

# **Assignment 4**

## **201014560 & 057931354**

### **Remark on Running Environment**

We ran on windows on python 3.6.4. We had to make few modifications on data\_util.py (add "rb" and "wb" when opening files) and in rnn.py (replace raw\_input() with input()). However, in Nova we left the original code untouched to allow it to run on python 2.7.

4 Že

3128 07/7  
x 026 780

(1)

Jack Daniel's main lawyer, Tim Cook : 1(a)  
S.A.

He is the 16th Jack Daniel's lawyer  
of the company (since 1916).

Apple has many fans around the  
world

If we look at Apple's fans, it is  
the fan entity.

It's about 100 million people who like  
Apple products. They are called "Appleheads".  
They are mostly young people, mostly  
men, mostly from the United States.  
They are very passionate about Apple.

It's part of the Apple community.  
Speech

If we look at the Apple community,  
it's very large and diverse.

$\frac{1}{2}(2n+1)D : e \approx 3.14 \cdot 1.6$   
 $(2n+1)DXH : n \approx 3.14 \cdot 1.6$   
 $HXC : 0 \approx 3.14 \cdot 1.6$

$O((2n+1)D)$  is approximately 16 e  
and  $e \approx 2.718 \approx 1.6$

$O((\partial_{W4}) DH + H)$   
 $O(HC + C)$   
 $O((\partial_{W4}) DH + HC)$   
 PL cold & sun & sea  
 21/2 83/2 CC

	Per	ORG	Loc	Misc	O	I	(d)
Per	2959	52	64	13	61		
ORG	149	1638	116	61	128		
Loc	44	104	1884	14	48		
Misc	us	60	55	995	105		
O	38	57	23	32	4269		

83%. (i) Fi f 1/6/18 (c) (d)

(d) (d) (d) (d) (d) (d)  
 (d) (d) (d) (d) (d) (d)  
 (d) (d) (d) (d) (d) (d)  
 (d) (d) (d) (d) (d) (d)  
 (d) (d) (d) (d) (d) (d)

(d) (d) (d) (d) (d) (d)  
 (d) (d) (d) (d) (d) (d)  
 (d) (d) (d) (d) (d) (d)  
 (d) (d) (d) (d) (d) (d)  
 (d) (d) (d) (d) (d) (d)

There

11/10/18

I like many in Tel AVV.

Tel AVV plus  $\partial_{W4}$   $\partial_{W4}$  Tel end  $\partial_{W4}$   $\partial_{W4}$   
 Tel  $\partial_{W4}$   $\partial_{W4}$  Tel  $\partial_{W4}$  Tel  $\partial_{W4}$  Tel  $\partial_{W4}$  Tel

I like apple; it's the best tech  
company in the world

I like apple it's Sri 1 first it's allowed me  
~~the 2nd place is sony~~  
Colony ~~it's very nice~~ 13282 located at  
105/1161 pali patti 17200 721306 132  
apple computers ~~apple~~ 121306 132

1.  $L = \text{softmax}(Wx + b)$  (I)  
 ~~$\approx D \times H \times W$~~   
 $\approx (QW+1)D \times H \times W$  -  
 $\approx QW(DH + W)$  (II)  
 $O(H^2 + DH + W)$   
 $O(H + C)$   
 $O((D+H)HT)$   
 ~~$O_{\text{out}}$~~

~~$O_{\text{per}} = \frac{1}{N} \sum_i \sum_j \sum_k \delta_{ijk}$~~  (I)  
 ~~$O_{\text{per}} = \frac{1}{N} \sum_i \sum_j \sum_k \delta_{ijk}$~~  (II)  
~~cross entropy loss~~  
~~accuracy~~  
~~precision~~  
~~recall~~

$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$  (I)  
 $F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$  (II)

3. (C)

$L_{\text{cls}} = \text{softmax}(Wx + b)$  (I)  
 $L_{\text{cls}} = \text{softmax}(Wx + b)$  (II)  
 $L_{\text{cls}} = \text{softmax}(Wx + b)$  (III)  
 $L_{\text{cls}} = \text{softmax}(Wx + b)$  (IV)

3172 (e)

R. H. P. Redichukas 272) תיבן פלזט מינרלים (f  
cohll

F, י' 3 85-86/.

3172 (e) מילון פלזט מינרלים (f) (g)  
TEL-AVIV University...  
Gard

תל אביב לוכן מינרלים נס, Tel Aviv  
פלזט מינרלים פלזט מינרלים Tel Aviv University

תל אביב לוכן מינרלים נס, Tel Aviv University  
פלזט מינרלים פלזט מינרלים Tel Aviv University

תל אביב לוכן מינרלים נס, Tel Aviv University  
TEL-AVIV University...  
תל אביב לוכן מינרלים נס, Tel Aviv University  
פלזט מינרלים פלזט מינרלים Tel Aviv University

תל אביב לוכן מינרלים נס, Tel Aviv University  
פלזט מינרלים פלזט מינרלים Tel Aviv University

### Question 3

(c) i – the maximum entropy value we can get is 1.0 for  $-\sum(i) y(t)[i] * \log(y(t)[i])$  where  $i$  is an index of the class (PER, LOC, etc.) and  $t$  is the time step (or a token of a given sentence). It can happen when  $p(i) = y(t)[i] * \log(y(t)[i])$  is  $1 / n_{\text{classes}}$ , and when this is the case we get the highest uncertainty level ( $y(t)[i]$  are completely randomized).

(c) ii – In the entropy graph we can see clearly that the value decreases as time step increases. It makes sense since as we train the input we wish to decrease the uncertainty (randomization) level and be able to predict well. In other words we wish to be able to distinguish clearly which element in the prediction vector is the maximum ( $\text{Max}(y(t)[i])$ ) and then, with high certainty, map it to the corresponding class.

The average loss graph (the  $J$  function in page 4) is very similar to the entropy graph (though a bit more noisy) Intuitively it makes sense since as the prediction vector  $y'(t)$  becomes more distinguished (significant), we expect  $y(i) * \log(y'(t)[i])$  to be higher ( $y(i)$  is the one hot vector), hence  $-y(i) * \log(y'(t)[i])$  to be smaller.

In the prediction-logits histogram graph we can see that in the back, the earlier epochs or time steps, the graphs are narrow and higher and as we move to the front, the graphs become wider and lower. It makes sense since these graphs capture, at the end of the day, the percentage of the classes in the training data (the bins somehow represent the different classes). Being wider and lower tells us that the proportion between the number of classes in the training data (namely, for each class, how many instances there are in the data training) are not extreme, and at any event, that these proportions are stabilized as we move along the time. As these proportions are stabilized we expect to see the higher level of certainty or lower entropy. At the end of the day, the percentage of the different

classes (for example, # ORG / total number of tokens) needs to be captured within the entropy.

(d) first we will identify the weakness of the model. We will observe few examples:

*Most of the Marines are on three ships in the Tarawa Amphibious Readiness Group*

$y^*$ :	O	O	O	MISC	O	O	O	O	O	ORG	ORG	ORG	ORG	O
$y'$ :	O	O	O	MISC	O	O	O	O	O	LOC	LOC	ORG	ORG	O
$p$ :	1.00	1.00	1.00	0.58	1.00	1.00	1.00	1.00	1.00	0.90	0.6	0.65	0.48	1.00

Here we can see that the model ( $y'$ ) incorrectly identified "*Tarawa Amphibious*" as LOC and not as ORG. In general, it is may be difficult to distinguish between organization and location especially in this case in which syntactically both are correct (prior to them we have the expression "in the" which can apply to both classes). In addition, it might be rare to see ORG of four words. In terms of predictions can see that the prediction of "Tarawa" as LOC is high and equal to 0.9 however the prediction decreases to 0.6 for the next word "Amphibious" (namely, we believe "Amphibious" is LOC with probability of 0.6 which is low).

*The expeditionary force would include nearly 1,000 Air Force personnel in ground and..*

$y^*$ :	O	O	O	O	O	O	ORG	ORG	O				
$y'$ :	O	O	O	O	O	MISC	MISC	MISC	O	O	O	O	O
$p$ :	1.00	1.00	1.00	1.00	1.00	0.80	0.93	0.66	1.00	1.00	1.00	1.00	1.00

Here the model treated the number 1,000 as a MISC (MISC are capaital letter nouns (somehow less concrete but not necessarily) which do not fall with one of the categories of LOC, ORG or PER such as Bangladeshi or even the expression "1,000 Lakes Rally").

It is clear the 1,000, in this example above, is, unlike "1,000 Lakes Rally", not part of "Air Force".

*A spokesman said he could substantially confirm a report in the news weekly Der Spiegel ,*  
 $y^*: O \quad ORG \quad ORG \quad O$   
 $y': O \quad PER \quad PER$   
 $p \quad 1.00 \quad 1.0 \quad 1.0 \quad 1.0 \quad 1.0 \quad 1.0 \quad 0.94 \quad 0.98$

Here we see the typical problem of words or, in fact expressions, being ambivalent. The model treats "Der Spiegel" as a person (PER) rather than a newspaper (ORG).

*Late on Friday Liege police said in a statement that on Thursday, Rachel Legeard, 18 , and Severine Potty, 19 , had gone shopping to the eastern town of Liege on Thursday , where Legeard 's wallet had been stolen .*

This example I believe shows the weakness and the strength of the model – the first occurrence of Liege was wrongly labeled as ORG but the second occurrence was labeled correctly as LOC. Hardly need to mention that Liege in both occurrences refers to the city Liege.