Assignment-Discussion
Vector-based POS Tagging

Goda Nagakalyani, 214050010
Karthika N. J, 21405007
Atul Kumar Singh, 22m0823

9 October,2022

Problem Statement

- Given a sequence of words, produce the POS tag sequence
- Technique to be used: HMM-Viterbi-vector and Word2Vec vectors – FFNN
- 5-fold cross validation
- Use Universal Tag Set (12 in number)
- ’.’ ,’ADJ’ ,’ADP’ ,’ADV’ ,’CONJ’ ,’DET’ ,’NOUN’ ,’NUM’ , ’PRON’ , ’PRT’ , ’VERB’ , ’X’

Overall performance – Viterbi symbolic

- Precision : 0.940198283308899
- Recall : 0.938524036355917З
- F-score (3 values)
  - F1-score : 0.9386931499010809
  - F0.5-score : 0.9394349573883816
  - F2-score : 0.938428994887414

Overall performance – Viterbi Word2Vec

- Precision : 0.9604942731149146
- Recall : 0.960550498867571
- F-score (3 values)
  - F1-score : 0.9604629662664248
  - F0.5-score : 0.960467122841927
  - F2-score : 0.9605015364677237

Overall performance – FFNN with BP

- Precision : 0.9492435479418486
- Recall : 0.9470962732745691
- F-score (3 values)
  - F1-score : 0.9469783997495549
  - F0.5-score : 0.9479818532819128
  - F2-score : 09467951214084014

Performance Comparison

	HMM Viterbi Symbolic	HMM Viterbi Word2Vec	FFNN – BP Using Word2Vec
Precision	0.940	0.960	0.949
Recall	0.938	0.960	0.947
F0.5 score	0.939	0.960	0.947
F1 score	0.938	0.960	0.946
F2 score	0.938	0.960	0.946

Classification report – Viterbi Symbolic

Tag	Precision	Recall	F1-score
.	0.98	1.00	0.99
ADJ	0.87	0.89	0.88
ADP	0.92	0.97	0.94
ADV	0.90	0.87	0.88
CONJ	0.99	0.99	0.99
DET	0.92	0.99	0.95
NOUN	0.95	0.92	0.93
NUM	0.99	0.80	0.88
PRON	0.93	0.96	0.94
PRT	0.91	0.85	0.88
VERB	0.97	0.92	0.94
X	0.17	0.35	0.23

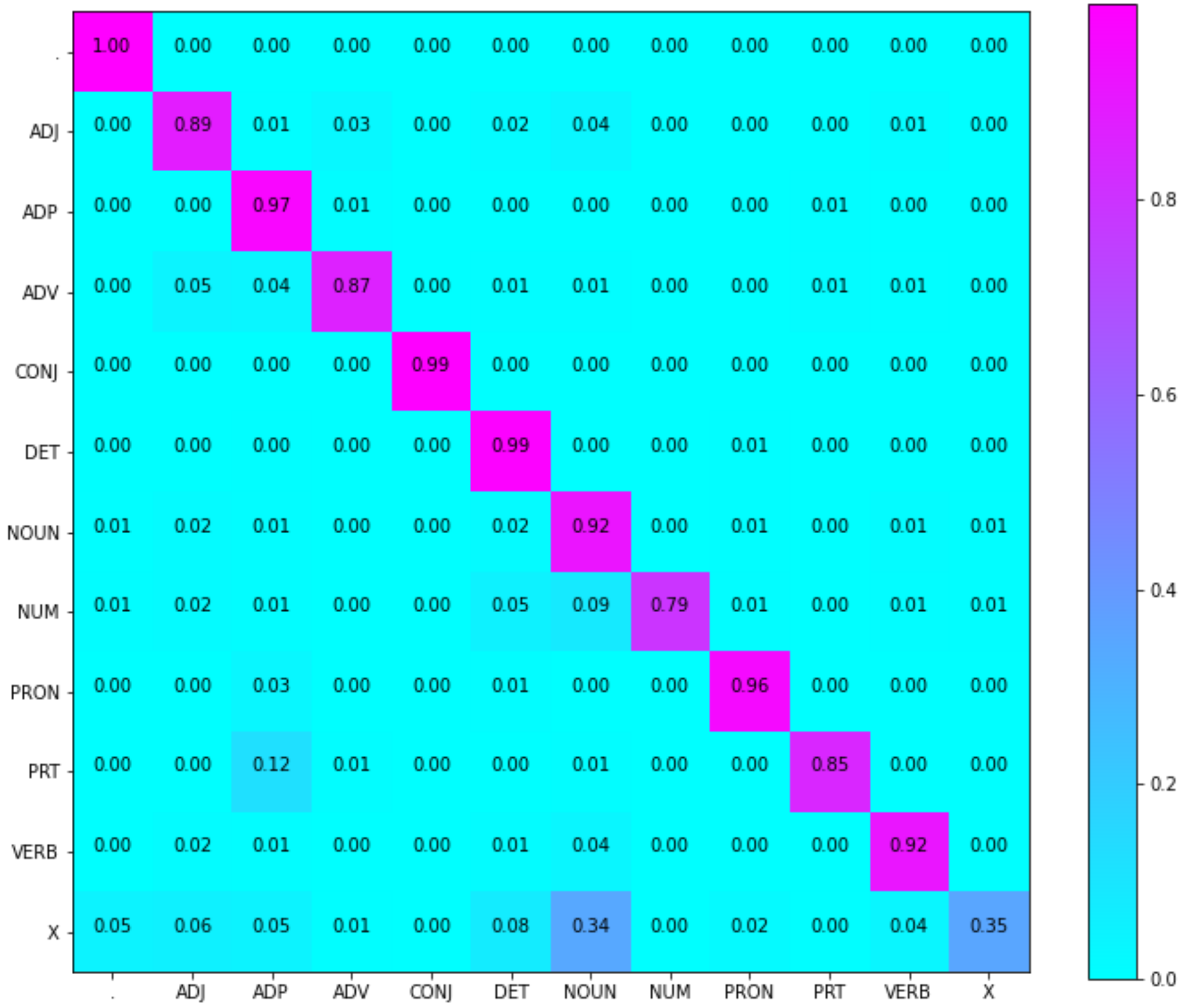
Classification report – Viterbi Word2Vec

Tag	Precision	Recall	F1-score
.	1.00	1.00	1.00
ADJ	0.92	0.92	0.92
ADP	0.95	0.97	0.96
ADV	0.91	0.89	0.90
CONJ	0.99	0.99	0.99
DET	0.97	0.99	0.98
NOUN	0.96	0.96	0.96
NUM	0.97	0.91	0.94
PRON	0.96	0.98	0.97
PRT	0.91	0.90	0.90
VERB	0.97	0.95	0.96
X	0.52	0.55	0.53

Classification report – Word2Vec with FFNN-BP

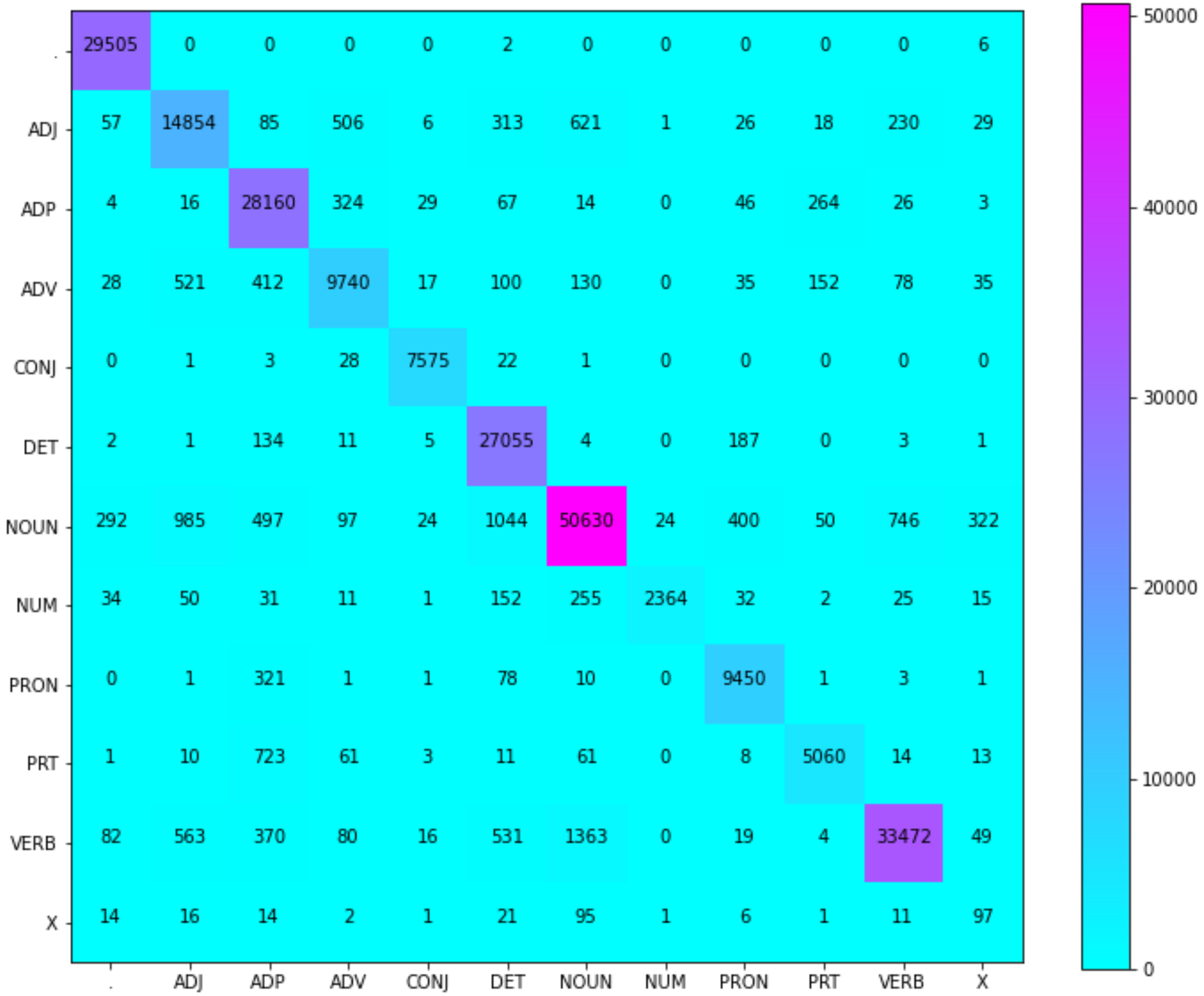
tag	Precision	Recall	F1-score
.	1.00	1.00	1.00
ADJ	0.91	0.89	0.90
ADP	0.93	0.91	0.92
ADV	0.90	0.86	0.88
CONJ	0.99	1.00	0.99
DET	0.99	0.98	0.99
NOUN	0.94	0.96	0.95
NUM	0.95	0.93	0.94
PRON	1.00	0.94	0.97
PRT	0.69	0.92	0.79
VERB	0.96	0.95	0.95
X	0.76	0.39	0.52

Confusion Matrix – Viterbi symbolic



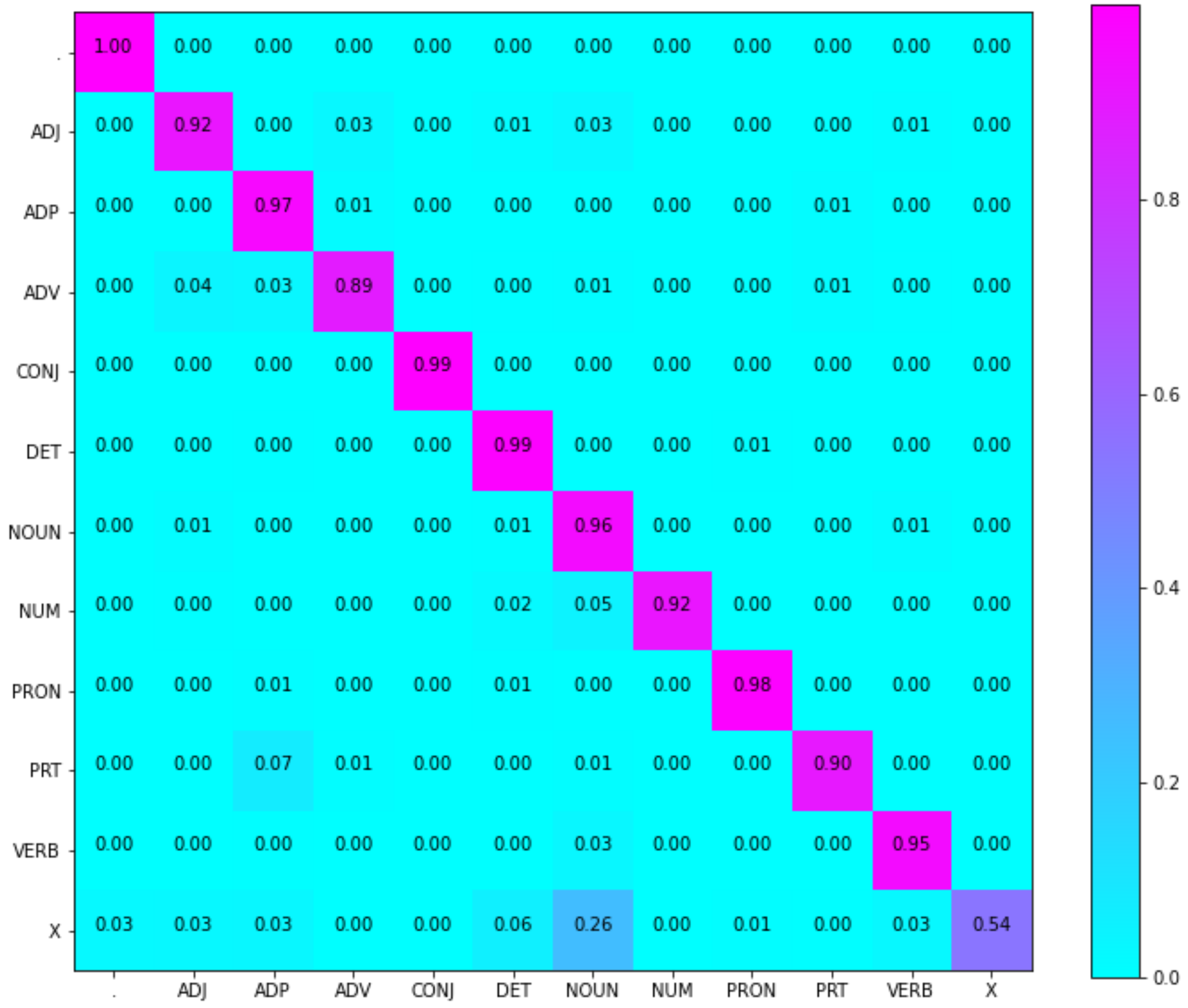
	.	ADJ	ADP	ADV	CONJ	DET	NOUN	NUM	PRON	PRT	VERB	X
.	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ADJ	0.00	0.89	0.01	0.03	0.00	0.02	0.04	0.00	0.00	0.00	0.01	0.00
ADP	0.00	0.00	0.97	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
ADV	0.00	0.05	0.04	0.87	0.00	0.01	0.01	0.00	0.00	0.01	0.01	0.00
CONJ	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DET	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.01	0.00	0.00	0.00
NOUN	0.01	0.02	0.01	0.00	0.00	0.02	0.92	0.00	0.01	0.00	0.01	0.01
NUM	0.01	0.02	0.01	0.00	0.00	0.05	0.09	0.79	0.01	0.00	0.01	0.01
PRON	0.00	0.00	0.03	0.00	0.00	0.01	0.00	0.00	0.96	0.00	0.00	0.00
PRT	0.00	0.00	0.12	0.01	0.00	0.00	0.01	0.00	0.00	0.85	0.00	0.00
VERB	0.00	0.02	0.01	0.00	0.00	0.01	0.04	0.00	0.00	0.00	0.92	0.00
X	0.05	0.06	0.05	0.01	0.00	0.08	0.34	0.00	0.02	0.00	0.04	0.35

Confusion matrix – Viterbi symbolic (Numbers)



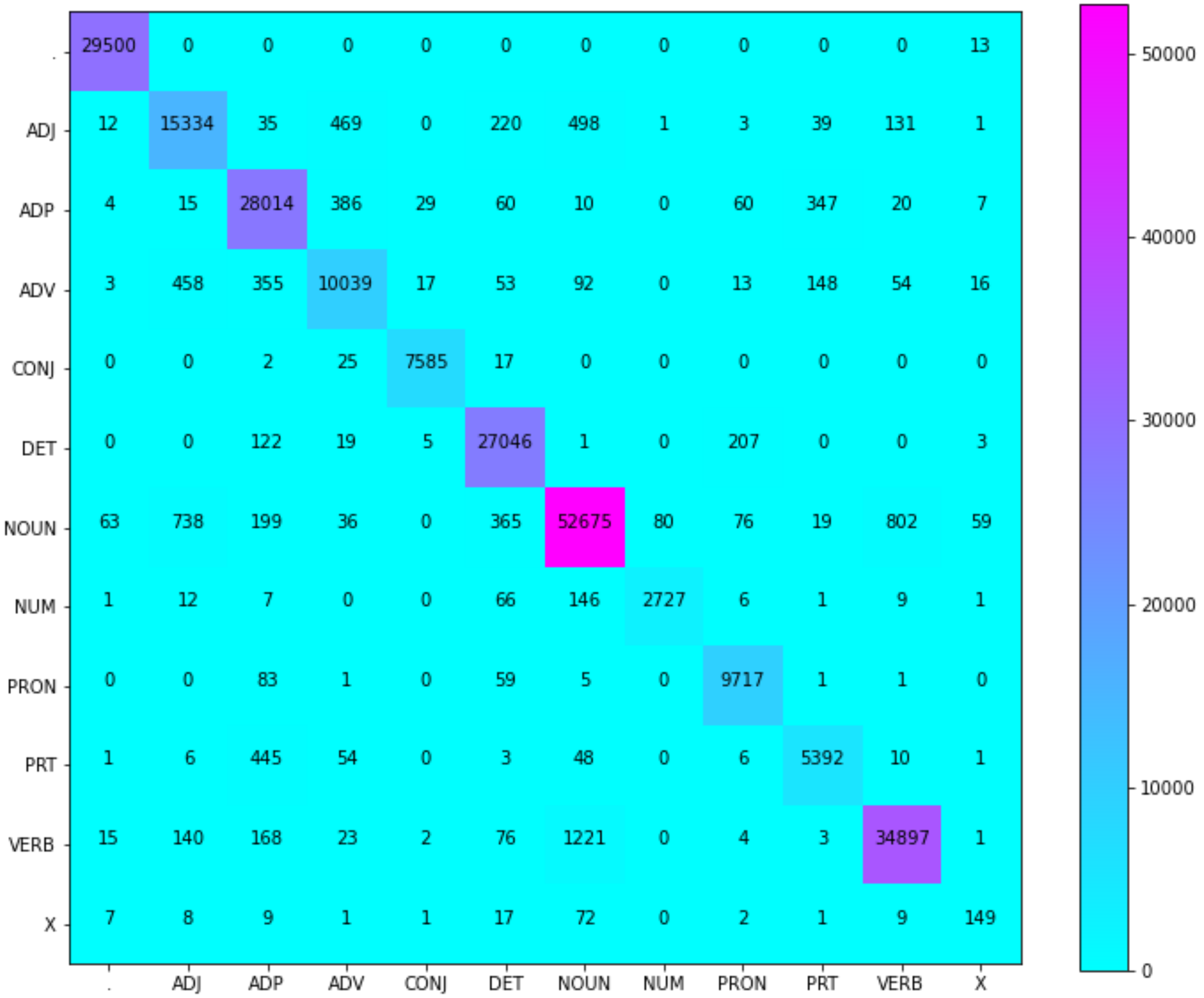
	.	ADJ	ADP	ADV	CONJ	DET	NOUN	NUM	PRON	PRT	VERB	X
.	29505	0	0	0	0	2	0	0	0	0	0	6
ADJ	57	14854	85	506	6	313	621	1	26	18	230	29
ADP	4	16	28160	324	29	67	14	0	46	264	26	3
ADV	28	521	412	9740	17	100	130	0	35	152	78	35
CONJ	0	1	3	28	7575	22	1	0	0	0	0	0
DET	2	1	134	11	5	27055	4	0	187	0	3	1
NOUN	292	985	497	97	24	1044	50630	24	400	50	746	322
NUM	34	50	31	11	1	152	255	2364	32	2	25	15
PRON	0	1	321	1	1	78	10	0	9450	1	3	1
PRT	1	10	723	61	3	11	61	0	8	5060	14	13
VERB	82	563	370	80	16	531	1363	0	19	4	33472	49
X	14	16	14	2	1	21	95	1	6	1	11	97

Confusion Matrix – Viterbi Word2Vec



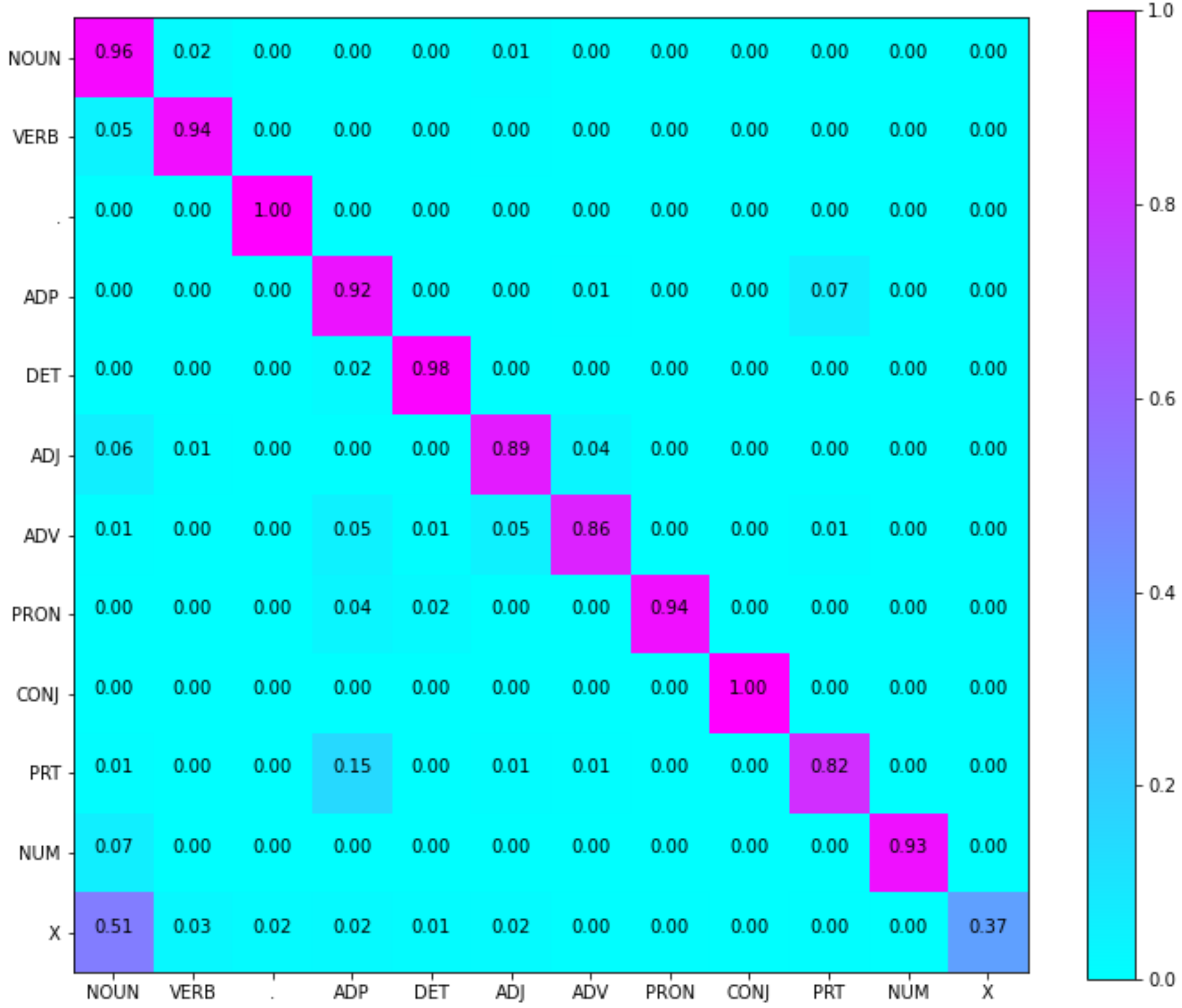
	.	ADJ	ADP	ADV	CONJ	DET	NOUN	NUM	PRON	PRT	VERB	X
.	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ADJ	0.00	0.92	0.00	0.03	0.00	0.01	0.03	0.00	0.00	0.00	0.01	0.00
ADP	0.00	0.00	0.97	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
ADV	0.00	0.04	0.03	0.89	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00
CONJ	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DET	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.01	0.00	0.00	0.00
NOUN	0.00	0.01	0.00	0.00	0.00	0.01	0.96	0.00	0.00	0.00	0.01	0.00
NUM	0.00	0.00	0.00	0.00	0.00	0.02	0.05	0.92	0.00	0.00	0.00	0.00
PRON	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.98	0.00	0.00	0.00
PRT	0.00	0.00	0.07	0.01	0.00	0.00	0.01	0.00	0.00	0.90	0.00	0.00
VERB	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.95	0.00
X	0.03	0.03	0.03	0.00	0.00	0.06	0.26	0.00	0.01	0.00	0.03	0.54

Confusion matrix –Viterbi word2vec (Numbers)



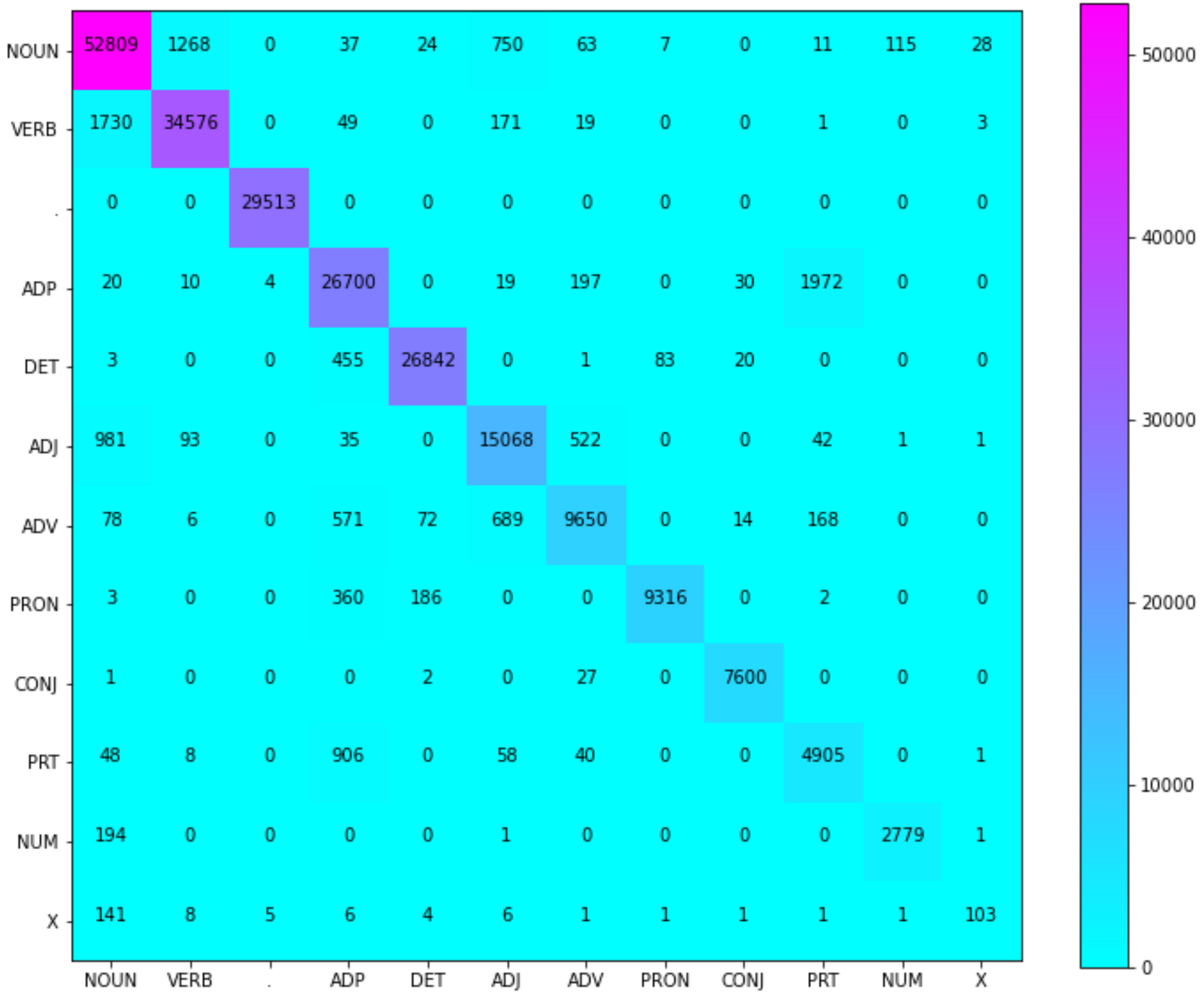
	.	ADJ	ADP	ADV	CONJ	DET	NOUN	NUM	PRON	PRT	VERB	X
.	29500	0	0	0	0	0	0	0	0	0	0	13
ADJ	12	15334	35	469	0	220	498	1	3	39	131	1
ADP	4	15	28014	386	29	60	10	0	60	347	20	7
ADV	3	458	355	10039	17	53	92	0	13	148	54	16
CONJ	0	0	2	25	7585	17	0	0	0	0	0	0
DET	0	0	122	19	5	27046	1	0	207	0	0	3
NOUN	63	738	199	36	0	365	52675	80	76	19	802	59
NUM	1	12	7	0	0	66	146	2727	6	1	9	1
PRON	0	0	83	1	0	59	5	0	9717	1	1	0
PRT	1	6	445	54	0	3	48	0	6	5392	10	1
VERB	15	140	168	23	2	76	1221	0	4	3	34897	1
X	7	8	9	1	1	17	72	0	2	1	9	149

Confusion Matrix – Word2Vec with FFNN-BP



	NOUN	VERB	.	ADP	DET	ADJ	ADV	PRON	CONJ	PRT	NUM	X
NOUN	0.96	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
VERB	0.05	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
.	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ADP	0.00	0.00	0.00	0.92	0.00	0.00	0.01	0.00	0.00	0.07	0.00	0.00
DET	0.00	0.00	0.00	0.02	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ADJ	0.06	0.01	0.00	0.00	0.00	0.89	0.04	0.00	0.00	0.00	0.00	0.00
ADV	0.01	0.00	0.00	0.05	0.01	0.05	0.86	0.00	0.00	0.01	0.00	0.00
PRON	0.00	0.00	0.00	0.04	0.02	0.00	0.00	0.94	0.00	0.00	0.00	0.00
CONJ	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
PRT	0.01	0.00	0.00	0.15	0.00	0.01	0.01	0.00	0.00	0.82	0.00	0.00
NUM	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.93	0.00
X	0.51	0.03	0.02	0.02	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.37

Confusion matrix – word2vec with FFNN-BP (Numbers)



	NOUN	VERB	.	ADP	DET	ADJ	ADV	PRON	CONJ	PRT	NUM	X
NOUN	52809	1268	0	37	24	750	63	7	0	11	115	28
VERB	1730	34576	0	49	0	171	19	0	0	1	0	3
.	0	0	29513	0	0	0	0	0	0	0	0	0
ADP	20	10	4	26700	0	19	197	0	30	1972	0	0
DET	3	0	0	455	26842	0	1	83	20	0	0	0
ADJ	981	93	0	35	0	15068	522	0	0	42	1	1
ADV	78	6	0	571	72	689	9650	0	14	168	0	0
PRON	3	0	0	360	186	0	0	9316	0	2	0	0
CONJ	1	0	0	0	2	0	27	0	7600	0	0	0
PRT	48	8	0	906	0	58	40	0	0	4905	0	1
NUM	194	0	0	0	0	1	0	0	0	0	2779	1
X	141	8	5	6	4	6	1	1	1	1	1	103

Confusion matrix Analysis

- Confusion of tag ‘X’ with ‘NOUN
  - There is no prior reason for X-NOUN because we predict tag X for extra characters like {ersatz, esprit, dunno, gr8, univeristy, etc} so there might some words like for word vector ignore spell mistakes and return word vector closed to NOUN tagged word.
- Confusion of tag ‘PRT’ with ‘ADP’
  - Keywords for PRT : at, on, out, over per, that, up, with
  - Keywords for ADP : on, of, at, with, by, into, under
  - Example
    - The city expects the higher rooming houses to bring **in** an additional $40000 a year. (PRT)
    - I like walking **in** the park during winters.(ADP)

Confusion matrix analysis

- Confusion of tag ‘NUM’ with ‘NOUN’
  - A cardinal number, five plus one (NUM)
  - A hit in which the ball crosses the boundary line of the field without a bounce, counting **six** runs for the batsman
- Confusion of tag ‘NUM’ with ‘DET
  - This is **one** of the best items of the city.
  - They had a strong attraction for **one** another.
- Confusion of tag ‘ADV’ with ‘ADJ’
  - He is travelling **underground** by subway. (ADV)
  - This is an **underground** vegetarian restaurant. (ADJ)
  - Take the dog **outside**. (ADV)
  - There is a news from the **outside** world. (ADJ)

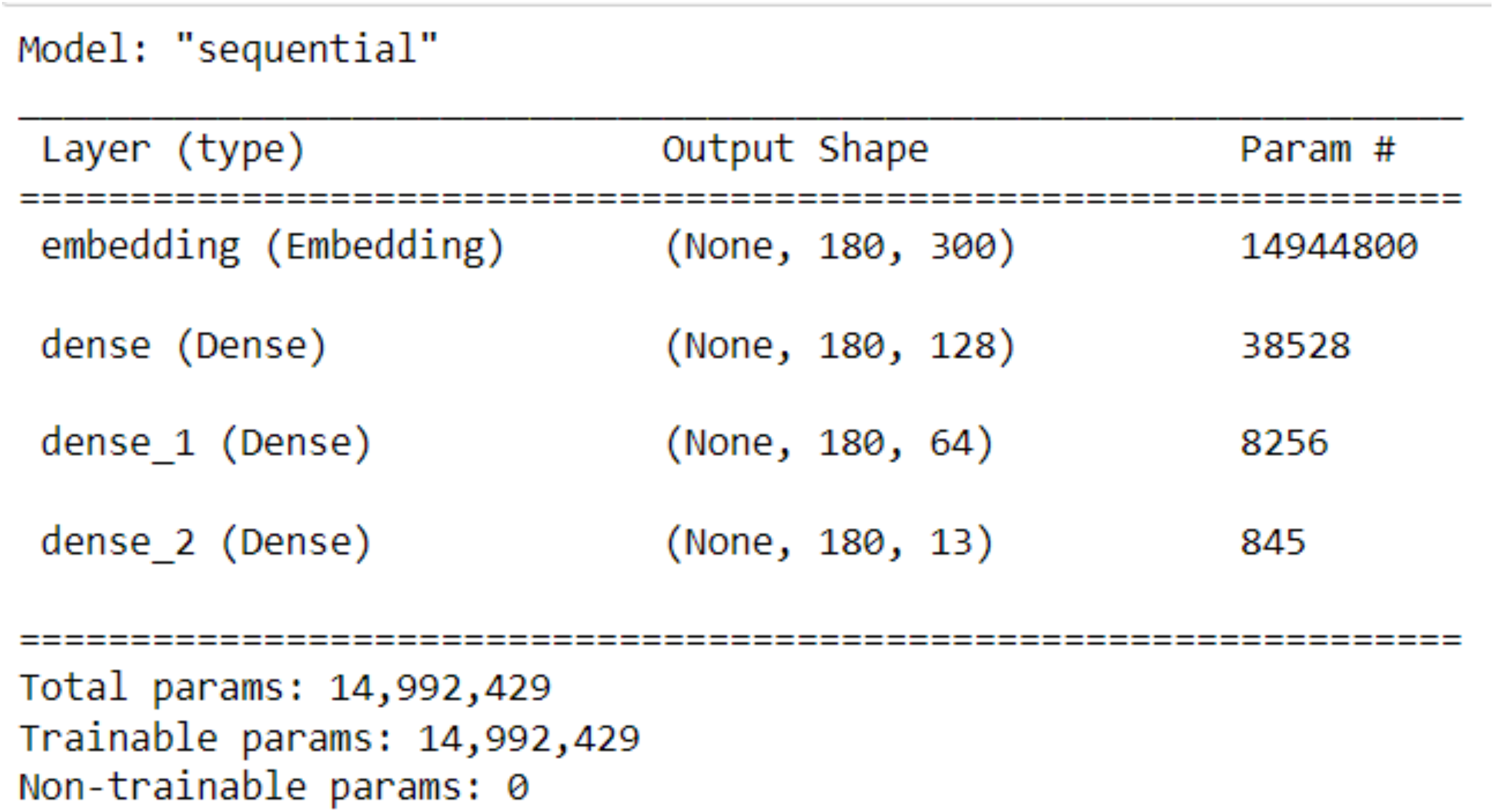
Confusion matrix analysis

- Confusion of tag ADJ and NOUN
  - Fill in the white space below(ADJ)
  - He has a speck in the white of his eye.(NOUN)
  - She is a young woman.(ADJ)
  - This is a game for young and old.(NOUN)
- Confusion of tag ADV and ADP
  - The school is close **by**.
  - He came **by** the highway.
- Confusion of tag VERB and NOUN
  - He likes to be in an excited **state**.
  - He came here to **state** a problem.
  - This **print** is too large for footnotes.
  - He’d rather **print** than use longhand.

Confusion matrix analysis

- Confusion of tag PRON and ADP
  - **That** is her mother.
  - **That** these magazines also deluded the krims of the world is unfortunate but inevitable.

FFNN-BP Model Details



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 180, 300)          14944800

 dense (Dense)               (None, 180, 128)          38528

 dense_1 (Dense)             (None, 180, 64)           8256

 dense_2 (Dense)             (None, 180, 13)           845

=================================================================
Total params: 14,992,429
Trainable params: 14,992,429
Non-trainable params: 0
```

Data Processing and Data Sparsity

- **<u>Data-Processing</u>**
  - For part-1
    - Firstly We added start token and end token to each sentences and make each word in lower letter for reducing computation.
    - We have created 12x300 matrix which stores word vector representing each tag That is useful in calculating cos-similarity while handing unknown word in lexical probability.
  - For part-2
    - We have used a Embedding layer in order to pass the embedding matrix which contains the word vectors for all the words in the vocabulary.
    - Then we have 2 hidden Dense layers with activation function as ‘relu’
    - Then we have a final output layer with its activation function set as ‘softmax’

- **<u>Obtaining Word-vectors</u>**
  - We used word2vec model trained on google news dataset and for extracting word vectors from that we used genism module.

- **<u>Handling Unseen Words</u>**
  - We have calculated 12x300 matrix which stores word vector corresponding to each tag (this word vector is sum of all word’s vector in train-set which tag as respective tags).
  - When algorithm encounter unseen word we calculate cos-similarity between tag vector and word vector and map it to [0,1] by exponential function.