NLP - Home Work 1

Pan Eyal
208722058

January 22, 2024

1 Question 1

To get all case-insensitive tokens:

```
tr -sc 'A-Za-z' '\n' < wizard.txt | tr 'A-Z' 'a-z'
```

To get all case-insensitive types:

```
tr -sc 'A-Za-z' '\n' < wizard.txt | tr 'A-Z' 'a-z' | sort | uniq
```

To get all starting words with "ea":

```
grep -E '^ea'
```

And to count all words:

```
wc -l
```

So overall, we get for (A):

```
tr -sc 'A-Za-z' '\n' < wizard.txt | tr 'A-Z' 'a-z' | sort | uniq | grep -E '^ea' | wc -l
```

overall 8 types.

And we get for (B):

```
tr -sc 'A-Za-z' '\n' < wizard.txt | tr 'A-Z' 'a-z' | grep -E '^ea' | wc -l
```

overall 26 tokens.

Question 2

2.1

1. For P precision and R recall, which are positive numbers in $[0, 1]$:

$$\begin{aligned} F_1 &\overset{?}{\leq} \text{avg}(P, R) \\ \frac{2PR}{P+R} &\overset{?}{\leq} \frac{P+R}{2} \\ \frac{4PR}{2(P+R)} &\overset{?}{\leq} \frac{(P+R)^2}{2(P+R)} \\ 4PR &\overset{?}{\leq} (P+R)^2 \\ 4PR &\overset{?}{\leq} P^2 + 2PR + R^2 \\ 0 &\overset{?}{\leq} P^2 - 2PR + R^2 \\ 0 &\leq \underbrace{(P-R)^2}_{\substack{\text{always} \\ \text{non-negative}}} \end{aligned}$$

   Therefore, we proved that $\boldsymbol{F_1 \leq \text{avg}(P, R)}$

2. In general, for some arbitrary β: $F_\beta \leq \text{avg}(P, R)$ is no longer true. For example, when $\beta = 0$:

$$\begin{aligned} F_0 &\overset{?}{\leq} \text{avg}(P, R) \\ \frac{PR}{0 \cdot P + R} &\overset{?}{\leq} \frac{P+R}{2} \\ \frac{2PR}{2R} &\overset{?}{\leq} \frac{(P+R)R}{2R} \\ 2PR &\overset{?}{\leq} PR + R^2 \\ PR &\overset{?}{\leq} R^2 \end{aligned}$$

   We can easily see that for $P = 0.2$ and $R = 0.1$ we get:

$$\begin{aligned} 0.2 \cdot 0.1 &\overset{?}{\leq} 0.1 \\ \mathbf{0.02} &\boldsymbol{\not\leq} \mathbf{0.01} \end{aligned}$$

   Therefore, we disproved the claim for $\boldsymbol{F_\beta}$ with some arbitrary $\boldsymbol{\beta}$.

2.2

The finished transition table is:

ה	פ	י	ח	י	ב	כ	מ	#	
8	7	6	5	4	3	2	1	0	#
8	7	6	5	4	3	2	2	1	ע
7	6	5	4	3	4	3	3	2	י
7	7	6	5	4	3	4	3.5	3	ב
8	7	7	6	5	4	5	4.5	4	ו
9	8	7.5	6.5	5.5	5	6	5.5	5	ד
10	9	8	7	6.5	6	7	6.5	6	ש
9	8	9	8	7.5	7	8	7.5	7	פ
8	9	10	9	8.5	8	9	8.5	8	ה

And the transition scheme based on it is:

ה	פ	ש	ד	ו	∅	ב	י	ע	∅
ה	פ	י	ח	∅	י	ב	∅	כ	מ
= (0)	= (0)	Sub (1.5)	Sub (1.5)	Del (1)	Add (1)	= (0)	Del (1)	Sub (1)	Add (1)

So overall, **the distance between the two strings is 8**

Question 3

3.1

For data:

- $d_1 =$ " **היתה תקופה כזו שהאושר בא בזעם** ", $y_1 = 1$
- $d_2 =$ " **את עולמי עם שחר את לי כל היום** ", $y_2 = 0$
- $d_3 =$ " **עכשיו יש את הזמן במזרחית** ", $y_3 = 1$

The feature values are:

- $\boldsymbol{f_1} = \langle \mathbf{1} \quad \mathbf{1} \quad \mathbf{1} \quad \mathbf{1} \quad \mathbf{1} \rangle$
- $\boldsymbol{f_2} = \langle \mathbf{1} \quad \mathbf{0} \quad \mathbf{3} \quad \mathbf{0} \quad \mathbf{1} \rangle$
- $\boldsymbol{f_3} = \langle \mathbf{0} \quad \mathbf{1} \quad \mathbf{0} \quad \mathbf{1} \quad \mathbf{1} \rangle$

(with 5th parameter as bias)

3.2

Remembering the parameter update function $\theta^{(t+1)} = \theta^{(t)} - \eta \cdot \nabla L_{CE}(\theta^{(t)})$ the iterations are as follows:

Iteration 1:

$$\begin{aligned} \hat{y}^{(1)} &= \sigma(f_1 \cdot \theta^{(0)}) \\ &= \sigma(0+0+0+0+0) \\ &= \sigma(0) = 0.5 \end{aligned}$$

$$\eta(1) = 0.8$$

$$\begin{aligned} \nabla L_{CE}(\theta^{(0)}) &= (\hat{y}^{(1)} - y_1) \cdot f_1 \\ &= (0.5-1) \cdot \langle 1 \quad 1 \quad 1 \quad 1 \quad 1 \rangle \\ &= -0.5 \cdot \langle 1 \quad 1 \quad 1 \quad 1 \quad 1 \rangle \\ &= \langle -0.5 \quad -0.5 \quad -0.5 \quad -0.5 \quad -0.5 \rangle \end{aligned}$$

$$\begin{aligned} \boldsymbol{\theta^{(1)}} &= \langle 0 \quad 0 \quad 0 \quad 0 \quad 0 \rangle - 0.8 \cdot \langle -0.5 \quad -0.5 \quad -0.5 \quad -0.5 \quad -0.5 \rangle \\ &= \langle \mathbf{0.4} \quad \mathbf{0.4} \quad \mathbf{0.4} \quad \mathbf{0.4} \quad \mathbf{0.4} \rangle \end{aligned}$$

Iteration 2:

$$\begin{aligned}\hat{y}^{(2)} &= \sigma(f_2 \cdot \theta^{(1)}) \\ &= \sigma(\langle 1 \quad 0 \quad 3 \quad 0 \quad 1\rangle \cdot \langle 0.4 \quad 0.4 \quad 0.4 \quad 0.4 \quad 0.4\rangle \\ &= \sigma(0.4 + 0 + 1.2 + 0 + 0.4) = \sigma(2) \approx 0.88\end{aligned}$$

$$\eta(2) = 0.4$$

$$\begin{aligned}\nabla L_{CE}(\theta^{(1)}) &= (\hat{y}^{(2)} - y_2) \cdot f_2 \\ &= (0.88 - 0) \cdot \langle 1 \quad 0 \quad 3 \quad 0 \quad 1\rangle \\ &= 0.88 \cdot \langle 1 \quad 0 \quad 3 \quad 0 \quad 1\rangle \\ &= \langle 0.88 \quad 0 \quad 2.64 \quad 0 \quad 0.88\rangle\end{aligned}$$

$$\begin{aligned}\boldsymbol{\theta^{(2)}} &= \langle 0.4 \quad 0.4 \quad 0.4 \quad 0.4 \quad 0.4\rangle - 0.4 \cdot \langle 0.88 \quad 0 \quad 2.64 \quad 0 \quad 0.88\rangle \\ &= \langle 0.4 \quad 0.4 \quad 0.4 \quad 0.4 \quad 0.4\rangle - \langle 0.352 \quad 0 \quad 1.056 \quad 0 \quad 0.352\rangle \\ &= \langle \mathbf{0.048} \quad \mathbf{0.4} \quad \mathbf{-0.656} \quad \mathbf{0.4} \quad \mathbf{0.048}\rangle\end{aligned}$$

Iteration 3:

$$\begin{aligned}\hat{y}^{(3)} &= \sigma(f_3 \cdot \theta^{(2)}) \\ &= \sigma(\langle 0 \quad 1 \quad 0 \quad 1 \quad 1\rangle \cdot \langle 0.048 \quad 0.4 \quad -0.656 \quad 0.4 \quad 0.048\rangle \\ &= \sigma(0 + 0.4 + 0 + 0.4 + 0.048) = \sigma(0.848) \approx 0.7\end{aligned}$$

$$\eta(3) = 0.27$$

$$\begin{aligned}\nabla L_{CE}(\theta^{(2)}) &= (\hat{y}^{(3)} - y_3) \cdot f_3 \\ &= (0.7 - 1) \cdot \langle 0 \quad 1 \quad 0 \quad 1 \quad 1\rangle \\ &= -0.3 \cdot \langle 0 \quad 1 \quad 0 \quad 1 \quad 1\rangle \\ &= \langle 0 \quad -0.3 \quad 0 \quad -0.3 \quad -0.3\rangle\end{aligned}$$

$$\begin{aligned}\boldsymbol{\theta^{(3)}} &= \langle 0.048 \quad 0.4 \quad -0.656 \quad 0.4 \quad 0.048\rangle - 0.27 \cdot \langle 0 \quad -0.3 \quad 0 \quad -0.3 \quad -0.3\rangle \\ &= \langle 0.048 \quad 0.4 \quad -0.656 \quad 0.4 \quad 0.048\rangle - \langle 0 \quad -0.081 \quad 0 \quad -0.081 \quad -0.081\rangle \\ &= \langle \mathbf{0.048} \quad \mathbf{0.481} \quad \mathbf{-0.656} \quad \mathbf{0.481} \quad \mathbf{0.129}\rangle\end{aligned}$$

3.3

For data:

- $d_4 =$ " **אם כבר אז שיירד כאן שלג** "

The feature values are:

- $\boldsymbol{f_4} = \langle \mathbf{0} \quad \mathbf{0} \quad \mathbf{1} \quad \mathbf{1} \quad \mathbf{1} \rangle$

(with 5th parameter as bias)

And the model prediction is:

$$\begin{aligned}\hat{\boldsymbol{y}}^{(\mathbf{4})} &= \sigma(f_4 \cdot \theta^{(3)}) \\ &= \sigma(\langle 0 \quad 0 \quad 1 \quad 1 \quad 1 \rangle \cdot \langle 0.048 \quad 0.481 \quad -0.656 \quad 0.481 \quad 0.129 \rangle \\ &= \sigma(0 + 0 - 0.656 + 0.481 + 0.129) = \sigma(-0.046) \approx \mathbf{0.049}\end{aligned}$$

Therefore, the model will predict that the data $\boldsymbol{d_4}$ is not a sentence from a Shlomo Artzi song.

Question 4

Feedback: In general, I already enjoy the class lectures and material as it is delivered. For me, I find it a bit confusing to have multiple definitions names, both in English and in Hebrew. I would prefer that we stick to a single defining name during the lectures (and preferable in English).