Óbudai Egyetem

Neumann János Informatikai Kar

Alkalmazott Informatikai Intézet

SZAKDOLGOZAT

FELADATLAP

Hallgató neve: **Lovas István**

Törzskönyvi száma: T/002145/FI12904/N

A dolgozat címe:

**Szoftver követelmények feltárását, analizálását és modellezést támogató intelligens eszköz**

**Intelligent Tool to Support Software Requirements Elicitation, Analysis and Modelling**

Intézményi konzulens: Dr. Tick József

Külső konzulens:

Beadási határidő: 2015. január 7.

A záróvizsga tárgyai: Számítógép architektúrák

Szoftverfejlesztés

**A feladat**

Készítsen szakdolgozatot, melyben bemutatja a szoftver követelménytervezés, illetve a követelmények modellezésének általános megközelítéseit. A szakirodalom alapján ismertesse az alkalmazott elveket, megoldásokat és azon eszközöket, melyek ezt a tevékenységet támogatják. Tervezzen meg, és valósítson meg egy olyan elosztott szoftver-eszközt, mely segíti az akár csoportos követelmény feltárást, követelmény analízist, a követelmények modellezését, összeköti a specifikáció szöveges részeit, a modellreprezentációk különböző absztrakcióival. Az eszköz könnyítse meg az ezek közti navigációt, illetve nyújtson intelligens módon segítséget a modellek elkészítésében, a szövegekből kinyert információk alapján. Az eszköz fejlesztése során alkalmazza a szoftvertechnológia modern megoldásait.

**A dolgozatnak tartalmaznia kell**:

* A feladat részletes leírását.
* Követelménytervezés elterjedt megközelítéseinek bemutatását a szakirodalom alapján.
* A kapcsolódó szövegbányászati eljárások szakirodalom segítségével történő bemutatását.
* A rendszer tervét, UML alkalmazásával.
* Az eszköz megvalósításának leírását
* A rendszer tesztelését, és annak eredményeit.
* Az eszköz továbbfejleszthetőségi lehetőségeit.

Ph.

.....……………….

Dr. Galántai Aurél

mb. intézetigazgató

A szakdolgozat elévülésének határideje: **2017. január 7.**

(OE TVSz 32.§ szerint)

A dolgozatot beadásra alkalmasnak tartom:

|  |  |
| --- | --- |
| ……………….. | ..……………………. |
| külső konzulens | intézményi konzulens |

# Tartalomjegyzék

[**Szoftver követelmények feltárását, analizálását és modellezést támogató intelligens eszköz** 1](#_Toc402464835)

[1 Tartalomjegyzék 3](#_Toc402464836)

[2 Ábrajegyzék 8](#_Toc402464837)

[3 Célkitűzések 9](#_Toc402464838)

[4 Irodalomkutatás 10](#_Toc402464839)

[4.1 Bevezetés 10](#_Toc402464840)

[4.2 A Követelménytervezés alapfogalmai 10](#_Toc402464841)

[4.2.1 Célkitűzések 10](#_Toc402464842)

[4.2.2 Szoftverkövetelmény 10](#_Toc402464843)

[4.2.3 Követelmények csoportosításai 11](#_Toc402464844)

[4.2.4 Felhasználói követelmények 11](#_Toc402464845)

[4.2.5 Rendszerkövetelmények 11](#_Toc402464846)

[4.2.6 Rendszer 12](#_Toc402464847)

[4.2.7 Funkcionális követelmények 12](#_Toc402464848)

[4.2.8 Nemfunkcionális követelmények 12](#_Toc402464849)

[4.2.9 Szakterületi követelmény 13](#_Toc402464850)

[4.2.10 Szükséges és javasolt követelmények 14](#_Toc402464851)

[4.2.11 A követelménytervezés folyamata 14](#_Toc402464852)

[4.2.12 Szoftverkövetelmények dokumentuma 14](#_Toc402464853)

[4.2.13 Verifikáció és Validáció 14](#_Toc402464854)

[4.2.14 Fogalomszótár 15](#_Toc402464855)

[4.3 A követelmények fontossága 15](#_Toc402464856)

[4.4 Követelménytervezés folyamata 15](#_Toc402464857)

[4.4.1 Megvalósíthatósági tanulmány elkészítése 15](#_Toc402464858)

[4.4.2 Követelmény feltárás és elemzés 16](#_Toc402464859)

[4.4.3 Követelmény specifikációs folyamat 16](#_Toc402464860)

[4.4.4 Követelmény validáció 16](#_Toc402464861)

[4.4.5 Követelmények kezelése 16](#_Toc402464862)

[4.5 A követelmény feltárás, és kezelés problematikái 16](#_Toc402464863)

[4.5.1 A természetes nyelv problémái 16](#_Toc402464864)

[4.5.2 A természetes nyelv problémáit kiküszöbölendő megfontolások 17](#_Toc402464865)

[4.6 A szoftveréletciklus modellek és a követelmények kapcsolata 18](#_Toc402464866)

[4.6.1 A hagyományos megközelítések problémái 18](#_Toc402464867)

[4.6.2 A manapság gyakori módszertanok és a követelmények kapcsolata. 18](#_Toc402464868)

[4.7 Követelménymodellezés 20](#_Toc402464869)

[4.7.1 Az UML rövid bemutatása 20](#_Toc402464870)

[4.8 Követelmény modellezés eszközei 20](#_Toc402464871)

[4.8.1 Use Case modell 20](#_Toc402464872)

[4.9 Az agilis követelménykezelés 20](#_Toc402464873)

[4.9.1 User Stories: A használati esetek agilis megközelítése 20](#_Toc402464874)

[4.10 Szakirány vezéreltség 20](#_Toc402464875)

[4.10.1 A TDD, DDD, BDD hármas értelmezése 20](#_Toc402464876)

[4.10.2 Szakirány vezérelt tervezés (Domain Driven Design) 20](#_Toc402464877)

[4.10.3 Girkin nyelven leírt követelmények 20](#_Toc402464878)

[4.11 Követelmény formátum sablonok és ajánlások 20](#_Toc402464879)

[4.11.1 Felhasználói követelmény sablon 20](#_Toc402464880)

[4.11.2 Rendszerkövetelmény sablon 20](#_Toc402464881)

[4.12 A CASE eszközök 20](#_Toc402464882)

[4.13 A szövegbányászatról általánosságában 21](#_Toc402464883)

[4.14 Szövegbányászati előfeldolgozás 22](#_Toc402464884)

[4.15 Az előfeldolgozás közben fontos alapvető dokumentumjellemzők 22](#_Toc402464885)

[4.15.1 A dokumentumot hordozó médium 22](#_Toc402464886)

[4.15.2 A dokumentum elérési helye 22](#_Toc402464887)

[4.15.3 A dokumentum mérete 23](#_Toc402464888)

[4.15.4 A dokumentum statisztikai jellemzői 23](#_Toc402464889)

[4.15.5 A dokumentummal kapcsolatos metaadatok 23](#_Toc402464890)

[4.15.6 A dokumentum formátuma 23](#_Toc402464891)

[4.15.7 A dokumentum karakterkódolása 24](#_Toc402464892)

[4.16 Szövegbányászati modellalkotás 25](#_Toc402464893)

[4.16.1 Az ismertebb dokumentum reprezentációs modellek bemutatása 25](#_Toc402464894)

[4.16.2 A vektortér modell 25](#_Toc402464895)

[4.16.3 A szó-dokumentum mátrix jellemző súlyozási sémái 26](#_Toc402464896)

[4.16.4 Módok a szöveg felbontására, illetve a szótár felépítésére 27](#_Toc402464897)

[4.16.5 A strukturális szegmentálás 27](#_Toc402464898)

[4.16.6 Mondatokra bontás 28](#_Toc402464899)

[4.16.7 Tokenizálás 28](#_Toc402464900)

[4.16.8 Stoppszószűrés 29](#_Toc402464901)

[4.16.9 Lemmatizálás és Szótövezés 30](#_Toc402464902)

[4.17 Információkinyerés 30](#_Toc402464903)

[4.17.1 Az információkinyerés összehasonlítása más főbb feladattípusokkal 31](#_Toc402464904)

[4.18 Az információkinyerés fontosabb részfeladatai 31](#_Toc402464905)

[4.18.1 A névelem felismerés 32](#_Toc402464906)

[4.18.2 A szereplők azonosítása 32](#_Toc402464907)

[4.18.3 A kereszthivatkozások azonosítása 32](#_Toc402464908)

[4.18.4 A szereplők közti relációk azonosítása 33](#_Toc402464909)

[4.18.5 Események felfedése és illesztése részfeladat 33](#_Toc402464910)

[4.19 Az információkinyerés során előforduló főbb nyelvészeti problémák 33](#_Toc402464911)

[4.20 A szabály és a statisztika alapú információkinyerési megközelítések összevetése 34](#_Toc402464912)

[4.20.1 A statisztikai megközelítések által használt elterjedt reprezentációs modellek 35](#_Toc402464913)

[4.21 Alternatív szövegelemzési megfontolások 35](#_Toc402464914)

[4.22 Az irodalomkutatás alapján leszűrt tanulságok 35](#_Toc402464915)

[4.22.1 Milyen lehet egy jó és hasznos követelményelemzést támogató eszköz. 35](#_Toc402464916)

[4.22.2 Az irodalomkutatás alapján milyen egy jó, és modern CASE eszköz 35](#_Toc402464917)

[4.22.3 A megismert szövegbányászati feladattípusok közül melyek, és milyen formában lehetnek relevánsak egy CASE eszközben 36](#_Toc402464918)

[5 A rendszerrel szemben támasztott követelmények 36](#_Toc402464919)

[5.1 Vázlatos felhasználói követelmények, illetve igények. 36](#_Toc402464920)

[5.2 A rendszerben megjelenő felhasználói szerepkörök 38](#_Toc402464921)

[5.2.1 A rendszer felhasználóinak használati eset nézete 38](#_Toc402464922)

[5.2.2 A rendszerben megjelenő, előre definiált felhasználói szerepkörök jellemzése 38](#_Toc402464923)

[5.3 Funkcionális követelmények 40](#_Toc402464924)

[5.3.1 A rendszer főbb felhasználói funkcióinak áttekintő használati eset nézete 40](#_Toc402464925)

[5.4 Nem funkcionális követelmények 41](#_Toc402464926)

[6 A rendszer kialakítása során felhasználható technológiák és jellemzésük 42](#_Toc402464927)

[6.1 Szerver oldali technológiák 42](#_Toc402464928)

[6.1.1 ASP.NET 42](#_Toc402464929)

[6.1.2 Java 42](#_Toc402464930)

[6.2 A kliens oldali megjelenítés kezelése 42](#_Toc402464931)

[6.2.1 HTML5 42](#_Toc402464932)

[6.2.2 CSS 42](#_Toc402464933)

[6.3 A kliens oldali dinamika kezelése 42](#_Toc402464934)

[6.3.1 JavaScript, és a TypeScript 42](#_Toc402464935)

[6.4 Kommunikációs technológiák 42](#_Toc402464936)

[6.5 Szövegbányászati csomagok 42](#_Toc402464937)

[7 Követelményanalízis 43](#_Toc402464938)

[7.1 A rendszer statikus modelljének elemzése 43](#_Toc402464939)

[7.1.1 Szerkezet áttekintő nézet 43](#_Toc402464940)

[7.1.2 Szerkezeti nézet 44](#_Toc402464941)

[7.2 A rendszer dinamikájának és folyamatainak elemzése 44](#_Toc402464942)

[7.3 Képernyő vázlatok 45](#_Toc402464943)

[8 Rendszerterv 45](#_Toc402464944)

[8.1 Képernyő tervek 45](#_Toc402464945)

[8.2 Csomagáttekintő nézet 45](#_Toc402464946)

[8.3 Részletes komponens leírások 46](#_Toc402464947)

[8.4 Részletes dinamikus nézetek 52](#_Toc402464948)

[8.5 Telepítési nézet és rendszerkörnyezet 56](#_Toc402464949)

[9 Az implementáció részletei 57](#_Toc402464950)

[10 Tesztelés 63](#_Toc402464951)

[10.1 Statikus tesztelés 63](#_Toc402464952)

[10.2 Dinamikus tesztelés 64](#_Toc402464953)

[10.2.1 Egység teszt 64](#_Toc402464954)

[10.2.2 Komponens teszt 65](#_Toc402464955)

[10.2.3 Integrációs teszt 66](#_Toc402464956)

[10.2.4 Rendszer teszt 67](#_Toc402464957)

[10.2.5 Átadási teszt 68](#_Toc402464958)

[11 Továbbfejlesztési lehetőségek 69](#_Toc402464959)

[12 A szakdolgozat tartalmi összefoglalója 70](#_Toc402464960)

[13 Irodalomjegyzék 72](#_Toc402464961)

[14 Mellékletek 73](#_Toc402464962)

# Ábrajegyzék

# Célkitűzések

Célunk egy olyan eszköz kifejlesztése, ami képes segíteni a szoftverfejlesztés egyes magas absztrakciós tevékenységeit. Gondolok itt főképp felhasználók által elkészítendő szoftver-rendszer, követelményeinek specifikációjával, analízisével, illetve modellezésével kapcsolatos teendőkre. A készítendő rendszernek célja továbbá, hogy támogassa a csoportos munkát, lehetőleg úgy, hogy a különböző felhasználók valós időben láthassák egymás munkáit, és közben képesek legyenek kommunikálni is egymással, a rendszert felhasználva. Végül, de nem utolsó sorban a rendszer legyen a lehető legkülönbözőbb környezetekből, és platformokról elérhető, és használható.

# Irodalomkutatás

## Bevezetés

Az elkövetkezendő szakaszokban, a feldolgozott irodalmak alapján bemutatom a szoftverkövetelményeket, a követelménytervezés folyamatát, kitérek a követelmények modellezésére, a szakirány központú szemléletekre, és az agilis követelménykezelésre. Majd röviden írok az egyes szoftver életciklus modellekről, illetve a CASE eszközökről. Ezután a szövegbányászat alapvető koncepcióit mutatom be, és kitérek néhány alternatív, egyszerű, a szoftverkövetelményekkel kapcsolatos információkinyerési megközelítésre.

## A Követelménytervezés alapfogalmai

Ebben a szakaszban igyekszem vázolni, a követelményekkel kapcsolatos összes fogalmat. Egyes fontosabb fogalmakra a későbbiekben külön fejezetben részletesebben kitérek. Nem foglalkozok ellenben a kritikus rendszerek speciális igényeivel, illetve a formális specifikációk mélyebb elemzésével, mert ezek a dolgozat hatókörén kívül esnek.

### Célkitűzések

Érdemes a project megkezdése elején egy rövid áttekintést adni, a készítendő szoftver rendszer alapvető céljairól, betöltendő szerepéről. Ezek sokkal inkább alapvető iránymutatásként szolgálnak, mint tényleges követelményként. Ezek lehetnek általános víziók a szoftver hasznosságáról, főbb feladatairól, vázolhatja a főbb használati szituációkat, és környezetet. Ez a legáltalánosabb vázlat, amit a rendszerről készítünk. Érdemes lehet itt megemlíteni, hogy a felhasználók munkáját a termékünk miképpen fogja megkönnyíteni, milyen fontosabb értékeket hordoz a megrendelőknek.

### Szoftverkövetelmény

A rendszer követelményei lehetnek a rendszer funkcionalitására vonatkozó elvárások magas és-vagy alacsonyabb szintű, részletesebb megfogalmazása, illetve a rendszerrel szemben támasztott megszorítások. Megmondja, hogy a rendszer a megrendelő problémáit milyen szolgáltatások révén oldja meg, és megoldás közben milyen általános és probléma centrikus megkötésekre kell odafigyelni. Részletesség szerint két csoportját érdemes megkülönböztetni, a felhasználói-, illetve a rendszerkövetelményeket. Ezeket a csoportokat szoktuk vegyíteni is tehát léteznek a funkcionális felhasználói követelmények, a nem-funkcionális felhasználói követelmények, és a rendszerkövetelmények is lehetnek funkcionálisak, illetve nem-funkcionálisak is. Ezeket az eltérő követelményeket fontos lehet elkülöníteni egymástól, akár jelöléssel, akár úgy, hogy a követelmény dokumentáció külön alfejezeteit alkotják az imént vázolt csoportok, ellenben jó ötlet lehet, az összekapcsolódó, de különböző csoportba sorolt követelmények gyors összevetésére is módot adni, de legalább egy hivatkozást elhelyezni, a kapcsolódó követelményre.

### Követelmények csoportosításai

Követelmények csoportosítása történhet a megfogalmazásuk mélysége szerint. Ugyanis a gyakorlatban fontos szétválasztani a követelmények azon szintjét, mely a megrendelőkkel, illetve a felhasználókkal történő egyeztetést segítik, ezek a felhasználói követelmények, és azt a szintet, melyben a rendszer követelményeit részletezzük, ezeket nevezzük rendszerkövetelményeknek. A követelmények egy másik csoportosítása történhet az alapján, hogy az adott követelmény egy a rendszertől elvárt szolgáltatást részletez, azaz funkcionális követelmény, vagy egy a rendszertől, esetleg annak egyes szolgáltatásaitól elvárt tulajdonságot ír le, tehát nem-funkcionális követelmény.

### Felhasználói követelmények

Absztrakt módon leírja a rendszertől elvárt szolgáltatásokat, külső viselkedését, és azok működési megszorításait. Közérthető, természetes nyelvű leírás, melyet a rendszer felhasználójának, a megrendelőnek, és ezek megbízottjainak szánnak. Esetenként a könnyebb érthetőség kedvéért kiegészítik magas szintű, vázlatos használati eset diagramokkal.

A természetes nyelven megfogalmazott követelmények, az olyan egyértelmű előnyeik mellett, mint az egyszerűségük, olyan problémákat is felvetnek, mint az egyértelműség nehéz biztosítása, ezekkel: A követelmény feltárás, és kezelés problematikái című részben bővebben foglalkozom.

### Rendszerkövetelmények

A rendszer szolgáltatásait, funkcióit, működési feltételeit és megszorításait részletezi. A felhasználói követelményeket fejtik ki bővebben. Érdemes megjegyezni, hogy a rendszerkövetelmények dokumentumát szokták specifikációnak is hívni.

Lényeges a specifikáció pontossága, hisz meghatározza, hogy mit kell implementálni, de ezen a szinten már kivédhetetlenül megjelenhet a hogyan is, emellett azért is fontos a teljes körű és precíz megfogalmazás mert a rendszerspecifikáció gyakran része a szerződésnek.

A rendszerspecifikációból lehetetlen kizárni minden a tervezéshez köthető információt, mert gyakran a követelményspecifikáció strukturálásához segítséget nyújt egy kiindulási architektúra elkészítése. A gyakorlatban emellett gyakran szükséges, hogy együtt működjünk más rendszerekkel, és ez bekorlátolja a tervezést és ezek a korlátok szükségszerűen új követelményekként is megjelenhetnek. Esetenként egy-egy nemfunkcionális követelmény is előírhat egy konkrét tervezési megközelítést, vagy egy konkrét fejlesztési platformot, így ilyenkor is megjelennek tervezési részletek, a rendszerkövetelmények szintjén.

A rendszerkövetelmények tehát megalapozzák a tervezést, illetve az implementációt, így a szoftver megalkotásában résztvevők a legfőbb olvasói. Ezért míg a felhasználói követelményeknél elengedhetetlen a természetes nyelv, és a külön képzettség nélküli érthetőség, addig a rendszerkövetelmények szintjén, az egyszerű formális szövegek mellett speciális jelöléseket és modelleket is alkalmazhatunk. Ilyenek lehetnek a stilizált, formázott és strukturált természetes nyelv, vagy a követelmények grafikus modelljei, mint a részletezett használati esetek vagy akár a matematikai formális nyelvek.

### Rendszer

Rendszer alatt olyan egymáshoz kapcsolódó komponensek, működési egységelemek, halmazát értjük, melyek egy közös cél érdekében dolgoznak együtt. Ezek a rendszerkomponensek munkájuk elérése érdekében, kommunikálnak a többi komponenssel és felhasználhatják azokat. A rendszer tartalmazhat szoftveres, mechanikus, és elektronikus egységeket is. A rendszer célja elérése közben emberi beavatkozást is igényelhet, de lehet teljesen automatikus is.

### Funkcionális követelmények

Leírja a rendszertől elvárt szolgáltatásokat. Az iménti fejezetekből világosan látható, hogy két szintje lehetséges, a funkcionális felhasználói követelmények magas szintű állítások, amik megfogalmazzák a rendszertől elvárt funkcionális viselkedést, míg a funkcionális rendszerkövetelmények, e funkcionalitás részletezései.

### Nemfunkcionális követelmények

Az egész rendszerre vonatkozó tulajdonságok. Olyan eredő rendszertulajdonságokra vonatkozhat, mint a megbízhatóság, a válaszidő vagy a tárfoglalás. Lehet megszorítás egyes használt technológiákra, eljárásokra, vagy szabványokra vonatkozólag. Nem csak a teljes rendszer, de akár annak egy-egy összetevőjére is vonatkozhatnak. A míg, egy funkcionális követelmény nem megfelelő támogatásával a rendszer egyes részei megkerülendővé válhatnak, addig a nemfunkcionális követelmények nem teljesítése akár a teljes rendszert is használhatatlanná teheti. A nemfunkcionális nem csupán magára a rendszerre vonatkozhat, de magára a kifejlesztésének folyamatára is vonatkozhatnak. Megszabhatja tehát a kifejlesztés közben használt metodikát, különböző minőségszabványok betartását írhatja elő, vagy akár megszabhatja a fejlesztés alatt használatos CASE eszközök sorát.

Az ide sorolható változatos követelmények, különböző forrásokból is származhatnak, mint például a szoftver termék kívánt jellemzőiből, a fejlesztő vállalattól, vagy valamilyen külső féltől, mely többnyire a megrendelő, de lehet akár valamilyen előírás is.

Lehetséges, hogy egy adott nemfunkcionális követelmény megléte, szükségessé teszi más, funkcionális követelmények felvételét. Például egy információbiztonsággal kapcsolatos követelmény, megkövetelheti különböző authentikációs és authorizációs funkciók felvételét.

A nemfunkcionális követelmények egyik problematikája, hogy a megrendelők gyakran csak általános célokként fogalmazzák meg. Ilyen nehezen mérhető jelzőket használnak, mint gyors, kisméretű, egyszerűen kezelhető, hordozható, megbízható, stb. Ezek helyett célszerű a követelményekben különböző objektív metrikákat használni. A gyors, jelző kifejezhető inkább a másodpercenkénti tranzakciók számával, egy eseményre adott átlagos válaszidővel, vagy akár a képernyő frissítési sebességével. A méret leírható elfoglalt tárterület alapján, mint például nem lehet nagyobb 64 kilobájtnál, vagy a maximális dinamikus memóriafogyasztással. A könnyű használatot leírhatjuk a betartott tervezési ajánlásokkal, a képzési idővel, vagy a segédképernyők számával. A megbízhatóságot a rendelkezésre állás, a hibák közt eltelt átlagos idő, és a hiba bekövetkeztének valószínűsége írhatja le precízen. A hordozhatóságot jellemezheti például a célrendszertől független utasítások aránya.

### Szakterületi követelmény

A szakterületi, vagy más néven környezeti követelmények, a rendszer alkalmazási területéről származnak. Ezek a követelmények nem éppen felhasználói igények, inkább a szoftver alkalmazásának szakterületéből adódó funkcionalitások, vagy megszorítások.

Ezek teljes értékű funkcionális, vagy nemfunkcionális követelmények, de mind a nyelvezetük, mind fogalomrendszerük az adott szakterülethez igazodik. Ha funkcionális követelmény, akkor előírhatja, hogy az adott feladatot hogyan kell végrehajtani, hogy az helyes legyen, illetve illeszkedjen az adott szakterület bevett gyakorlatához. Lehet akár egy szabvány, de lehet akár egy képlet is, amivel az adott szakterületen számolnak.

Ezen követelmények nagy jelentőségűek, míg az adott szakterület képviselőinek egyértelműek addig a rendszer tervezőinek némiképp idegenek lehetnek a nyelvezetük idegensége és a mögöttes implicit háttértudás hiányában.

### Szükséges és javasolt követelmények

A javaslat, nem mindig követelmény, lehet, hogy csupán egy igény a rendszerre vonatkozólag. Esetleg egy teoretikus elképzelés, ami jelenleg nem fontos, vagy kritikus és nem döntő fontosságú a rendszer szempontjából. Érdemes lehet ezeket is rögzíteni, hisz nyerhetünk belőlük újabb jó ötleteket vagy az is lehet, hogy egy átpriorizálás közben felértékelődik, és szükségessé nyilváníthatják.

A javasolt követelmények nyelvezetében érdemes, ilyen kifejezéseket használni, mint: „jó lenne ha”, vagy „érdemes lenne”, „javasolt lenne”, „javallott” stb.

Míg a szükséges követelmények a „kell”, a „szükséges”, a „fontos”, a „kívánt” az „elvárt”, és hasonló szavak használata különíti el az előző csoporttól.

### A követelménytervezés folyamata

Ebben a folyamatban megértjük és definiáljuk a rendszer által biztosítandó szolgáltatásokat, illetve a fejlesztési és az üzemeltetési megszorításokat. A folyamat végeredménye a követelménydokumentum, mely rend szerint külön tárolja a felhasználói illetve a rendszerkövetelményeket.

### Szoftverkövetelmények dokumentuma

A szoftverkövetelmények dokumentuma, amit szoktak szoftverkövetelmény specifikációnak is hívni, az a dokumentum, mely a követelmény specifikáció folyamata során jön létre. A specifikáció tehát a követelmények egy szabványos formába való szedésének és leírásának folyamata. Maga a dokumentum tartalmazza a felhasználói-, illetve a rendszer követelményeket. Szöveges dokumentum, így a megrendelő külön előképzettség nélkül is olvashatja, de a kötött szerkezete elősegíti a lehető legteljesebb információreprezentálást.

A funkcionális és a nemfunkcionális követelményeket célszerű megkülönböztetni a dokumentumon belül. Ez lehet elszeparálás, ilyenkor a külön elhelyezkedő követelmények között nehezebb megtalálni az esetleges összefüggéseket, viszont az egyes követelmények nem mosódnak egybe, és jól elkülöníthetőek a funkcionális, illetve a nemfunkcionális megfontolások. A megkülönböztetés történhet esetleg más vizuális jellemzőkkel is.

### Verifikáció és Validáció

Ezt a két lépést együtt szokták V&V-nek is rövidíteni. Míg a verifikáció azt ellenőrzi, hogy a szoftvert jól, azaz a specifikációknak megfelelően készítettük-e el, addig a validáció azt ellenőrzi, hogy tényleg a megfelelő, a felhasználó által óhajtott, értékes terméket alkottuk meg.

### Fogalomszótár

A követelmények között, vagy a különböző modelleken megjelenhetnek olyan szavak, melyeket érdemes lehet definiálni. Ez több célt is szolgálhat, növeli a szövegek precizitását, megóvhat az egyes félreértésektől, kiküszöböli a kétértelműségeket. A szakterületi kifejezések leírása segítheti a fogalmak megértését, és ezáltal segítheti a hatékonyabb fejlesztést. A rendszerben használt rövidítéseket is érdemes lehet ide felvenni.

## A követelmények fontossága

A követelmények meghatározása, elemzése és karbantartása a megvalósítandó komplex szoftverrendszerek életciklusában egy kritikus fontosságú feladat, hisz ez a szakasz segít megérteni, és hogy a megrendelő mit is vár el a készítendő rendszertől. A feltárt követelmények alapján döntünk arról, hogy a rendszer megvalósítható-e és hogy megközelítőleg mennyibe fog kerülni a megrendelőnek. A szoftver komplexitása is becsülhető általa. Általában a szoftver specifikáció része a szerződésnek.

Hagyományosan a specifikáció köti össze a felhasználó igényeit és a fejlesztőket, így a kommunikációs szerepe is jelentős. Amennyiben a rendszer megrendelői mi magunk vagyunk, vagy a kommunikáció igen jó és gyakori a megrendelővel, illetve egyes metodikákat alkalmazva, a követelmények kifejtése a fejlesztés során több részletben, iteratívan történik.

## Követelménytervezés folyamata

Az a folyamat, melyben felderítjük, elemezzük, dokumentáljuk és ellenőrizzük a szoftverkövetelményeket. Ez a szoftverspecifikáció megalkotásának folyamata, itt készítjük el és tartjuk karban a rendszerkövetelmények dokumentumát.

A folyamat négy nagy tevékenységre bontható. Ezek a részfolyamatok a megvalósíthatósági tanulmány elkészítése, a követelmények feltárása és elemzése, a követelmények validálása, illetve azok kezelése és követése.

### Megvalósíthatósági tanulmány elkészítése

Megvizsgálja és becslést ad arról, hogy a rendszerrel kapcsolatos elvárások kielégíthetőek-e az adott szoftveres és hardveres technológiák segítségével. Eldönti, hogy a rendszer költséghatékony-e az adott üzleti szempontokat figyelembe véve, illetve hogy a költségvetési megszorítások mellett kivitelezhető-e. Lehetőség szerint minél inkább gyors, és olcsó folyamatnak kell lennie. A tanulmány elkészítése során kiszámításra kerül egy ROI (Return Of Investment) érték, mely megadja, hogy milyen mértékben és mikor térül meg az elkészítendő rendszer, és mekkora annak az üzleti haszna. A megvalósíthatósági tanulmány információt biztosít a rendszer elkészítésének költséghatékonyságáról.

A tanulmány elkészítésének végeztével döntés születik a fejlesztés folytatásáról. A megvalósíthatósági jelentés a részfolyamat kimenő dokumentuma, mely tartalmazza a lépésben összegyűjtött fontosabb információkat és döntéseket. Ez után a lépés után következhet a követelmények feltárása részfolyamat.

### Követelmény feltárás és elemzés

A folyamat során a potenciális felhasználókkal és megrendelőkkel történő megbeszélések, és egyeztetések során, illetve az esetleges már működő rendszerek és folyamatok megfigyelése által az elemzők feltérképezik és megértik a készítendő szoftver követelményeit. Ez a folyamat magában foglalhatja egyes rendszermodellek, illetve prototípusok elkészítését, melyek elősegítheti a követelmények pontosabb megértését.

### Követelmény specifikációs folyamat

Az elemzési tevékenység során összegyűjtött információk egységes dokumentummá történő szerkesztésének folyamata.

### Követelmény validáció

A követelmények valószerűségét, konzisztenciáját és teljességét ellenőrző tevékenység.

### Követelmények kezelése

## A követelmény feltárás, és kezelés problematikái

### A természetes nyelv problémái

A természetesen nyelven írt dokumentumok, mint amilyen lehet a felhasználói követelmények is, számos problémával rendelkeznek. Az egyik ilyen probléma az egyértelműség hiánya. Olykor nehéz a nyelvet pontosan használni, egy dolgot többféleképpen is leírhatunk, és a szavainknak is lehet számos jelentése. A szövegeink emellett terjengőssé is válhatnak, ezzel megnehezítve a lényeges információk kiszűrését. A folyó szövegben összemosódhatnak az egyes követelmények, illetve a különböző típusú, funkcionális és nemfunkcionális követelmények keveredhetnek így nehezebb ezeket elválasztani egymástól.

### A természetes nyelv problémáit kiküszöbölendő megfontolások

A követelmények megfogalmazásakor érdemes lehet minden követelményhez egy egyszerű magyarázatot fűzni, ami kifejti, hogy miért került be a megjegyzés, mi a szerepe annak. Ez nem csak érthetőbbé teszi az adott követelményt, de változásakor is segítséget nyújthat.

Érdemes egy szabványos követelményformátum elkésztése. Ami megadja, milyen formai és tartalmi szerkezettel adjuk meg az egyes követelményeket. Egyes ajánlásokban például a követelményeket kártyákra vették fel, minden egyes ilyen kártya egy darab követelményt tartalmaz, illetve tartalmazza az adott követelmény magyarázatát, más követelményekkel való kapcsolatát, azoktól való függését, illetve a követelmény forrását, azaz azt a személyt, akitől a követelmény ered. Így könnyen megtalálható az, akivel a követelmény változásakor érdemes lehet konzultálni. A formátum előírhatja továbbá a követelmény félkövér típussal való kiemelését, esetleg azt, hogy a felhasználói szintű követelmény tartalmazzon egy linket az azt részletező rendszerkövetelményre. Az ilyen kötöttebb, szabványos formátum megelőzheti az esetleges véletlen elhagyásokat, és könnyebben ellenőrizhetővé teszi magukat a követelményeket is.

Mint már egy korábbi szakaszban említettük érdemes lehet szétválasztani a szükséges illetve a javasolt követelményeket, például a „kell” vagy a „javallott” szavak felhasználásával és ezek következetes használatával.

A követelmény kulcsfontosságú részeit érdemes lehet kiemelni, félkövér, dőlt, vagy egyéb szövegformázási eszközökkel, ezzel elősegítve a követelmény későbbi olvasását, és az fontosabb információk hangsúlyozását.

Elkerülendőek a számítástechnikai szakzsargonok, de a rendszer felhasználási területének felhasználása elkerülhetetlen. Érdemes lehet az egyértelműség kedvéért fogalomszótárat használni.

Teljesség, ellentmondás mentesség, konzisztencia.

Analízis paralízis.

Változás.

Nyelvezet.

Elkülönítés, csoportosítás.

## A szoftveréletciklus modellek és a követelmények kapcsolata

A szoftverfejlesztési módszertanok olyan módszerek, melyek meghatározzák a fejlesztés menetét és lépéseit. Keretet biztosítanak ahhoz, hogy milyen sorrendben végezzük az egyes a rendszer kifejezését elősegítő lépéseket. A módszertanok tartalmazhatnak különböző eszközöket a modellezésre, jelölési konvenciókat és ajánlásokat. Tanácsokat, esetlegesen betartandó szabványokat, tartalmazhatnak specifikációra, a tervezésre, illetve a fejlesztésre vonatkozólag.

### A hagyományos megközelítések problémái

Az olyan hagyományos módszertanokban, mint a vízesés modell, a fejlesztési lépések szigorú egymásutániságban követik egymást. Ezek a lépések a következők: követelményelemzés és meghatározás, a rendszer- és szoftvertervezés implementáció és egységteszt, az integráció és rendszerteszt, és végül egy működtetés és karbantartás, lépés. A módszer egyik nagy hibája abban áll, hogy ebben a merev rendszerben, gondot okoz, a követelmények módosulása, ugyanis ilyenkor esetleg az összes eddigi lépéseket újra végre kell hajtani. A követelmények minősége, és teljessége, kritikus, hisz a többi fázis ezekre építkezik, és a végső termék minősége is főként ezektől függ. A gyakorlatban számtalanszor előfordult, hogy a követelmények elemzése annyi időt és energiát emésztett fel, hogy az a teljes project bénulásához vezetett. Ezt a jelenséget hívja a szakirodalom elemzési bénultságnak (Analysis paralysis).

Egy másik probléma, hogy későn vesszük észre, ha esetleg nem pont azt a terméket készítjük, amit a felhasználó kíván, tehát későn következik be egy átfogó validáció, és későn látja működés közben a rendszert. Így a project kockázata döntő mértékben a fejlesztés végére csoportosul.

Az ilyen módszertanok, csak lassan változó követelményekkel rendelkező, teljesen leírható területekre szánt szoftvereknél alkalmazható hatékonyan.

### A manapság gyakori módszertanok és a követelmények kapcsolata.

A manapság legelterjedtebb életciklus modellek mind iteratív jellegű, ciklikus folyamatok. Ezekre általában igaz, hogy egyszerre a rendszer egy részét veszik, és ezen végigviszik, a fejlesztés lehetőleg összes lépését, majd az elkészült produktumot a felhasználónak meg is mutatják véleményezésre.

Az egyik ilyen metodika a RUP (Rational Unified Process), melyet az UML kifejlesztésén is fáradozó „három amigó”, James Rumbaugh, Grady Booch és Ivar Jacobson készítettek, a Rational cégnél. Ez a metodika egy iteratív és inkrementális modell. A szoftvert növekményekre bontja, ezeket az inkremenseket, a felhasználó szemszögéből fontossági sorrendbe állítja, és ezeken a részeken viszi végig a fejlesztési fázisokat. Így biztosított az, hogy a fontosabb rendszerkomponensek előbb elkészüljenek, ezeket a felhasználó, akár a teljes rendszer elkészülte előtt, igénybe is veheti. Ez a metodika ajánlja a teljes rendszer, különböző részletességű vizuális modelljeinek elkészítését. A módszer az üzleti folyamatok, illetve a követelmények modellezésére a használati eseteket ajánlja. És ezeket a modelleket fel is használja a teljes fejlesztési folyamat során. Fontosnak tartja a megrendelő követelményeinek explicit dokumentálását, és azok folyamatos karbantartását.

A napjainkban divatos Agilis elveket követő modellek is igyekeznek a rendszert kisebb a felhasználó számára értékeket hordozó egységekre bontani, és ezeket súlyozni, és fontosságuk szerinti sorrendben elkészíteni. Ezek az értékek egyenként áramolnak át a fejlesztés összes fázisán. A kezdeti célok és magas szintű üzleti modellezés és követelmény leírások itt is szerepet kapnak, de a végső követelmények a termékkel együtt fejlődnek. Az agilis szemléletmódot követő Scrum metodika, vagy az egyre inkább tért nyerő Lean megközelítés is a fejlesztés természetes velejárójaként tekint a követelmények változására, ugyanis sokszor a fejlesztés közben jobb képet kaphatunk a rendszerről, nagyobb tudás birtokában vagyunk. Emellett a világunk is folytonos változás állapotában van, így nem kérdés, hogy egy szoftver követelményei is megváltozhatnak. Az agilis projectekben inkább a felhasználó érdekeit igyekeznek szem előtt tartani, minthogy a követelmények szövegeihez próbálnának csökönyösen ragaszkodni, de ez nem jelenti azt, hogy az ilyen szemlélet szerint készülő szoftvereknek ne lenne specifikációja, sőt a követelmények modellezése itt is bevett szokás, de a formális diagramok helyett előszeretettel választják a formába öntött szövegek használatát.

Azzal, hogy a ciklikus modellek, a szoftvert több szakaszban adják át a megrendelőnek, a project kockázatát, a hagyományos modelleknél jobban képesek időben elosztani, és a project összkockázatát is sikeresen csökkentik. Mindemellett a gyakoribb kommunikációval és átadásokkal a követelmények esetleges félreértéséből adódó kockázatot is eredményesen csökkentik. Emellett viszont megnövelik a követemének karbantartásának és követésének a fontosságát. A nagyobb és komplexebb szoftverfejlesztési projecteknél elengedhetetlen, hogy a követelmények kezelését megfelelően támogató szoftvereszközöket alkalmazzanak.

## Követelménymodellezés

### Az UML rövid bemutatása

## Követelmény modellezés eszközei

### Use Case modell

## Az agilis követelménykezelés

### User Stories: A használati esetek agilis megközelítése

Scenario

## Szakirány vezéreltség

### A TDD, DDD, BDD hármas értelmezése

### Szakirány vezérelt tervezés (Domain Driven Design)

### Girkin nyelven leírt követelmények

## Követelmény formátum sablonok és ajánlások

### Felhasználói követelmény sablon

### Rendszerkövetelmény sablon

## A CASE eszközök

A CASE (Computer-Aided Software Engineering) rövidítés számítógéppel támogatott szoftvertervezést jelent. Ellenben ezen eszközök jelenleg koránt sem csak a tervezés folyamatát támogatják.

A CASE rövidítés (Computer-Aided Software Engineering) számítógéppel támogatott szoftver tervezést jelent, de ezen eszközök jelenleg koránt sem csak a tervezés folyamatát támogatják.

A CASE eszközök olyan szoftverek vagy szoftver rendszerek, melyek a szoftverfejlesztés különböző fázisait és/vagy ezek közti kapcsolat megteremtését segítik elő, lehetőleg minél több automatikus eszköz és eljárás segítségével. A CASE eszközök többnyire egy fejlesztési metodikához illeszkednek, és ennek a szoftverfolyamatnak a tevékenységeit, vagy azok egy részét támogatják. Persze léteznek általánosan használható CASE eszközök is, melyesk nem kötődnek szorosan egy adott metodikához. A szoftverfolyamat tevékenységei közül támogatják például a követelményelemzést, a követelmények változásainak kezelését, a rendszermodellezést és a modellek esetleges ellenőrzését, a tervezést, a fejlesztés nyomkövetését, és akár a tesztelést. CASE eszközök segíthetik a követelmény specifikációs, és a szoftvertervezésnél használt modellek és diagramok előállítását. Egyes CASE eszközök a tervezés elemeiről adatszótárakat tartalmaznak, amik leírják az adott elemeket és azok kapcsolatait. Segíthetik a felhasználói interfész vázlatszerű, vagy teljes értékű elkészítését.

A CASE eszközök, a rutinfeladatok automatizálásával képesek a szoftver minőségén javítani, illetve a fejlesztés sebességét fokozni.

Bár nagymértékben megkönnyítik a szoftverek kifejlesztését, a CASE eszközök nem voltak képesek oly mértékben segíteni a munkát, mint amennyire azt a 1980-as, '90-es években prognosztizálták. Ennek fő okai közül az egyik, hogy a szoftverek kifejlesztése egy egyedi, és kreatív folyamat, melyet, akár mesterséges intelligencia bevonásával is, csak nehezen, vagy egyáltalán nem lehet teljesen automatizálni. A másik jelentős probléma, hogy egyes eszközök nem képesek elégségesen megtámogatni a csoportos munkavégzést, mely a nagy és komplex feladatoknál elengedhetetlen. Szerintem törekedni kell továbbá ezen CASE eszközök minél nagyobb fokú integráltságának, illetve együttműködésének biztosítására. Hisz hiába van egy olyan eszközünk, ami az adott részfolyamatot tökéletesen segíti, de az adott rendszer feltöltése a szükséges adatokkal több időt elvesz, mintha egy olyan rendszert használnánk, amely esetleg nem olyan hatékony céleszköz, de a környező rendszerekkel való kommunikációja révén a feladatot mégis gyorsabban képes ellátni, és a fejlesztési folyamatban az értékek áramlását, kevésbé gátolja.

## A szövegbányászatról általánosságában

Az emberek, már a kezdeti, ősi civilizációkban is, a szóbeli mellett, jellemzően írásbeli szövegek segítségével tárolták, és adták át egymásnak az ismereteiket. Napjainkban a rögzített tudásanyagainknak jelentős hányada egyszerű szöveges dokumentumokban található. Ezt a feltevést támasztják alá, többek között a Merill Lynch elemzései is, melyek becslése szerint az üzleti információk körülbelül 85%-a található strukturálatlan, illetve gyengén strukturált szövegekben. Az általunk kezelt szövegek növekvő arányban digitálisan tárolt dokumentumok.

Így talán nem meglepő, hogy a szövegek, és főképp a bennük lévő információk kezelésének hatékonyabbá tétele napjaink egyre fontosabbá váló informatikai tevékenysége. Az egyik tudományág mely ezzel foglalkozik, a szövegbányászat.

A szövegbányászatot definiálhatjuk úgy, mint szöveges adatokon végzett feldolgozási és elemzési tevékenység, mely célja a dokumentumban rejtett információk feltárása, azonosítása, és elemzése. A szövegbányászat interdiszciplináris szakterület, mely olyan informatikai eszközök mellett, mint a gépi tanulás és a hatékony algoritmusok, a matematika és a nyelvészet eszközeit is felhasználja.

A szövegbányászat két nagy alaptípusa a keresés és a rendszerezés. A keresésnél kiválasztjuk azokat a dokumentumokat, ahol egy adott keresőkifejezés előfordul, míg a rendszerezésnél valamilyen kategóriákba, vagy előre nem definiált csoportokba soroljuk azokat. A szövegbányászat főbb feladattípusai is ilyen jellegű feladatokat, vagy ezek kombinációját hajtják végre céljaik elérése során.

Ezek a főbb feladattípusok a kereséstámogatás és információ-visszakeresés, az információkinyerés, az osztályozás, a csoportosítás, az összegzéskészítés, a kivonatolás, a válaszkereső rendszerek, a szövegelemzés, és a napjainkban egyre inkább tért nyerő webes tartalomkeresés. Ezek közül csak a dolgozat szempontjából leginkább hangsúlyos információkinyerést fogom mélyebben részletezni.

## Szövegbányászati előfeldolgozás

Számos szövegbányászati feladat megoldható, már létező, adatbányászati eszközökkel, és algoritmusokkal. Ehhez a szöveges adatokat úgy kell transzformálni, olyan alakra kell hozni, hogy ezek a bejáratott eljárások, lehetőleg hatékonyan, képesek legyenek működni rajta. Ezt a transzformációs lépést szokták előfeldolgozásnak is nevezni. A folyamat végeredménye a dokumentumot reprezentáló modell. Az előfeldolgozás egységesítési, formalizációs és normalizációs feladatokat is tartalmaz.

## Az előfeldolgozás közben fontos alapvető dokumentumjellemzők

### A dokumentumot hordozó médium

Ha digitális, gondoskodni kell az elérhetőségről, ha pedig analóg, akkor előbb digitalizálni kell. Ha képi, akkor a képen látható szövegeket ki kell nyerni, lehetőleg automatikus karakterfelismeréses vagy akár manuális begépeléses módszerrel.

### A dokumentum elérési helye

A hagyományos világban lehet egy könyvtár egy polca, míg a digitális világban jellemzően lehet egy URL (Uniform Resource Locator) mely egy egységes erőforrás-azonosító és az interneten keresztüli adatelérésre szolgál, vagy lehet egy háttértárolón egy fájl elérési útvonala. Szövegbányászatnál mindenképpen fontos, hogy rendelkezzünk legalább olvasási jogokkal az adott dokumentumon.

### A dokumentum mérete

A nyers digitális állományok legelterjedtebb általános méretmegadási módja az adattárolón elfoglalt terület alapján, bájt alapon történik. A szövegfájlok mérete megadható még benne szereplő karakterek, szavak vagy az oldalaik számával. A szöveges korpuszok méretét a benne szereplő dokumentumok összesített mérete adja.

Az előfeldolgozási fázis után a méretet jellemzően szavaik, ritkábban karaktereik számával adhatjuk meg. A modellbeli reprezentációjuk méretét pedig olyan a felírásukra használt elemek számával jellemzik, mint a tokenek vagy az egyedi szavak száma, esetleg a modell indextömbjének mérete.

### A dokumentum statisztikai jellemzői

A későbbi feldolgozás során szükségünk lehet egyes statisztikai adatokra is a dokumentumunkról. A legfontosabb ilyen jellemző lehet a szavak eloszlása, de esetenként fontos lehet még a karaktereinek eloszlása, a szóhosszoknak az eloszlása vagy átlaga. A magán és mássalhangzók száma és eloszlása, vagy a mélyebb nyelvi elemzéssel kideríthető zöngés és zöngétlen magán- illetve mássalhangzók száma, melyekből például következtetni lehet az adott dokumentum nyelvére.

### A dokumentummal kapcsolatos metaadatok

A dokumentum adathordozón való tárolásakor keletkeznek metaadatok, mint a keletkezési ideje, az utolsó módosítás ideje, a fájlnév, illetve a hozzáférést módosító attribútumok.

Szövegbányászati szempontból egyéb jellemzők is fontosak, ilyen például a dokumentum tartalmára vagy a keletkezési körülményeire vonatkozó metaadatok, mint a nyelv, a keletkezési idő/hely, a dokumentum témája vagy más szempont szerinti besorolása, a stílusa, a nyelvezete vagy zsánere. Ezek osztályozáskor vagy csoportosításkor lehetnek segítségünkre. Ezek közül az adatok közül nem áll rendelkezésünkre minden, de néhányat, ha hiányzik, szövegbányászati eszközökkel meg tudunk határozni. A nyelv vagy a téma is lehet ilyen.

### A dokumentum formátuma

A dokumentum formátumát többnyire a készítésére használt eszköz típusa bekorlátozza. Egyes eszközök egyszerű szövegként (plain text, txt) mentenek, míg másik eszközök formátumai különböző formázási és szerkezetre vonatkozó adatok, de akár a dokumentum régebbi változatai is lehetnek. Az ilyen tárolási formátumok, a teljesség igénye nélkül, például a doc/docx, a pdf, rtf, de akár a HTML és az XML is ide sorolható.

A dokumentum készítői által kiemelt szövegrészek lehetnek külön formázásúak, mint például a félkövéren szedett, a dőlt, vagy az aláhúzott szövegrészek, és lehetnek szerkezetileg meghatározó helyen, mint a különböző szintű címekben, vagy például a webes elérésű dokumentumok az elérési útvonalában is szerepelhetnek. A dokumentumban kiemelt kifejezések a dokumentum szempontjából relevánsabbak lehetnek, ezért érdemes lehet ezeket az információkat a modell reprezentációba is átvinni.

### A dokumentum karakterkódolása

Az általunk használt adatok különböző karakterkódolásokban fordulhatnak elő, hiszen a számítógép, a szöveges adatokat is kettes számrendszerű számokként tárolja, ezért szükséges annak a meghatározása, hogy egy-egy szám milyen karaktert kódol.

Az egyik első széles körben használt kódtábla az ASCII (American Standard Code for Information Interchange), amit az Amerikai Egyesült Államokban alkottak meg, először hét biten ábrázolt, ami 128 karakter ábrázolását teszi lehetővé, melyek némelyike vezérlő karakter, mint az új sor, vagy a tabulátor, illetve vannak nem látható vezérlő karakterek is melyek a korabeli nyomtatókat vezérelték. Az ASCII 8. bitjét eleinte paritásbitnek szánták, de később általánossá vált, hogy a karakterkészlet kibővítésére használták, az így nyert további 128 lehetséges állapottal az ISO 8859-1 Latin-1-es kódolás a nyugat európai, míg az ISO 8859-2 inkább a kelet európai nyelvek speciális ékezetes karaktereit írja le. Napjainkban egyre inkább az Unicode veszi át a vezető karakterkódolási szabvány szerepét, mely arra hivatott, hogy legyen képes az ismert nyelvek összes karakterét kódolni. Legelterjedtebb megvalósítása az UTF-8, mely változó hosszon kódolja a karaktereket, ezért tömör kódolást biztosít, mindeközben kompatibilis a 7 bites ASCII-vel. A karakterkódolások különbözősége mellett olyan dolgok is nehezítik a feladatunkat, mint hogy a sor végének jelölése egyes platformokon eltérhet.

Az ismert kódolások részben, vagy teljesen inkompatibilisek, ezért a konvertálásnál kiemelt kérdés a karakterkódolás, mivel ennek a helytelen használata, a beolvasott szöveg torzulásához vezethet, vagy az akár teljesen használhatatlanná is válhat.

A karakterkódolási információt esetenként maga a fájl tartalmazza, ha nem ismerjük, akkor az előfeldolgozás helyessége nem garantálható.

## Szövegbányászati modellalkotás

### Az ismertebb dokumentum reprezentációs modellek bemutatása

A legelterjedtebb modellek jellemzően valamilyen numerikus objektumok. Ez számos előnnyel jár. Az egyik előny a kisebb tárolási méret. Ugyanis ha a szavakat karakterenként letároljuk, és jellemzően egy-két bájtos egy karakter, akkor nagyobb helyet vennénk igénybe, mint a numerikus tárolásnál, ahol szavanként egy darab 2 vagy 4 bájtos számmal számolunk. Mindemellett, a számok használatának van egy másik jelentős előnye, méghozzá az, hogy matematikai műveleteket, és transzformációkat hajthatunk végre az ilyen modelleken. Ráadásul a modellben a matematikai eszközökkel való munka elősegíti a dokumentumok hatékony kezelését.

Azt, hogy ténylegesen milyen modellt és adatábrázolást használunk, befolyásolja a megoldandó feladat típusa. Keresés jellegű feladatoknál egy megfelelő szóelőfordulás táblázat is nagy szolgálatot tehet, míg a rendezés jellegű feladatoknál összetettebb dokumentum összehasonlító módszerekre van szükség.

A modellalkotásnál használt három nagy matematikai elméleti megközelítés a halmazelméleti, az algebrai, illetve a valószínűségelmélet alapú. A halmazelmélet alapú modellek jó szolgálatot tehetnek az egyes keresőrendszerekben, hisz kereséskor minden dokumentumra fennáll, hogy része az eredményhalmaznak, vagy sem. Az algebrai modellben a dokumentumokat olyan algebrai objektumokként reprezentáljuk, mint a vektor, vagy a mátrix. Ezeket algebrai műveletekkel össze is hasonlíthatjuk, ezért ezek már használhatók, rendszerezési feladatok megoldásánál is. A legelterjedtebb megvalósítása a vektortér modell és annak változatai. A valószínűségi modellben maguk a dokumentumok valószínűségi események által reprezentáltak, míg a kapcsolataik feltételes valószínűségi becslések eredménye. Ezek mellett léteznek komplex függvénytani eszközöket alkalmazó spektrális szövegbányászati modellek is.

### A vektortér modell

A vektortér modellben hatékonyan meg lehet határozni a dokumentumok távolságát, illetve hasonlóságát. A szövegbányászatnál gyakran élünk különböző intuitív heurisztikákkal, melyek meghatározzák, az ezeket felhasználó eszközök felhasználhatósági körét és korlátait. A vektortér modellnél azt jelentjük ki intuitív módon, hogy azokat a dokumentumokat tekintjük hasonlónak, melyek szókészlete átfedi egymást, és ennek a hasonlóságnak a mértéke arányos az átfedés mértékével. A modell egy sokdimenziós vektortérben, vektorokkal reprezentálja a dokumentumokat. A vektortér egyes dimenzióit a dokumentumgyűjtemény egyedi szavai adják. Tehát egy-egy dokumentum a szavaiból álló vektor, abban a vektortérben, ahol az egyes szavak a teret kifeszítő vektorok. A dokumentumgyűjteményt a szó-dokumentum mátrixszal reprezentáljuk. Az egyedi szavak összessége a szótár, vagy más néven lexikon.

Általában a teljes szótár csak egy kis része fordul elő egy-egy dokumentumban, ezért a szó-dokumentum mátrix ritka. Mindemellett az egyedi szavak száma meglehetősen nagy, akár több százezres, vagy akár milliós nagyságrendet is elérheti. Ezért a mátrix méretét nyelvtechnológiai, illetve matematikai eszközökkel érdemes lehet csökkenteni. A vektortérmodell másik nehézsége az alaphipotézisből származik, miszerint a szavak megléte, illetve száma adja a dokumentumok hasonlóságát. De nem tér ki, és nem is tárolja a szavak sorrendjére se a szövegen belüli pozícióira vonatkozó információkat. Ezért hívják szózsákmodellnek is. Ez a modell a feladatok nagy részében jól használható, egyszerű és hatékony megoldást nyújt. A szósorrend elvesztéséből való hátrányát sokszor más modellekkel való együttes alkalmazásával küszöbölik ki.

### A szó-dokumentum mátrix jellemző súlyozási sémái

A legegyszerűbb módszer, ami csak a szó dokumentumbeli esetleges meglétét jelöli, a bináris reprezentáció. Ahol a mátrix egy adott dokumentumot reprezentáló oszlopvektorában, egy adott szóhoz tartozó sorban nullát írunk, ha az adott szó nem szerepel a dokumentumban és egyet, ha igen, és nem számít az, hogy hányszor, ez az információ elvész. Egy másik egyszerű módszer, ha a súly maga az előfordulások száma. Itt viszont felvetődhet a kérdés, miszerint tényleg lineáris-e a függés a szó adott dokumentumra vonatkozó relevanciája, és az adott szó előfordulásainak száma között. Tehát egy jóval gyakrabban előforduló szó tényleg ugyanilyen mértékben fontosabb-e. Az esetek nagy többségében a válasz egyértelműen nem, sokkal nagyobb megkülönböztető értéke van, ha az adott szó előfordul a dokumentumban, mint annak, hogy az tízszer, vagy százszor fordul elő. Ezt kiküszöbölendő kezdték el használni a logaritmikus súlyozást. Ezen egyszerű súlyozási sémák nagy hátránya, hogy nem veszik figyelembe az egyes dokumentumok hosszát. Pedig feltételezhetően egy rövidebb dokumentumban egy szó bizonyos számú előfordulása fontosabb, mint az ugyanekkora számú előfordulása egy jóval hosszabb szövegben. Így a bonyolultabb súlyozási függvények figyelembe veszik a szó dokumentumbeli előfordulási gyakoriságát is. Ezt a gyakoriságot az angol irodalomban term frequency kifejezéssel illetik, és a szakirodalmakban gyakran TF-nek rövidítik. Ezen elven alapuló jellemző megvalósítások a gyakoriság alapú, illetve a normalizált logaritmikus súlyozások. Ez a megközelítés már közelebb áll a valósághoz, de nem veszi figyelembe azt, hogy vannak olyan szavak, melyek minden dokumentumban egységesen sűrűn fordulnak elő és nincsenek semmilyen összefüggésben az adott dokumentumok tartalmával. Az angol terminológiát átvéve ezeket stopszavaknak hívjuk. Ezért a szó fontosságát befolyásolhatja a teljes gyűjteményben való előfordulásainak száma is, amit szoktunk gyűjteménytámogatottságnak is nevezni, vagy angolul collection frequency az elterjedt megnevezése. Fontos lehet továbbá az előfordulások eloszlása a korpuszon, ugyanis a koncentráltan kevés dokumentumban, de azokon belül nagy számban szereplő szavak valószínűleg fontosabbak is az adott dokumentumokra nézve, illetve nagyobb a dokumentumok közti megkülönböztető szerepük is. Ha nk-val jelöljük a tk szót tartalmazó dokumentumok számát, és N pedig a dokumentumok száma a korpuszban, akkor a dokumentum gyakoriságot a df = nk/N képlettel számolhatjuk. Ez a szám megadja, az adott szó ritkaságát, azt hogy a szó megléte, és annak száma mennyire különbözteti meg a dokumentumokat egymástól az adott korpuszban. Az idf (inverse document frequency) súlyozási sémában ennek inverzével számolunk. Maga a dokumentumgyakoriság inverzét számoló egyik elterjedt súlyozási függvény:

idf(tk) = log(N/nk)

A leggyakrabban használt tf-idf (term frequency and inverse document frequency) súlyozást úgy kapjuk, hogy a vektortérmodell szavakat reprezentáló tengelyeit az adott szavak idf által megadott relevanciájával arányosan súlyozzuk.

Tehát a szó-dokumentum mátrix adott dik súlya kiszámítható a

dik= fki \* idf(tk) képlettel

E mellett számos más képlete, és módosulata létezik, amikre most nem térek ki.

A tf-idf súlyozás értéke magas lesz a nagy megkülönböztető képességű, adott dokumentumra gyakori, de a korpuszra ritka szavaknál. Alacsonyabb lesz a korpuszban gyakoribb, vagy az adott dokumentumban, ritkább szavakban, és elhanyagolhatóan alacsony, akár zérus, az olyan szavakban melyek az egész korpuszban gyakran fordulnak elő.

### Módok a szöveg felbontására, illetve a szótár felépítésére

Mielőtt felírhatnánk a választott modellbe a dokumentumainkat, számos az azt megelőző előfeldolgozási lépésre lehet szükségünk. Az első ilyen lépés a reprezentációs egységekre való bontás. Ez leggyakrabban egyszerű szavakra való bontás, de egyes esetekben, mint például a kivonatolás a dokumentumok bekezdésekre történő, illetve mondat szintű felbontására is szükség van. Ezután a vektortér modell használata előtt meg kell határozni a szótár szavait is. Az összes egyedi szó szövegekből való kinyerése után, célszerű lehet ezek egy kis fontosságú, de esetenként sok szót tartalmazó részét elhagyni, ezzel jelentősen csökkentve, az indexelt szavak számát és a vektortér méretét.

### A strukturális szegmentálás

Dokumentumaink szövege számos hierarchiaszintbe sorolható. Ilyen strukturális egységek lehetnek például a kötet, rész, fejezet, szakasz, pont, illetve a bekezdés a mondatok és a szavak. Ezek és az ezekből összeállítható dokumentumtérkép automatikus meghatározása egyes fájlformátumokból kinyerhető információk alapján viszonylag egyszerű feladat is lehet, míg például az egyszerű szövegekből nehézkes, és esetenként koránt sem egyértelmű feladat is lehet.

### Mondatokra bontás

A szövegeink mondatokra való bontása automatizálható, de nem triviális feladat. Kézenfekvő lenne egyszerűen a mondatvégi írásjeleket nézni, és ezek szerint elszeparálni a mondatokat. Ilyenkor viszont álmondathatárokkal is találkozhatunk, hisz például a pont is előfordulhat mondat közepén, rövidítéseknél, sorszámoknál, dátumokban, vagy akár IP-címekben is. A feladat megoldására általában szabály alapú döntési algoritmust alkalmaznak, mely az ilyen téves mondathatárokra utaló, vagy ezeket cáfoló jelekre és tulajdonságokra tartalmaz szabályokat. Ezeket a szabályokat próbáljuk illeszteni a szövegeinkre. A tulajdonságok lehetnek felszíniek, mint például a jelet követő szó nagy kezdőbetűs e vagy sem, esetleg a jel egy szó közepén van e, stb. Az algoritmus az illesztés mellett kereshet gyakori álmondathatárokat tartalmazó listákban is, mint például a rövidítéslisták. Az ilyen listák nyelvenként eltérőek lehetnek, és időnként a lista karbantartására, frissítésére is szükség lehet. A szabályok tartalmazhatják, hogy a vélt határtól milyen pozícióra elhelyezkedő és milyen jellegű szavakra, vagy pontosabban tokenekre vonatkozik, és szokás megadni hozzájuk egy súly értéket is, hogy amennyiben egy vélt mondatvégre több szabvány is illeszkedik a nagyobb súllyal rendelkező juthasson érvényre. Legyen a szabályrendszerünk bármilyen komplex, és kifinomult is, előfordul, hogy az algoritmus nem képes jól dönteni, mert a mondat, vagy mondatok értelmezésével lehet csak jól meghatározni, a mondatok valódi határát.

### Tokenizálás

A token egy bizonyos karaktersorozat konkrét előfordulása, míg típusnak nevezzük az azonos tokenek osztályát. Így a típusok összessége alapján állítható elő a szótár. Ez a tokenizációs lépés minden lényeges szövegbányászati feladatnál, mind a keresés, mind pedig a rendszerezés jellegűeknél, egy fontos előfeldolgozási lépés. Hisz míg például a keresésénél ezekre illesztünk, és információkinyerésnél ezek potenciális információt hordozó entitások lehetnek, addig a dokumentumok rendszerezésénél, ezek olyan attribútumok lehetnek, melyek segítségével mérhetjük az egyes dokumentumok összetartozását, hasonlóságát. Ezek a feldolgozásunk legkisebb részelemei. A gyakorlatban sokszor egyszerű szavak, de mint a fenti definícióból látható lehetnek más szövegegységek, és karaktersorozatok, ritkán és szélsőséges esetekben egész dokumentumok is.

A tokenek alatt leggyakrabban mégis elszeparált szavakat értünk. Az ilyen tokenek előállítása elsőre egyszerű is lehet, hisz elég csak az írásjeleket elhagyva, a szóközöknél elválasztani a szöveget. De számos további kérdés is felmerül, mint, hogy hogyan kezelhetjük a kötőjellel elválasztott vagy aposztrófot tartalmazó szavakat, egyben vagy külön tokenekként. Kérdés továbbá, hogy az egyéb írásjeleket tartalmazó szavakat, mint a „kukac” és a „pont” karaktert tartalmazó e-mail címekre milyen szabályt használjunk. Továbbá hogyan kezeljük azokat a jeleket, amilyen az informatikusok többségének egyértelműen külön fogalmat reprezentáló C++, itt sem hagyható el a két pluszjel, hisz úgy már egy merőben más fogalmat kapnánk. Kérdés továbbá, hogy miként kezeljük a tulajdon neveket, különböző együttesek vagy csoportok több tagból álló neveit vagy a konkrét eseményt jelölő dátumokat. Természetesen az előbbi kérdésekre több helyes válasz is elképzelhető a feladatunktól függően, illetve a nyelvi és a felhasználási környezet is nagyban befolyásolhatja, hogy konkrétan milyen módon tokenizálunk.

Egy alternatív megoldás is létezik, amit karakter n-gramm módszernek nevezünk. Ekkor a szöveget n hosszúságú karaktersorozatokra bontjuk. Például: a „karakter 4-gramm” szöveget 4-grammokra bontva a következő eredményt kapjuk: kara, arak, rakt, akte, kter, ter˽, er˽4, r˽4-, ˽4-g, 4-gr, -gra, gram, ramm. Ez az eljárás jól használható nyelvek meghatározásakor, vagy más egyszerű osztályozási problémánál, esetleg olyan jellemzően távol-keleti nyelveknél szótárépítésre, ahol nincsenek a különböző szavak szóközökkel elválasztva egymástól. De nem célszerű a használata az európai eredetű nyelveknél olyan feladatoknál, ahol a szótár elemeinek értelmes egységeknek kell lenniük.

A tokenek típusai lesznek a nyelvi elemzés alapvető szövegreprezentáló egységei, és különböző szűrő eljárások után a típusok építik fel magát a szótárat is.

### Stoppszószűrés

Egyes szavak, illetve tokenek nagy számban fordulnak elő a dokumentumgyűjtemény dokumentumaiban és nincsenek szoros kapcsolatba azok témájával, ilyenek a névelők, a határozószavak, és a névutók. Ezek a stopszavak. A stopszószűrésre, vagy azok feldolgozás közbeni kiküszöbölésére láttunk egy matematikai jellegű módszert, a td-idf súlyozás szerepében. Tehát a stopszavakat a tokentípusok előállítása után, különböző dokumentumgyakorisági adataik alapján határozzák meg, majd a gyakorlatban esetenként manuálisan ellenőriznek, és ezek egy részét, a feladattól függően elhagyhatják. Napjainkban például keresésénél, esetleg maximum néhány tízes vagy akár tíz alatti szót hagynak el. Míg az elhagyások száma egyes osztályozási és csoportosítási feladatokban ez több százas vagy akár ezres nagyságú is lehet.

Ennek az az oka, hogy a keresésénél, az elhagyott stopszavak szerepelhetnek ismert kifejezésekben, különböző irodalmi vagy más művészeti művek címeiben, és elhagyásukkal ezekre való keresés is nehézkessé válna. Az indexállományok tömörítése, illetve a megnövekedett egyre olcsóbb tárkapacitás együttesen elérték, hogy az idők folyamán egyre kevesebb és kevesebb stopszót kelljen elhagynunk. Mindeközben a rendezési, és más pontos egyezést nem igénylő feladatoknál, ezek a szavak, konkrét jelentés nélkül, csak megnehezítenék a feldolgozást, nélkülük nagy mértékben lecsökkenhet a reprezentációs modell mérete. Gyakran szabják az adott korpuszra, az elhagyható szavak listáját, így biztosítva a lehető leghatékonyabb és mégis biztonságos feldolgozást.

### Lemmatizálás és Szótövezés

A stopszószűrés mellett léteznek nyelvészeti eszközök is a szótárak szűrésükre és méretének redukálására. Az alapötlet az, hogy a legtöbb nyelvben vannak a szavaknak különböző módosult alakjai. Ha ezeket a különböző alakokat egy közös alakként tárolnánk, egyes feladatokban elhanyagolható szóalak információ elvesztése mellett, nagy mértékben redukálhatnánk a modell méretét. A csökkentés mértékét befolyásolhatja az adott nyelv morfológiájának gazdagsága. Például ez a mérték az Angol nyelvnél megközelítőleg 40-70 százalék között lehet, addig a Magyarnál elérheti akár a 90%-ot is. A lemmatizálás, és a szótövezés is ilyen közös, úgymond kanonikus alakra hozó módszerek. A szó módosult alakjait a toldalékok, mint a képző, a jel, és a rag, adják. Ezek a toldalékok a szó előtt prefix-ként, és a szó után is állhatnak, körül is vehetik a szótövet, vagy a magyartól idegen módon, be is ékelődhetnek a szótőbe. A közös alakra hozásnál, ezeket a toldalékokat igyekszünk valamilyen módon leválasztani az adott szóról.

A lemmatizálás közben a szó normalizált szótári alakját azaz a lemmáját határozzuk meg. Nyelvészeti motivációjú eljárás közben, mindig értelmes szótári alakot hozunk létre. A lemmatizálás nem egyértelmű, ezért az eredmény több szó is lehet. Például a falunk szó szótári alakja a falu, de a fal szó is.

A szótövezés egyfajta szócsonkolást eredményez. Nem az a célja, hogy értelmes szótári szó jöjjön létre, sokkal inkább az, hogy a kialakított új szó a valódi szövegben is megtalálható legyen. A szótövező eredménye függ az adott algoritmustól.

Hasonlítsuk össze néhány példán a két megközelítést. A munkát és a munkám szavak lemmája a munka, szótöve algoritmustól függően a munka, vagy a munká. A ló lemmája a lovak, lovát és maga a ló szavaknak, de ezek szótöve a választott szótövező függvényében lehet lahet a lo, vagy a ló is.

## Információkinyerés

Az információkinyerésnél (Information Extraction - IE) a fő célunk nagy mennyiségű szövegből kigyűjteni a legfontosabb információkat, tesszük ezt olyan formában, hogy azt később akár egy relációs adatbázisba is beírhatjuk. Tehát a strukturálatlan adatokat kívánjuk valamilyen struktúrában összefoglalni. Napjaink meghatározó szövegbányászati kutatási iránya, hisz kiválóan alkalmas lehet nagy mennyiségű emberi munka kiváltására. Az adatok strukturált formába öntésével segíti a folyamatosan növekvő mennyiségű információinkat könnyebben kezelhető, és jobban áttekinthető, jobban ellenőrizhető és feldolgozhatóbbá tenni. Egyik dinamikusan fejlődő altípusa a nyelvközi információkinyerés (Cross-Language IE), melynél az adatokat több nyelvű szövegekből is összeszedjük, és táblázatba öntjük, majd elég csak a táblázatfejléceket lefordítani a kívánt nyelvre.

Az információ-kinyerés nagy mértékben feladatfüggő megoldásokat kíván, mert többnyire csak előre rögzített típusú elemeket vagyunk képesek a szövegekből hatékonyan kinyerni. Fontos lehet ismernünk a felhasználási szakirány egyes jellemzőit, ahhoz, hogy kideríthessük, hogy az adott feladat szempontjából mik a leginkább fontos attribútumok, amiket a szövegből ki szeretnénk gyűjteni, és azokat milyen módon, és formában lehet célszerű a felhasználók számára azokat prezentálni. A megoldásunk továbbá függeni fog az alkalmazási terület jellemző korpuszától is.

### Az információkinyerés összehasonlítása más főbb feladattípusokkal

Az információkinyerés a rendezési alaptípusokkal szemben, nem dokumentumokat akar valamilyen szempont szerint besorolni valamilyen kategóriába, mint az osztályozás, és nem célja továbbá dokumentumok valamilyen szempont szerinti csoportosítása sem.

A keresési alaptípusba tartozó információ-visszakereséssel szemben nem lokalizálja a felhasználó számára az információt, nem választja ki a dokumentumkollekció egy kifejezés alapján releváns dokumentumait. Ezzel szemben kiválasztja a feladat szempontjából fontos dokumentumokat, ezeket önmaga elemzi, és a kivont és formába öntött információkat juttatja vissza a felhasználónak. Az információkinyerő rendszerek feladatfüggőek, míg az információ-visszakeresés problémája és megoldása lehet általános. Így a keresőrendszerek kialakításához kevesebb a felhasználási szakterülettel kapcsolatos háttérismeretre van szükségünk. A keresés gyorsabb, és az eredményét a felhasználó manuálisan tekinti át így, ha esetlegesen hibát vét a rendszer annak nincs különösebb kockázata. Ellenben kevesebb munkát is igényel a felhasználótól az információ kinyerő által feldolgozott tömörebb adatmennyiség áttekintése. Az információkinyerő rendszerekben gyakran kell a többi feladattípus megoldásait felhasználni.

## Az információkinyerés fontosabb részfeladatai

A probléma megoldása során számos olykor egymásra épülő feladatot kell végrehajtanunk. Ezek rendre a névelem felismerés, a kereszthivatkozás azonosítás, a szereplők azonosítása, a szereplők közti relációk azonosítása, és az események illesztése.

### A névelem felismerés

A névelem, vagy angolul Named Entity, egy a valóságban előforduló, névvel ellátott objektum-egyed megnevezése. Az ilyen egyed lehet egy személy tulajdonneve, egy telefonszám, egy e-mail cím, vagy bármilyen más azonosító. Ezek felismerése olykor egyszerű reguláris kifejezésekkel történik, de főleg a tulajdonnév-felismerésnek nevezett alfajában bonyolultabb módszerekre is szükség lehet. A tulajdonnév-felismerés problémakörébe, nem csak a tényleges tulajdonnevek felismerését szoktuk érteni, de különböző együtt kezelendő tokensorozatokat is ide értünk. Például egyes főneveket, mint igazgatóság, vagy bevásárlóközpont esetenként érdemes lehet, a név részeként tekinteni. Manapság a névelem-felismerésre gyakran különböző gépi tanulókat alkalmaznak. Ezek a névelemeket, számukra új, de azért a tanulóadatokhoz hasonló tulajdonságú szövegeken, is eredményesen fedezhetik fel.

További példák lehetnek névelemekre, például hírekben a személyek, helyek, konkrét időpontok, szervezetek, egyes eszközök konkrét megnevezése, de biológiai szövegekben lehetnek akár gének, fajok, fehérjék megnevezései is. A specifikációs szövegekben is előfordulhatnak névelemek, különböző cégek, vagy konkrét személyek formájában, de ennél jellemzőbb, hogy különböző felhasználói szerepköröket kell jelölnünk.

### A szereplők azonosítása

A szövegbányászatban szereplők alatt többnyire, bizonyos előre definiált sablonnak megfelelő névelemek előfordulásait értjük. Az ilyen sablonokat a szövegbányászati terminológiában szokták esemény sablonoknak is nevezni. Az ilyen felderítendő szereplők lehetnek például vásárlók, termékek, vagy orvosi szövegekben például a betegségek tünetei vagy a gyógyszerek. További gyakori felderítendő szereplők a helység, és a szervezet.

A specifikációs szövegekben szereplők alatt különböző felhasználói kategóriákat értünk. Ezek a szereplők inkább, mint kategória osztályok jelennek meg a szövegekben, semmint egyedek közös sablonra illeszkedő sokaságaként, ezért a megtalálásukra is némiképp eltérő módot lehet célszerű keresni.

### A kereszthivatkozások azonosítása

Ez egy olyan lépés, melyben az azonos egyedre utaló tokensorozatokat és névelemeket rendeljük össze. Ennek egyik válfaja az anaforafeloldás, ami a különböző névmások és utalószavak a megfelelő névelemhez való rendelését jelenti.

### A szereplők közti relációk azonosítása

Ebben a lépésben az a célunk, hogy az előre meghatározott, és felderített szereplők közti kapcsolatokat ismerjünk fel.

Például üzleti hírekben vásárló és vásárolt árú, esetleg az árúhoz tartozó ár meghatározása, vagy orvosi folyóiratokban lehet egy adott betegség, és az azt gyógyító gyógyszer összerendelése. Specifikációkban ilyen feladat lehet egyes aktorok és az általuk végezhető használati esetek összerendelése, esetlegesen egyes használati esetek és aktorok egymás közti viszonyainak feltárása.

### Események felfedése és illesztése részfeladat

Esemény alatt többnyire egy bizonyos történést, vagy cselekvést értünk. Az ilyen események szövegeinkben gyakran igék formájában jelennek meg, és fontos lehet az azonosításuk. Miután az összes eddigi részfeladatot elvégeztük, rendelkezünk előre definiált események konkrét dokumentumbeli előfordulásával, illetve az azokhoz köthető szereplőkkel, és köztük fennálló különböző kapcsolatokkal. Ebben a lépésben pedig megoldást adunk arra, hogy ezek alapján különböző kérdésekre válaszolhasson a rendszer.

Ilyen kérdések lehetnek, hogy egy konkrét piaci szereplő milyen termékeket vett egy adott évben, vagy a szoftverkövetelmények követelmények területén, hogy egy adott használati esetben mely szereplők vehetnek részt, vagy éppen fordítva, egy adott aktor milyen feladatokat végezhet az adott szoftverrendszeren.

## Az információkinyerés során előforduló főbb nyelvészeti problémák

Az információkinyerő rendszer a működése érdekében szükség szerint számos nyelvtani jellegű problémát old meg.

A keresett eseményeket, vagy funkciókat gyakran cselekvések azaz igék írják le. Így fontos lehet ezeknek a szövegben való azonosítása. Előfordulhat, hogy egy szóalakhoz tartozik főnévi értelmezés is, mint például a nyúl szó is lehet egy cselekvést leíró ige, de főnévi értelmében már egy állatot jelöl. Egyértelmű, hogy a két különböző jelentésnek megfelelően, az ilyen szavak különböző feldolgozást is igényelnek. Egyes esetekben az ige helyett névszói frázis fejezi ki az adott cselekvéseket. Ezen feladatokban lehetnek segítségünkre a morfológiai, illetve szófaji elemzések.

Az igéknél gyakran el kell dönteni, hogy azok jelen jövő, vagy múlt idejűek. Fontos lehet továbbá, hogy az adott ige módja kijelentő, vagy feltételes. Emellett a felderített információkról el kell dönteni, hogy állító vagy tagadott értelemben szerepelnek.

Általában igaz az, hogy a különböző szereplők, események és relációk, a nekik megfelelő helyen, a rájuk jellemző mondatrészben találhatóak. Így fontos lehet meghatározni a mondatok szerkezetének meghatározására. Az alany az állítmány és a mondat további bővítményei, mint például a tárgy és a határozók, közti kapcsolatok ismerete is hasznos lehet.

Gyakran fordul elő, hogy a különböző az eseménnyel kapcsolatos névelemeket különböző beékelődött főnévi frázisok szeparálják el az eseményt jelző igétől. Fontos lehet ezek kiküszöbölése.

## A szabály és a statisztika alapú információkinyerési megközelítések összevetése

A legegyszerűbb mód az entitások, a relációk, és a különböző események kinyerésére, ha különböző nyelvi mintákat definiálunk, és ezek segítségével végezzük a kinyerést. Ilyen minta lehet például, ha egy tetszőleges hosszú nagybetűs szót a Kft. rövidítés követ, abból arra következtetünk, hogy a nagybetűs szó egy szervezet típusú névelem. Ennek a megközelítésnek a fő előnye, hogy meglehetősen nagy pontossággal és hatékonyan képes, az előre definiált típusok felfedezésére. De hátránya, hogy csak a szabályrendszer által lefedett típusokat deríti fel, illetve a szabályrendszer, folyamatos karbantartást igényel. A gyakorlatban az ilyen szabályokkal általában nehéz komplexebb problémákra jól működő megoldást adni. Ennek a fő okai, hogy a nyelvekben ugyanazt az információt többféleképpen is kifejezhetjük, minden lehetősség lefedésére irreálisan nagy, és nehezen kezelhető szabályrendszert kéne alkotnunk, különben megkockáztatjuk, hogy a jelenlévő hasznos információ csak töredék részét fedezzük fel. A szövegek nyelvtani elemzésének felhasználása nagy mértékben képes javítani a módszer használhatóságát. A szabályrendszer ilyen módú általánosítására kereteket alkalmazunk. A keret továbbá leírhat egy szereplőt, vagy akár szereplők egymáshoz való viszonyát. Egy keret meghatározhat nyelvi mintákat, megadva például olyan nyelvtani tulajdonságokat mint a szófaj, az eset vagy, hogy az adott szó egyes vagy többes számban szerepelhet. A keretek illesztésénél nem kell, hogy a vizsgált mondatban, a keret által meghatározott összes szerep jelen legyen, ez nagy mértékben megnöveli a keretek rugalmasságát. Léteznek általánosan felhasználható metakeret adatbázisok, mint például a Berkley egyetem által létrehozott FrameNet.

A szövegbányászat során, a legtöbb problémára lehet találni statisztikai eszközöket alkalmazó megoldást. Ez alól az információkinyerés sem kivétel. Gyakran alkalmazunk különböző gépi tanulókat, melyeket előre megjelölt, felcímkézett szövegeken tanítunk be. Ezek képesek az ismert tanító címkék alapján, a számukra új teszt, és remélhetőleg a végleges működési adatokon, az eddigieknek megfelelő, vagy valamilyen azokhoz hasonló új elemek felderítésére. Napjainkban ez a megközelítés egyre szélesebb körben terjed, ugyanis a szabály alapú rendszerek számos problémáját kiküszöböli. Az ilyen statisztikai alapú módszer általános, problémafüggetlen megoldást biztosíthat. Persze a rendszer jó működésére a továbbiakban is szükség van a jó minőségű, és lehetőleg a felhasználási környezethez illeszkedő tanítóadatra. Emellett a gyakorlatban a felidézése, szóval az, hogy milyen mértékben találja meg a jelenlévő információt, és a pontossága azaz, hogy mekkora mértékben tipizálja helyesen az információkat az esetek nagy részében jobb, mint a hasonló, de szakértői szabályokat használó rendszereké.

### A statisztikai megközelítések által használt elterjedt reprezentációs modellek

Két elterjedt megközelítés létezik, az egyik tokenek szekvenciáit, jellemzően teljes mondatokat használ a modellek alapegységeként, ezt hívják szekvencia alapú modellnek, míg a másik megközelítés magukat a tokeneket tekinti alapegységnek, ezt hívják token alapú modellnek. Az első módszer célja, hogy teljes mondatok osztálycimke-sorozatát, azok mondatstruktúráját, képes legyen előre jelezni. A szakirodalomban ezt szokták strukturált predikciónak is nevezni.

A másik módszer, mely külön kezeli a tokeneket, képes lehet új, számára eddig ismeretlen tokenek címkéjének automatikus előállítására. Az adott token környezetében lévő szavakra vonatkozó információk itt sem vesznek el, legfontosabb jellemzőikkel, az adott szó jellemzővektorát bővíthetjük.

A szekvencia alapú modellek többnyire valamelyik Markov-modell, és hasonló statisztikai modellek alapján működnek, míg a token alapú megközelítés a problémát felügyelt gépi tanulók segítségével igyekszik megoldani.

## Alternatív szövegelemzési megfontolások

„Abbot”; clean codeing;

## Az irodalomkutatás alapján leszűrt tanulságok

### Milyen lehet egy jó és hasznos követelményelemzést támogató eszköz.

### Az irodalomkutatás alapján milyen egy jó, és modern CASE eszköz

Véleményem szerint a jövő CASE eszközeinek a csoportos munka támogatására kell törekedniük, illetve arra, hogy a szoftverfejlesztés minél nagyobb területét lefedjék, és ezt úgy tegyék, hogy egymással a lehető legnagyobb összhangban működnek. Persze emellett a továbbiakban is törekedniük kell a munka, mimnél teljesebb, minél hatékonyabb, a lehető legkevesebb emberi beavatkozást igénylő, segítésére.

### A megismert szövegbányászati feladattípusok közül melyek, és milyen formában lehetnek relevánsak egy CASE eszközben

# A rendszerrel szemben támasztott követelmények

## Vázlatos felhasználói követelmények, illetve igények.

(VÁZLATOS)

A megvalósítandó rendszer egy követelmény specifikációs és követelmény analízist támogató intelligens, elosztott eszköz.

Az eszköz kell, hogy rendelkezzen webes felhasználói felülettel, mely lehetőleg minél több platformon helyesen jelenik meg.

Könnyen kezelhető és intelligens eszközöket biztosít a követelmény feltárás, elemzés és modellezés megkönnyítésére.

Módot kell adnia a különböző követelmény- és modell elemek egymáshoz való kapcsolására, és a meghatározott kapcsolatok segítségével történő navigálásra,

Esetleg jó lenne, ha lehetőséget biztosítana az összekapcsolt elemek egy azon nézetben való megjelenítésére.

Jó lenne, ha a rendszerben, különböző szótárak kezelésével lehetőséget biztosítana a rendszerben jelen lévő entitások azonosítására, esetleges szövegekben való kijelölésére, írás közbeni szöveg-kiegészítésére.

Jó lenne, ha a rendszerrel végzett munkát, minél előbb, akár a munkavégzés pillanatában, valós időben látnák a rendszer további felhasználói.

A rendszer módot biztosít a rendszer felhasználóinak különböző felhasználói szerepkörökhöz való rendelésére.

Jó lenne, ha a rendszer, az alap felhasználói szerepkörök finomhangolására is módot adna, az arra jogosultak számára.

A rendszer felhasználói interfészének törekednie kell egyes modern felülettervezési ajánlások betartására.

Jó lenne, ha a felhasználó által kijelölt elemek környékén megjelenő környezetfüggő menü átlátható módon tartalmazná a legfontosabb teendőket az adott elemen, ezzel elősegítve a kényelmes és hatékony kezelést. Az itt nem megjelenő menüelemek, egy a modern Office programokból megismert szalagmenüben kaphatnának helyet.

A rendszernek módot kell biztosítania a további bővítésének megkönnyítésére.

Jó lenne, ha egy követelmény megváltozása kijelölné felülvizsgálatra a kapcsolódó követelmény, és modell elemeket.

Jó lenne, ha le lehetne kérdezni, hogy mely követelmények lettek kifejtve az adott szinten, és melyek azok, amelyek még „érintetlenek”.

Az eltérő követelménytípusok különböző színű „kártyákon” jelenhetnének meg.

## A rendszerben megjelenő felhasználói szerepkörök

### A rendszer felhasználóinak használati eset nézete



### A rendszerben megjelenő, előre definiált felhasználói szerepkörök jellemzése

A felhasználó szerepköre megszabja, hogy az adott felhasználó milyen módon és milyen jogokkal férhet a rendszerhez. A különböző szerepkörök képviselői hasonló felhasználói felülettel rendelkeznek, de a számukra nem engedélyezett funkciók, nem jelennek meg előttük.

A **General Reader** (általános olvasó) van a hierarchia tetején és a legáltalánosabb jogokkal rendelkezik a rendszeren belül.

A **Software Architect**, a **Business Analyst**, illetve a **Convention Manager** az olvasó szerepkört specializálják, az ő funkcionalitását öröklik, illetve saját felelősségi körökkel bővítik. Tehát a General Reader lehetősségein felül további jogokkal, és lehetősségekkel rendelkeznek.

A **Convention Manager** a rendszerben megjelenő sablonokat, konvenciókat, és megkötéseket olvashatja.

A **Business Analyst** a felhasználói követelmények feltárására, és szerkesztésére specializálódott szerepkör.

A **Software Architect** a rendszerkövetelményeket, és rendszermodelleket állítja elő a magas szintű követelmények alapján.

A Convention Manager konkretizálása a **Company Contact**, aki a cégre vonatkozó általános konvenciókat szerkeszti, illetve a **Project Manager**, aki project szinten finomítja ezt. A **Project Manager** a Business Analyst leszármazottja is, így a magas szintű követelmények szerkesztésére is módja nyílik.

## Funkcionális követelmények

### A rendszer főbb felhasználói funkcióinak áttekintő használati eset nézete



A rendszer fő funkciói, az UML használati esetet modelljének OMG általi ajánlásainak megfelelően, kerettel el van határolva, a rendszer külső felhasználóitól.

A General Reader általános olvasási jogokkal rendelkezik, illetve megjegyzéseket is hozzáfűzhet a rendszer elemeihez.

A Software Architect kezeli a rendszerkövetelményeket, míg a Business Analyst a felhasználói követelményeket szerkeszti.

Míg a Convention Manager olvashatja a céges, illetve a project szintű megszorításokat, addig a Company Contact, illetve a Project Manager szerkeztheti is a felelősségi körének megfelelő konvenciókat.

A szerepkörök egy másik ága az Administrator, aki a felhasználókat, és azok szerepköreit és jogosultságait kezeli.

## Nem funkcionális követelmények

# A rendszer kialakítása során felhasználható technológiák és jellemzésük

## Szerver oldali technológiák

### ASP.NET

### Java

EE vagy Spring MVC esetleg valami egyszerűbb REST FW?

## A kliens oldali megjelenítés kezelése

### HTML5

SVG vagy Canvas

### CSS

## A kliens oldali dinamika kezelése

### JavaScript, és a TypeScript

## Kommunikációs technológiák

REST – a WinApi, illetve egy Javas megoldás;

SignalR

## Szövegbányászati csomagok

# Követelményanalízis

## A rendszer statikus modelljének elemzése

### Szerkezet áttekintő nézet

### Szerkezeti nézet

## A rendszer dinamikájának és folyamatainak elemzése

## Képernyő vázlatok

# Rendszerterv

## Képernyő tervek

## Csomagáttekintő nézet

## Részletes komponens leírások

## Részletes dinamikus nézetek

## Telepítési nézet és rendszerkörnyezet

# Az implementáció részletei

# Tesztelés

## Statikus tesztelés

## Dinamikus tesztelés

### Egység teszt

### Komponens teszt

### Integrációs teszt

### Rendszer teszt

### Átadási teszt

# Továbbfejlesztési lehetőségek

# A szakdolgozat tartalmi összefoglalója

# Irodalomjegyzék

Ian Somerwille: Szoftver rendszerek fejlesztése

Vég Csaba: Alkalmazásfejlesztés a Unified Modelling Language szabványos jelöléseivel

# Mellékletek