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

# Bevezetés

A téma lényege a természetes nyelven leírt mondatok reprezentációja, illetve az ezen mondatok közti kapcsolat meghatározása. A feladat folyamán a számítógépnek 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).

# 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.

A az első esetben mondatok szavainak megfeleltetett 40 dimenziós sűrű vektorokat összeátlagoltam. Így kaptam két vektort, amely a két mondatot reprezentálja. Majd 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 meg a Maxent osztályt.

# Eredmények

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

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, mivel minden mondatpárra ellentmondást mond.

***tényleges típus***

**cont. neut. ent.**

**cont.** *1469 1335 1508*

***előrejelzett*** **neut.** *0 0 0*

**ent.** *0 0 0*