DEEP LEARNING ALAPÚ SZERZŐAZONOSÍTÁS

Füleki Fábián, Jani Balázs Gábor, Torner Márton

Csapat: LoremIpsum

Szerzőazonosítás

- Egyedi stílus felismerése
  - Kézírás
  - Szavak és szófordulatok használata
  - Mondat szerkezet
- Stilometria
- Plágiumdetekció

Korábbi munkák

- Sentence level vs. Article level megközelítés(46% vs 69.1%)
- GRU, LSTM (69.1% vs 62.7%)
- (Sziámi hálózat)
  - Plágiumdetekció 99.8%!

Dataset - reuter_50_50 (C50)

- nyers szöveg
- 50 szerző 50+50 cikke (5000db)
- Összehasonlíthatóság más cikkekkel

- https://archive.ics.uci.edu/ml/datasets/Reuter_50_50

Adatok előkészítése

- Szavak reprezentálása számokként
  - Szóvektorok
  - Part of Speech tag-ek
- Azonos hosszúságú mondatok
- További trükkök

SPACY

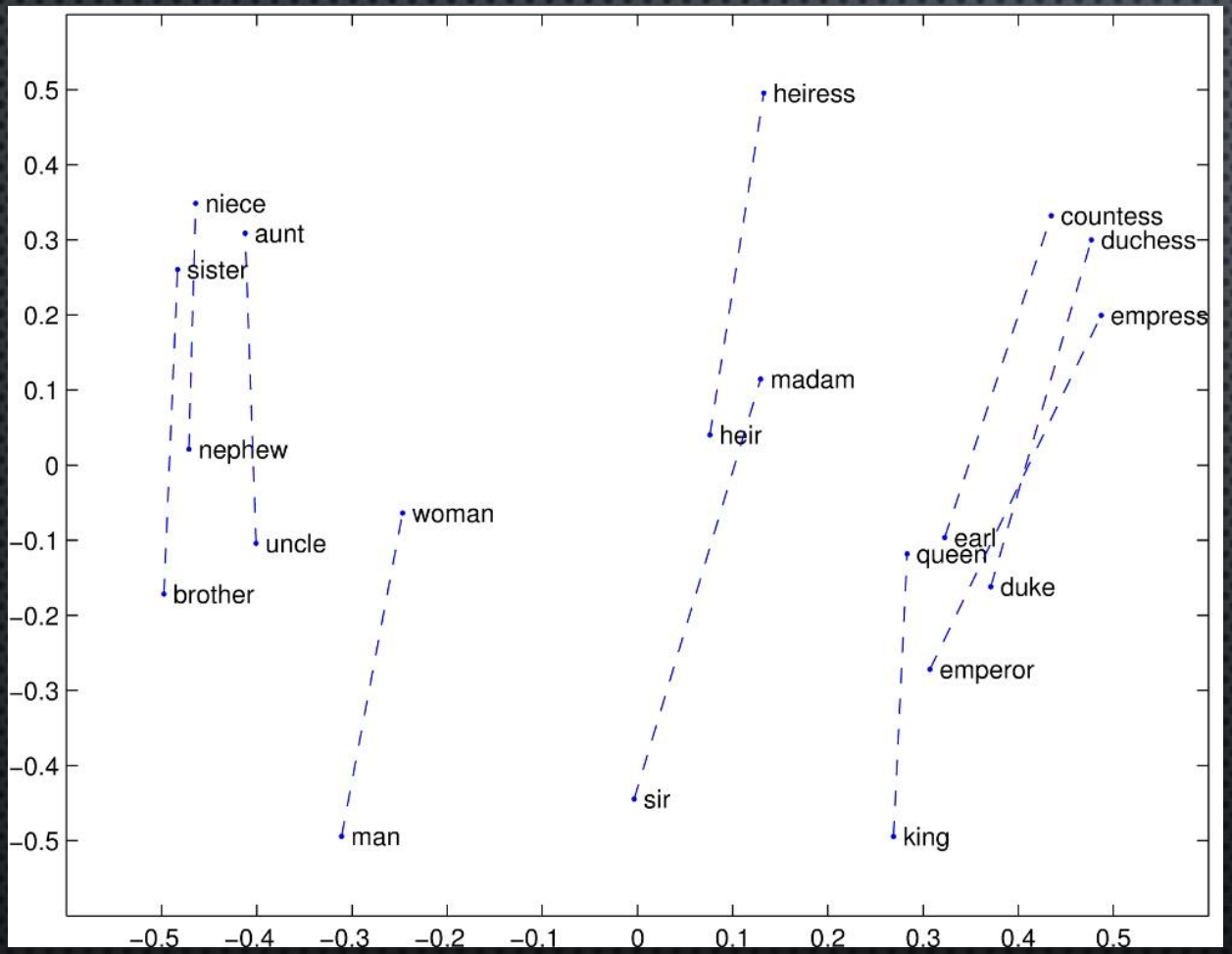
- GloVe
- Part of Speech Tagging
- 301 dimenziós vektor
  - szó reprezentáció + pos tag

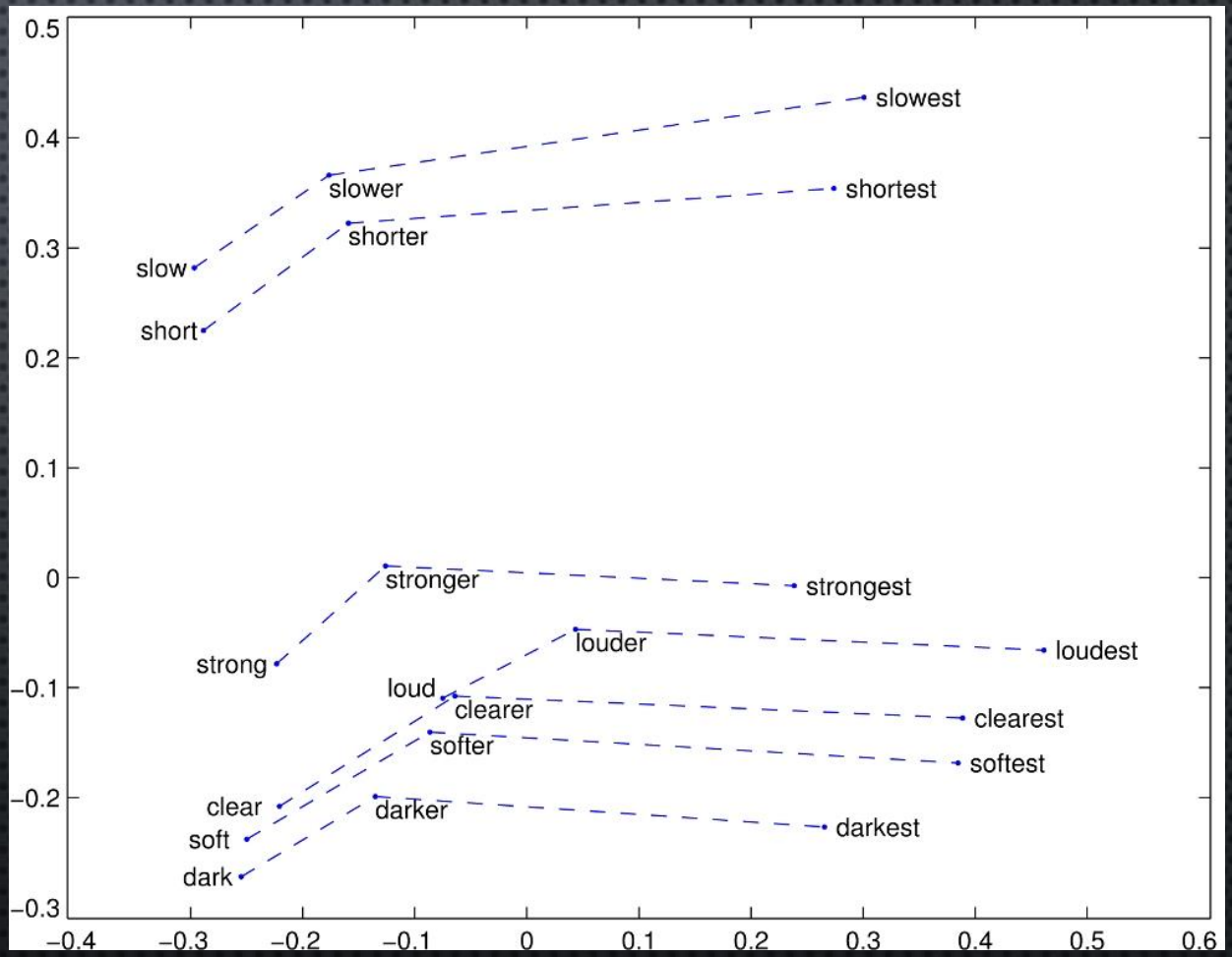
- Kiszűrt részek
  - Stop words, punctuation
  - Ritka szavak (nincs megfelelő reprezentáció az adatbázisban)

GLOBAL VECTOR FOR WORDS

- *„GloVe is an unsupervised learning algorithm for obtaining vector representations for words.”*
- A szavak reprezentálása vektorokként
- Általunk használt bázis: 685 000 egyedi vektor (szó)
  - 300 dimenziós vektorok

- https://nlp.stanford.edu/pubs/glove.pdf
- https://nlp.stanford.edu/projects/glove/

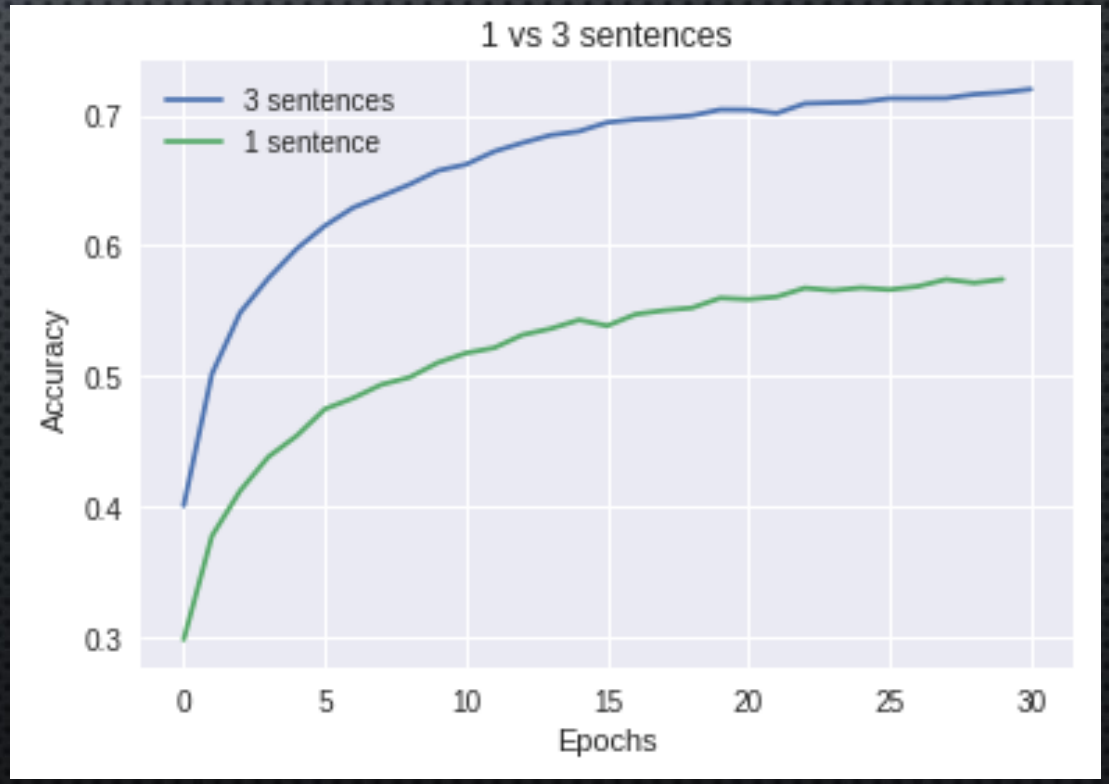

heiress
niece
aunt
countess
duchess
sister
empress
madam
nephew
heir
woman
uncle
earl
queen
brother
duke
emperor
sir
man
king


slowest
slower
shortest
shorter
slow
short
stronger
strongest
louder
loudest
strong
loud
clearer
clearest
softer
softest
clear
darker
soft
darkest
dark

	0	1	2	3	4	5	6	7	8	...
text	The	Internet	may	be	overflowing	with	new	technology	but	...
vector	[0.27204, -0.06203, -0.1884, 0.023225, -0.0181...	[-0.50955, 0.088231, -0.32273, -0.40398, 0.003...	[-0.042501, 0.090773, -0.11918, 0.12372, -0.19...	[-0.059177, 0.10653, -0.21613, -0.086178, 0.00...	[0.074908, -0.036973, 0.082992, -0.31622, 0.22...	[-0.099534, 0.028202, -0.23189, 0.094477, 0.12...	[0.34046, 0.13752, -0.20643, -0.4555, 0.19251,...	[-0.32298, 0.38883, 0.4586, -0.5227, -0.064451...	[-0.01689, 0.17402, -0.30247, -0.30063, 0.2141...	...
pos_str	DET	NOUN	VERB	VERB	VERB	ADP	ADJ	NOUN	CCONJ	...

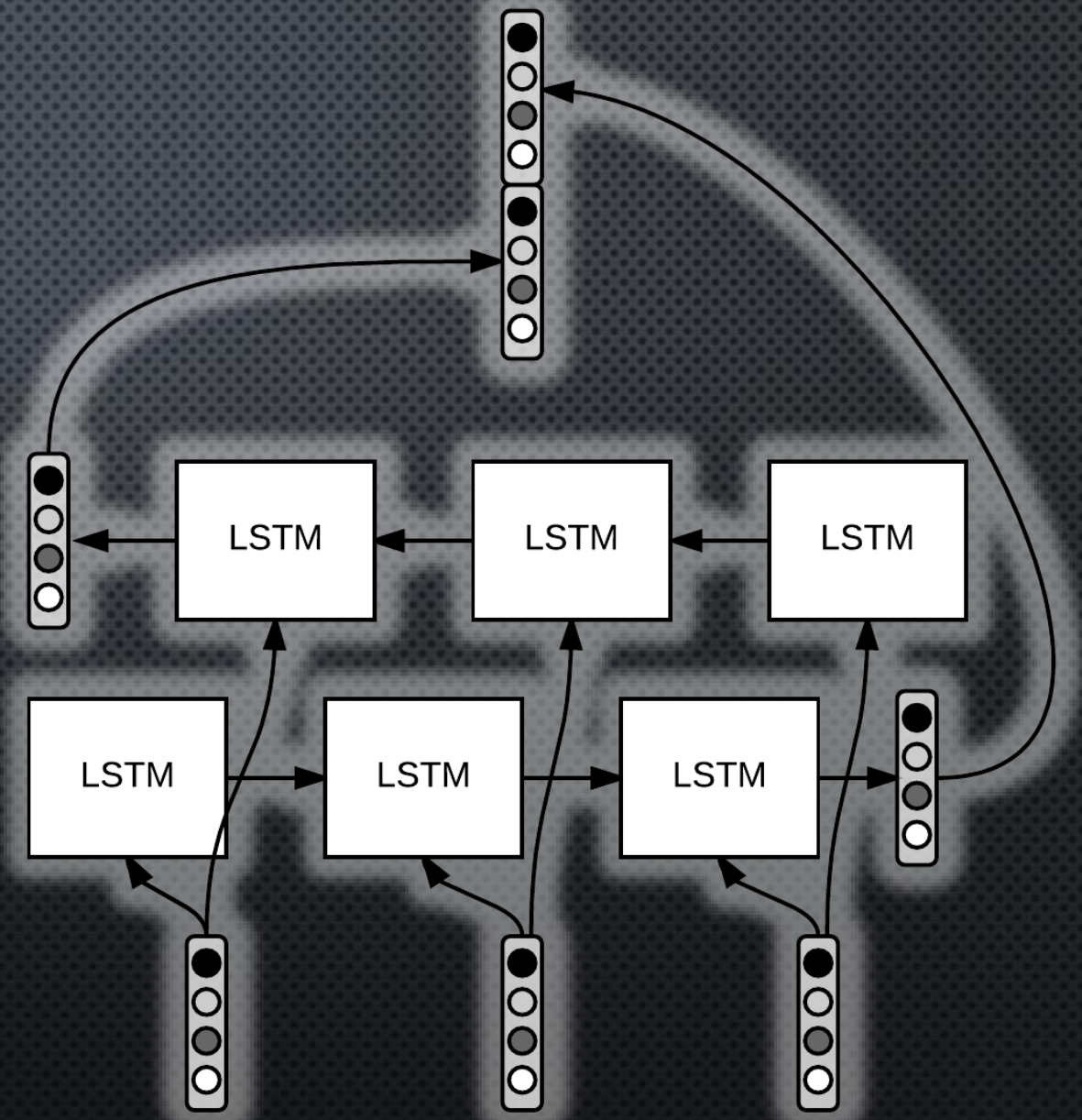
MONDATOK ÖSSZEFOGÁSA

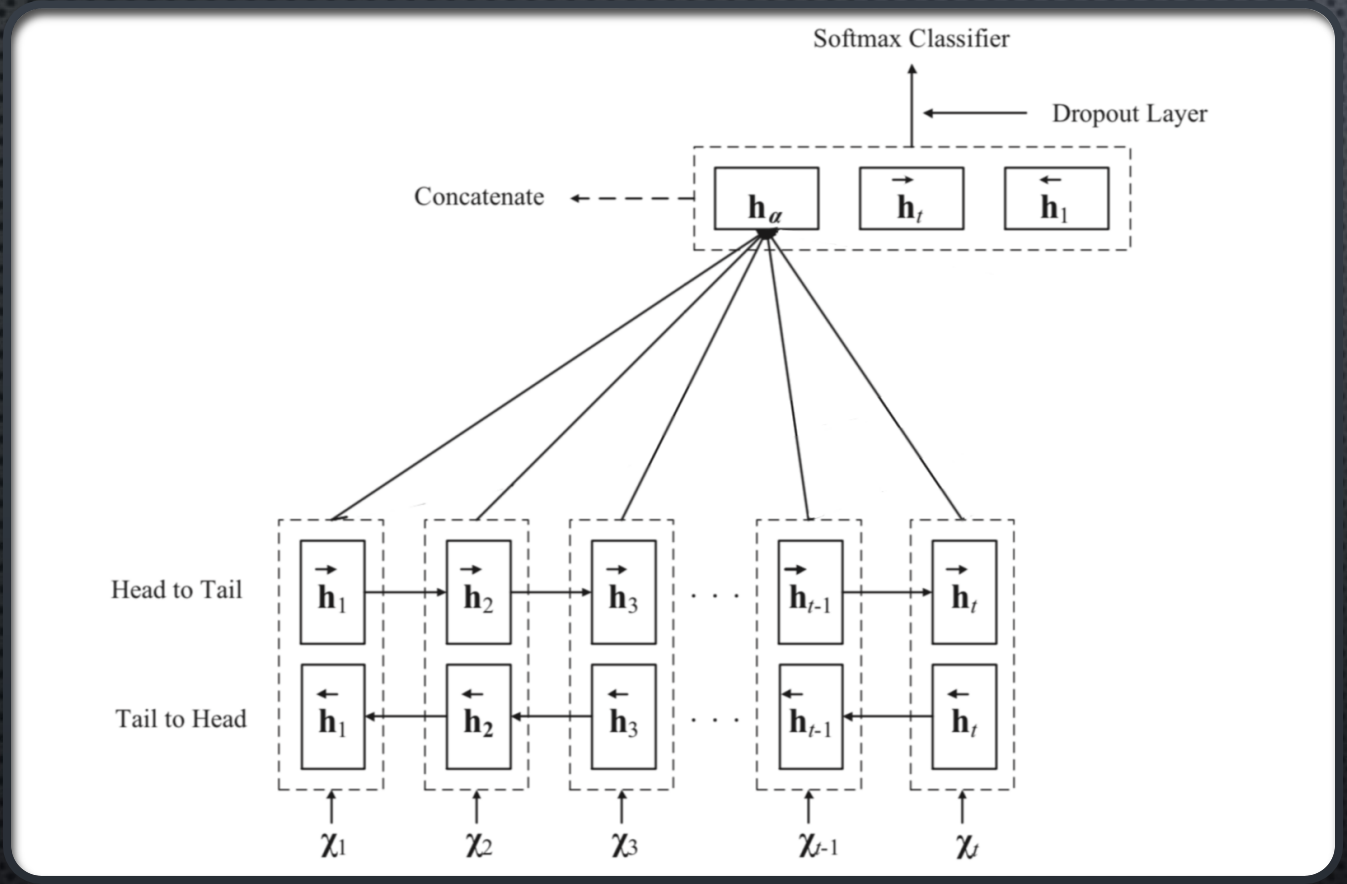
- AZONOS HOSSZÚSÁG
  - RÖVIDRE VÁGÁS
  - BŐVÍTÉS/KIEGÉSZÍTÉS
- 3 MONDAT EGYÜTTES HASZNÁLATA
  - KIFEJEZÉSEK ÁTFEDÉSE



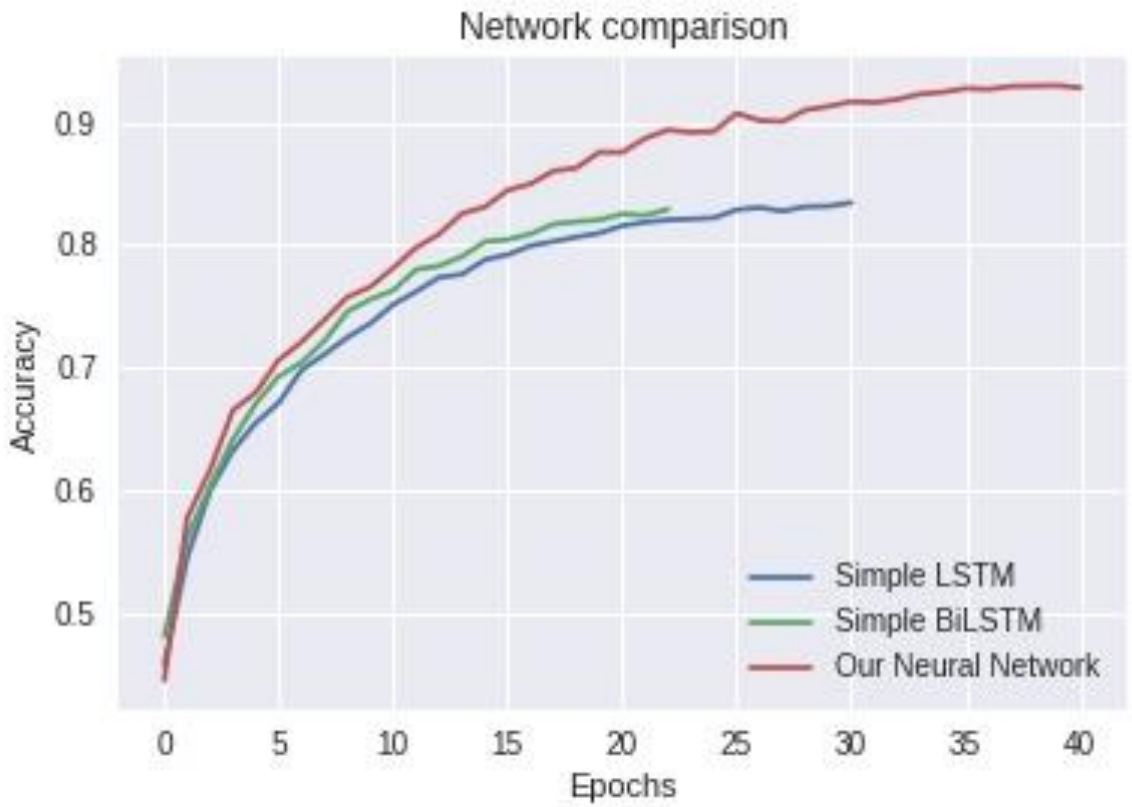
ARCHITEKTÚRÁK

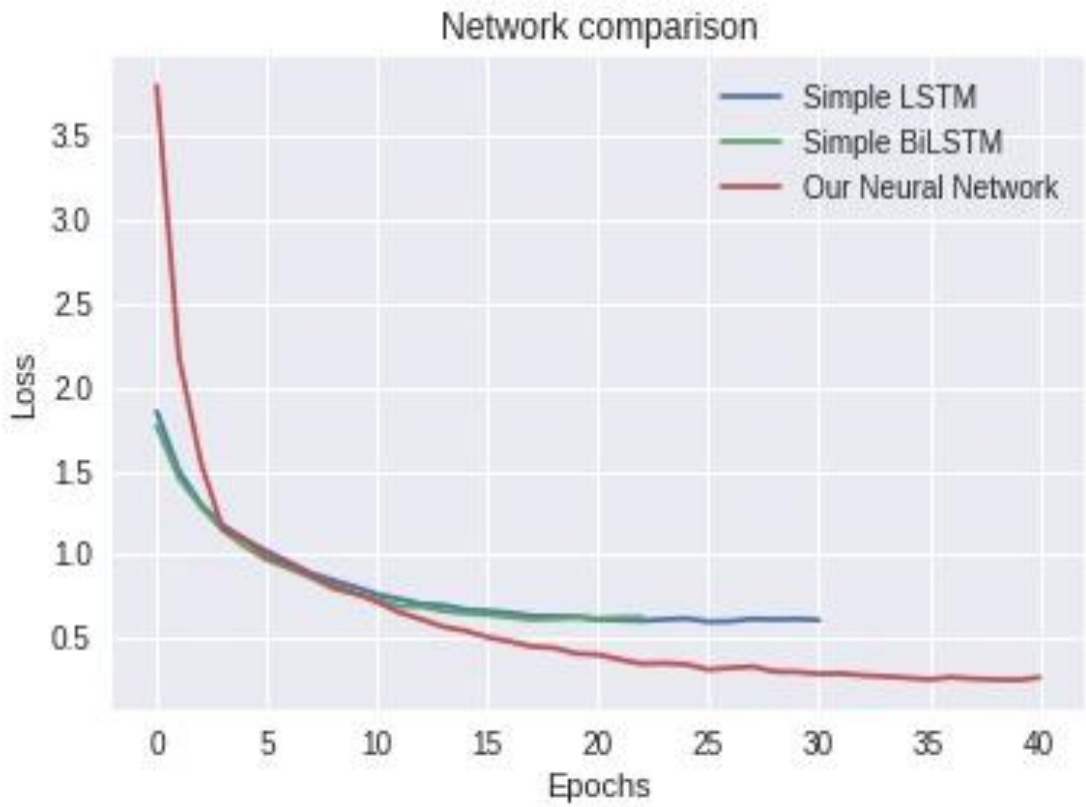
- LSTM / BiLSTM
  - Time distributed
  - Average pooling
  - Softmax classifier
- BiLSTM
  - Kataforák és anaforák





SAJÁT MEGKÖZELÍTÉS


Network comparison
Accuracy
Epochs
Simple LSTM
Simple BiLSTM
Our Neural Network


Network comparison
Loss
Epochs
Simple LSTM
Simple BiLSTM
Our Neural Network

Hiperoptimalizálás paraméterei

- Batch-ek mérete: 256, 512, 1024, 2048
- Neuronok száma: 128, 256, 512, 1024
- Dropout, rekurrens dropout: 0 ⟶ 1
- Optimizer: sgd, rmsprop, adagrad, adam

Hiperoptimalizálás konklúziói

- Legjobb optimalizációs módszerek: adam, rmsprop
- Legjobb futtatás paraméterei: 512-es batch méret, 512 neuron, (dropout: 0,5 körül), adam
- Túl sok neuron / túl kicsi batch méret: sokkal lassabban tanul csak egy kicsit jobban

Összehasonlítás

- Qian, Chen, Tianchang He, and Rao Zhang. "Deep Learning based Authorship Identification."
  - Article level LSTM – 69.1% accuracy (BEST)
  - http://web.stanford.edu/class/cs224n/reports/2760185.pdf
- Saját: 94%
  - 3 sentence bundles
  - 512 batch size, 512 neuron, 0.53 dropout, 0.5 recurrent dropout

Merre tovább?

- Ismeretlen csoport
- Attention
- Más előfeldolgozás (nagyobb bundle-ök, article level, stb.)
  - Plusz szemantikai információk

KÖSZÖNJÜK A FIGYELMET!