Neural Network in NLP

Task1

S_1: The man saw a car in the park

S_2: ~~I saw the man park in the car~~

I saw the man park the car.

Procesare:

S_1 tokens = {the, man, saw, a, car, in, the, park}

S_2 tokens = {I, saw, the, man, park, the, car}

$V = S_1$ tokens $\cup$ S_2 tokens = {the, man, saw, a, car, in, park, I}

Reprez. vectorială (frecvență)

cuvânt	S_1	S_2
the	2	2
man	1	1
saw	1	1
a	1	0
car	1	1
in	1	0
park	1	1
I	0	1

$S_1 = [2, 1, 1, 1, 1, 1, 1, 0]$

$S_2 = [2, 1, 1, 0, 1, 0, 1, 1]$

a). Distanța euclidiană: $d = \sqrt{\sum_i (S_{1_i} - S_{2i})^2}$

$$S_1 - S_2 = [2-2, 1-1, 1-1, 1-0, 1-1, 1-0, 1-1, 0-1]$$

$$= [0, 0, 0, 1, 0, 1, 0, -1]$$

$$\Rightarrow d = \sqrt{1^2 + 1^2 + (-1)^2} = \sqrt{3} \simeq 0,366$$

b). Vector cosinus: $\cos = \frac{S_1 \cdot S_2}{\|S_1\| \cdot \|S_2\|}$

$$S_1 \cdot S_2 = 2 \cdot 2 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 0 + 1 \cdot 1 + 1 \cdot 0 + 1 \cdot 1 + 0 \cdot 1 = 4 + 4 = 8$$

$$\|S_1\| = \sqrt{2^2 + 1^2 \cdot 6} = \sqrt{4+6} = \sqrt{10}$$

$$\|S_2\| = \sqrt{2^2 + 1^2 \cdot 5} = \sqrt{4+5} = \sqrt{9} = 3$$

$$\Rightarrow \cos = \frac{8}{3\sqrt{10}} \simeq 0,843$$

c). Jaccard

$$J(S_1,S_2)=\frac{|S_1\cap S_2|}{|S_1\cup S_2|}$$

$$|S_1\cap S_2| = |\{\text{the, man, saw, car, park}\}| = 5$$

$$|S_1\cup S_2| = 8$$

$$\Rightarrow J = \frac{5}{8} \simeq 0,625$$

d). Overlap

$$\left.\text{Overlap} = \frac{|S_1\cap S_2|}{\min(|S_1|,|S_2|)}\right\} \Rightarrow \text{overlap} = \frac{5}{6} \simeq 0,833$$

$$|S_1| = 7$$
$$|S_2| = 6 \quad \Rightarrow \min(|S_1|,|S_2|) = 6$$