, G. and Gábor, A. (2014), "ProKEX – Integrated Platform for Process-Based Knowledge Extraction", in Gómez Chova, L., López Martínez, A. and Candel Torres, I. (Eds.), *ICERI2014 Proceedings*, presented at the 7th International Conference on Education Research and Innovation, IATED Academy, Seville, Spain.

Castello, V., Mahajan, L., Flores, E., Gabor, M., Neusch, G., Szabo, I., Guerrero, J., *et al.* (2014), "THE SKILL MATCH CHALLENGE. EVIDENCES FROM THE SMART PROJECT", in Gómez Chova, L., López Martínez, A. and Candel Torres, I. (Eds.), *ICERI2014 Proceedings*, IATED Academy.

Weber, C., Neusch, G. and Vas, R. (2016), "Studio: A Domain Ontology Based Solution for Knowledge Discovery in Learning and Assessment", *Proceedings of the 2016 AIS SIGED International Conference on Information Systems Education and Research*, pp. 1-13., available at: https://aisel.aisnet.org/siged2016/12.

Neusch, G. (2016), "Ontology Tailoring for Job Role Knowledge", in Gábor, A. and Kő, A. (Eds.), *Corporate Knowledge Discovery and Organizational Learning*, Springer International Publishing, pp. 105–130, available at: https://doi.org/10.1007/978-3-319-28917-5_5.

Egyéb szakmai teljesítmények

Neusch, G., 2015. 'Studio-User Help' saját számítógépi programalkotás, Szellemi Tulajdon Nemzeti Hivatala, Önkéntes műnyilvántartásba vételi szám: 003871