**NLP - HOMEWORK 1**

I.D: 208722058

**Q1)**

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) types:**

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

* **8**

**And we get for (B) tokens:**

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

* **26**

**Q2.1**.1**)**

For precision and recall, which are positive numbers in :

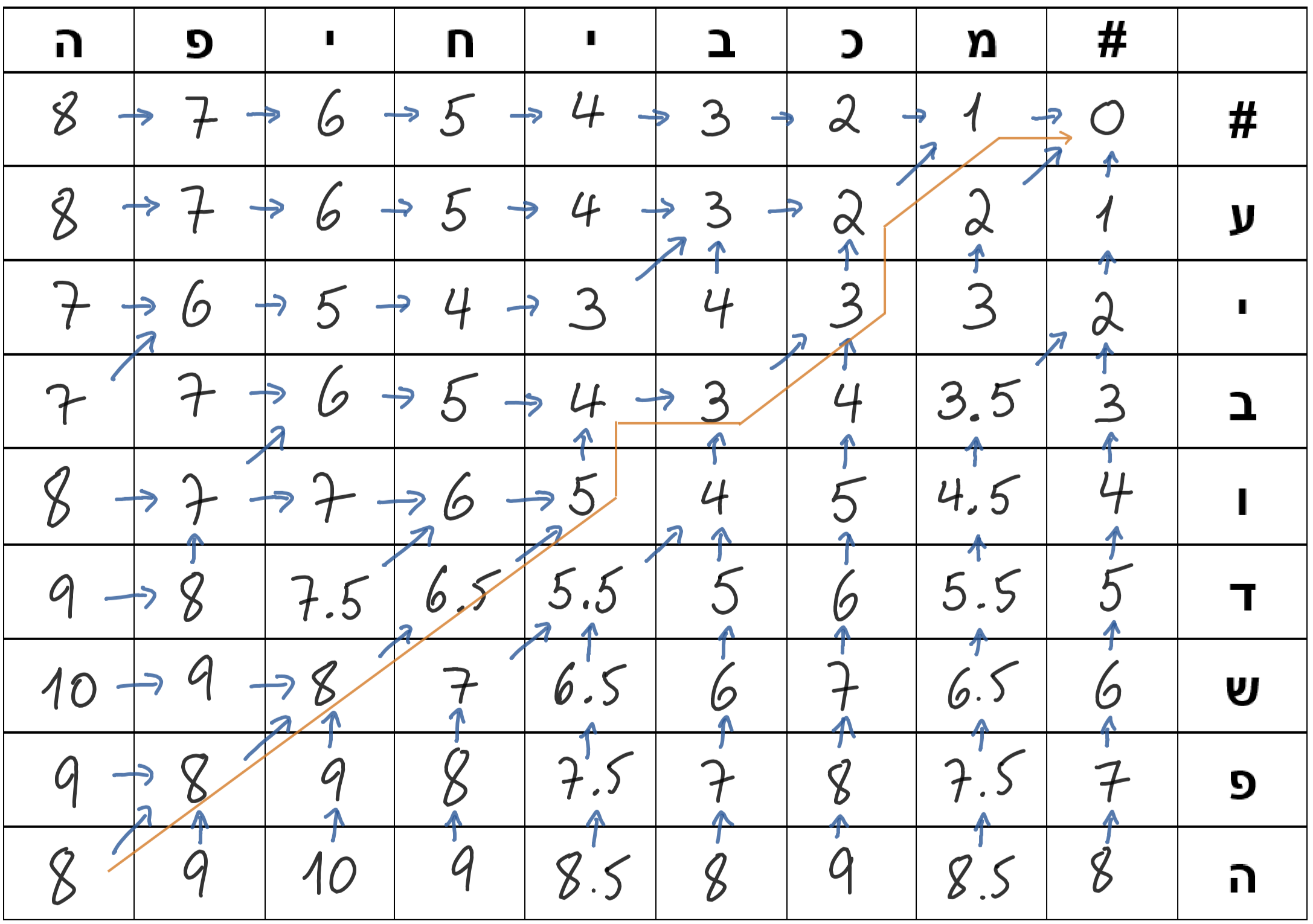
**Q2.1**.2**)**

In general, for some arbitrary is no longer true.

For example, when :

And for and we get:

**Q2.2)**

****

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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)** |

Overall:

**Q3**.1**)**

For data:

* "היתה תקופה כזו שהאושר בא זעם" ,
* "את עולמי עם שחר, את לי כל היום" ,
* "עכשיו יש את הזמן במזרחית" ,

**The features values are:**

(With 5th parameter as bias)

**Q3**.2**)**

The parameter update function:

Iteration 1:

Iteration 2:

Iteration 3:

**Q3**.3**)**

For data:

* " אם כבר אז שיירד כאן שלג"

The feature values are:

(With 5th parameter as bias)

And now for the model:

**Therefore, the model will predict that the data is a sentence from a Shlomo Artzi song.**

**Q4)**

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 and I would prefer that we would stick to a single defining name during the lectures (and preferable in English).