**Tanulmány a   
„Természetes nyelvek Inferenciája”  
című témakiíráshoz.**

# Bevezetés

A téma lényege a természetes nyelvű mondatok közti kapcsolat meghatározása. A feladat folyamán az ágensnek három féle lehetőség közül kell eldöntenie, hogy milyen kapcsolat van a neki megadott mondatpárok között. Az egyik mondat lehet az másik következménye (*Entailment*), ellentmondhatnak egymásnak (*Contradiction*), vagy nincs semmilyen jelentésbeli kapcsolat köztük (*Neutra*l). Ennek az eldöntéséhez szkség van amondatok közti kapcsolato

# Kísérleti körülmények

A mondatok szavait egy szóbeágyazásos módszerrel előállított adatbázisból származó szóvektorokkal ábrázolom. Kísérletemhez a MultiNLI előállított 40 dimenziós sűrű vektorait, és 100 dimenziós ritka (sparse) vektorait használtam. A mondatokat az NYU MultiNLI adatbázisából használom tanítópéldaként, és az általuk szolgáltatott tesztelő adatbázison tesztelem a logisztikus regresszióval előállított modellt. A modell kiértékelése végén egy 3\*3-as konfúziós mátrixot kapunk, amelynek oszlopaiban a modell által előrejelzett típusok vannak, soraiban meg a mondatpárok valódi típusai. A mátrix főátlójában találhatóak a modell által helyesen megjósolt értékek száma.

Az első esetben mondatok szavainak megfeleltetett 40 dimenziós sűrű vektorokat összeátlagoltam. Így kaptam két vektort, amely a két mondatot reprezentálja. Majd az ezen vektorok keresztszorzatából állítottam elő a mondatpárt reprezentáló vektort. Ezekre mondatpár vektorokra tanítottam be egy logisztikus regressziós modellt. A kísérletek elvégzéséhez a Java-s Smile csomagokat használtam. A 40 dimenziós sűrű vektorokhoz, és az azokból előállított 1600 dimenziós mondatpár vektorokhoz a LogisticRegression osztályt, a 100 dimenziós ritka vektorokhoz és az azokból előállított 10’000 dimenziós mondatpár vektorokhoz a Maxent osztályt használom.

A Logisztikus regresszióhoz a mondat szavainak szóbeágyazásos vektorainak átlagát használom. De a mondatokban előfordulhatnak olyan szavak, is amelyek nem fordulnak elő a szóbeágyazásos adatbázisban. Ezek a szavak nem játszanak szerepet a mondatot reprezentáló vektor kiszámításában. Tehát az adatbázisban szereplő szavak vektorainakaz összegét veszem,majd ezt osztom az adatbázisban megtalálható szavak számával. Majd az így előállított mondatok vektorai közül az elsőt oszlopvektorként kezelem, a másodikat sorvektorként. A két vektort a mátrixszorzás szabályainak megfelelően összeszorzom, majd az így kapott mátrix sorait egymás után fűzöm, hogy előálljon a korábban említett 1600 dimenziós mondatpár vektor.

A 100 dimenziós esetben is hasonlóan járok el az átlagolásnál. De a Java-s Maxent osztály csak az azt kapja meg, hogy hol nem nulla a vektor értéke, ezért az átlagolás után a vektornak azon értékeit, amelyek egy bizonyos küszöbértéknél nagyobbak megtartom. Ezután a 40 dimenziós esetben leírt módszerrel előállítom a 10000 dimenziós mondatpár vektorokat.

A harmadik kísérletben visszatértem a sűrűvektoros szóábrázoláshoz, de itt figyelembe vettem a MultiNLI-s adatbázisban szereplő szintaxis fákat. A mondat szavai ezen szintaxisfák leveleiben helyezkednek el. Ezekhez a levélelemekhez tartoznak a szavak szófajai. A többi csúcspontban a szóösszetételek illetve mondatrészek típusai találhatók meg, amelyeknek gyerekei az adott egységhez tartózó szavak, illetve mondatrészek tartoznak. A mondatok szavaihoz tartozó súlyokat ezen szintaxisfák alapján határozom meg úgy, hogy a gyökérelemből kiindulva, a gyökérelemet nem beleszámolva az egyes típusokhoz tartozó súlyok számtani közepét veszem. A súlyokat az egyes típusok relatív gyakorisága alapján határozom meg, úgy hogy az egyes relatív gyakoriságokat még megszoroztam, egy egynél nagyobb számmal. Ezek alapján már különbséget tudok tenni a szavak mondaton belül betöltött szerepében. Ezek után a két mondathoz kiszámítom a szavak vektorainak súlyozott átlagát, majd ezeknek a keresztszorzat helyett a két mondat vektor különbségét veszam.

# Eredmények

Az első esetben előállított mondatpár vektorokra betanított Logisztikus regressziós modell által helyesen osztályozott vektorok aránya 36.45%, 100’000 tanítópélda alapján.

A kapott eredmény mondatpár típusokra is le van bontva. Azaz azt is figyelembe veszem, hogy az adott típusúnak tippelt mondatpárok közül mennyi helyes.

Ellentmondásoknak tippelt 35.42% ténylegesen ellentmondás

Semlegedsekből 37.18% ténylegesen semleges mondatpár

Következményekből 37.49% ténylegesen következmény.

Az egyszerű átlagolással, és keresztszorzattal számolt mondatpár vektorokra fölállított modell nem sokkal lett jobb a véletlenszerű találgatásnál.

A kísérlet Konfúziós mátrixa:

***tényleges típus***

**cont. neut. ent.**

**cont.** *1469 1335 1508*

***előrejelzett*** **neut.** *777 771  806*

**ent.** *967 1017 1165*

A második esetben a 100 dimenziós ritka vektorokat használtam a Maximum entróptiával (Maxent), amely még rosszabbul teljesített mint a sűrűvektoros megoldás a logisztikus regresszióval. A lenti konfúziós mátrixból az látszik, hogy a maximum entrópiás módszer minden mondatpárra ellentmondást tippelt, amely így egy 32.74% eredményt produkált. Ez az eredmény megmaradt a 0.75, 0.5, illetve a 0.25 küszöbértékeknél is.

***tényleges típus***

**cont. neut. ent.**

**cont.** 3213 3123 *3479*

***előrejelzett*** **neut.** *0 0 0*

**ent.** *0 0 0*

A harmadik kísérlet az előzőekhez képest jobban teljesített, mivel a tesztadatbázison futatva 40.62% egyezőséget mutatott, amely egy lényeges előrelépés, a véletlenszerű találgatáshoz képest. A jobb eredmény eléréséhez nagyobb szerepet játszhatott az, hogy a mondatok vektorainak nem keresztszorzatát, hanem különbségét veszem, hiszen a három féle osztály a mondatok jelentésének egymástól való eltérése alapján próbálja osztályozni a mondatokat.

A jobb eredmény ellenére a modell továbbra is eléggé rossz eredményt mutat az ellentmondások megtalálásában, ahogyan ez a lenti típusokra lebontott eredményeken is jól látszik.

Ellentmondások: 38.93%

Semlegesek: 42.43%

Következmény: 41.37%

***tényleges típus***

**cont. neut. ent.**

**cont.** 1619 1126 *1413*

***előrejelzett*** **neut.** *696 1104 802*

**ent.** *898 893 1264*

# Projekt megvalósításához felhsznált irofalmak:

Java-s Smile-packege:

<https://haifengl.github.io/smile/>

<http://haifengl.github.io/smile/api/java/index.html>  
<http://haifengl.github.io/smile/api/java/smile/classification/Maxent.html>

Felhasznált adatbázisról:

<http://www.nyu.edu/projects/bowman/multinli/>

Szóbeágyazásokról:

<http://ruder.io/word-embeddings-1/>

<https://ronxin.github.io/wevi/>

<http://allennlp.org/>

Konfúziós mátrixokról:

<https://en.wikipedia.org/wiki/Confusion_matrix>