23. NLP and building your own Embeddings

9 януари 2024

За какво си говорихме предния път

- Основни разлики между Data Science, AI и ML
- Разни полезни инструменти (но не заместители)
- Доста за NumPy
- Доста и за Pandas
- Започнахме да разглеждаме данни

Малко въпроси

```
a = np.array([0,1,2,3,4,5,6,7,8,9])
b = a.reshape(3,3)
b
```

```
ValueError: cannot reshape array of size 10 into shape (3,3)
```

Малко въпроси

```
a = np.arange(20).reshape(4,5)
a[1:2, 3]
```

```
array([[ 0,  1,  2,  3,  4],
       [ 5,  6,  7,  8,  9],
       [10, 11, 12, 13, 14],
       [15, 16, 17, 18, 19]])
```

Малко въпроси

```
arr1 = [0, 1, 2, 3, 4, 5]
arr2 = [6, 7, 8, 9, 10, 11, 12]

s1 = pd.Series(arr1)
s2 = pd.Series(arr2)
s2.add(s1)
```

```
0     6.0
1     8.0
2    10.0
3    12.0
4    14.0
5    16.0
6     NaN
dtype: float64
```

Малко за НЛП-то

- Клон в АИ, който се фокусира в това да разреши на машините да отговарят на “човешкия” език по ценен начин.
- Защо това е важно?
- Примери: Преводачи, Виртуални асистенти, Сърч Енджини, Сентимент Анализ, Филтриране на имейли, Чатботове и тн.

Embeddings

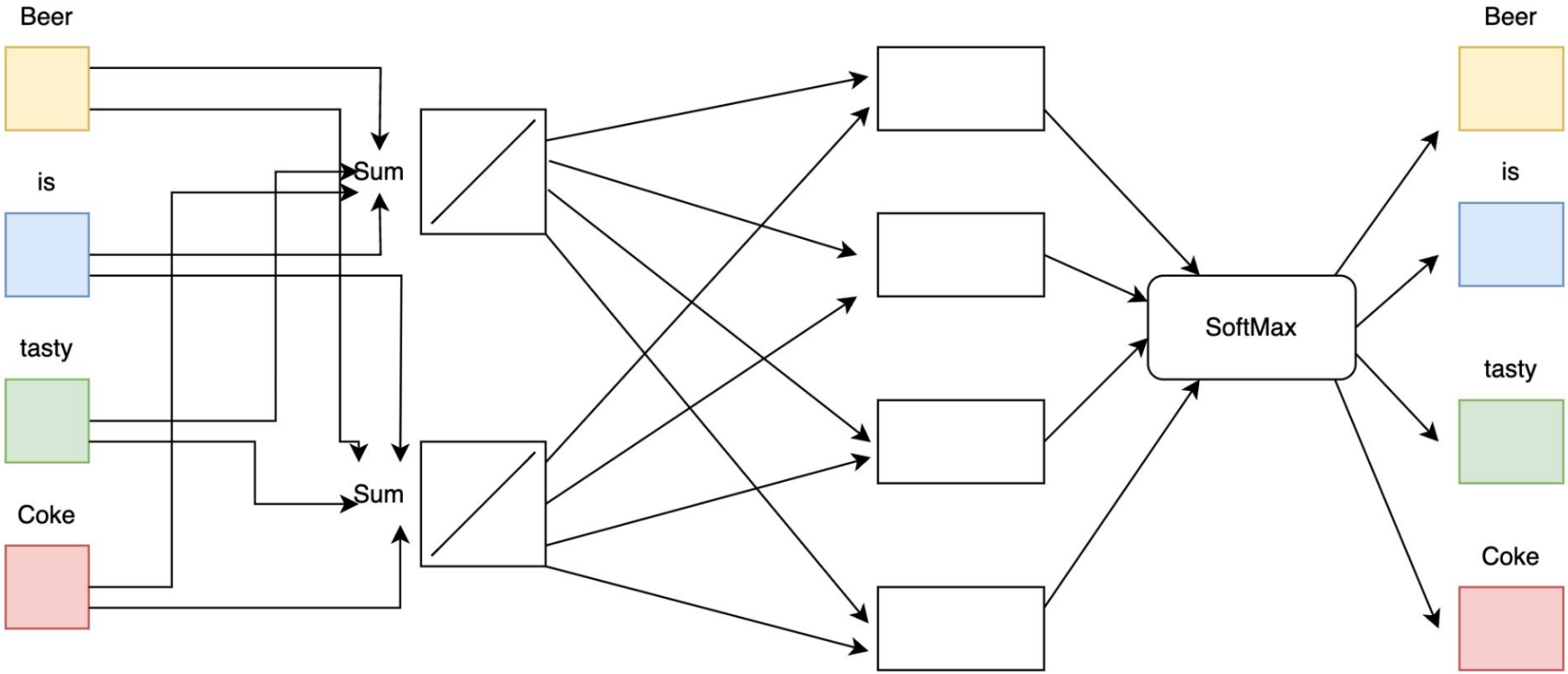
- Числено представяне на думи, където всяка дума е репрезентирана от вектор в непрекъснато векторно пространство. Обикновено се състоят от десетки или стотици измерения.
- Улавянето на семантична връзка между думите.
- Улавяне на контекстуална информация в изречения.
- Позволяват ни да намалим нужните пространства, за да работим с думи.
- Използват Невронни мрежи.

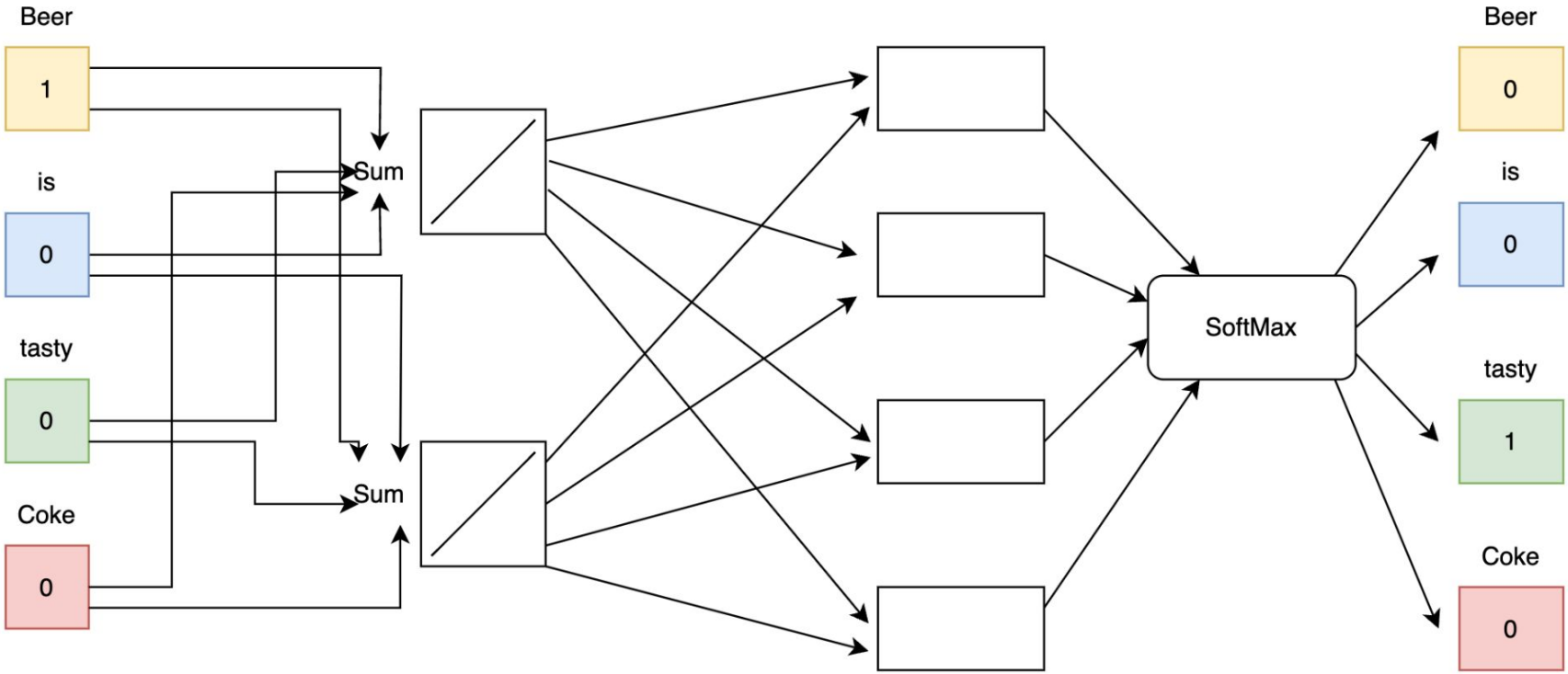
Word2Vec

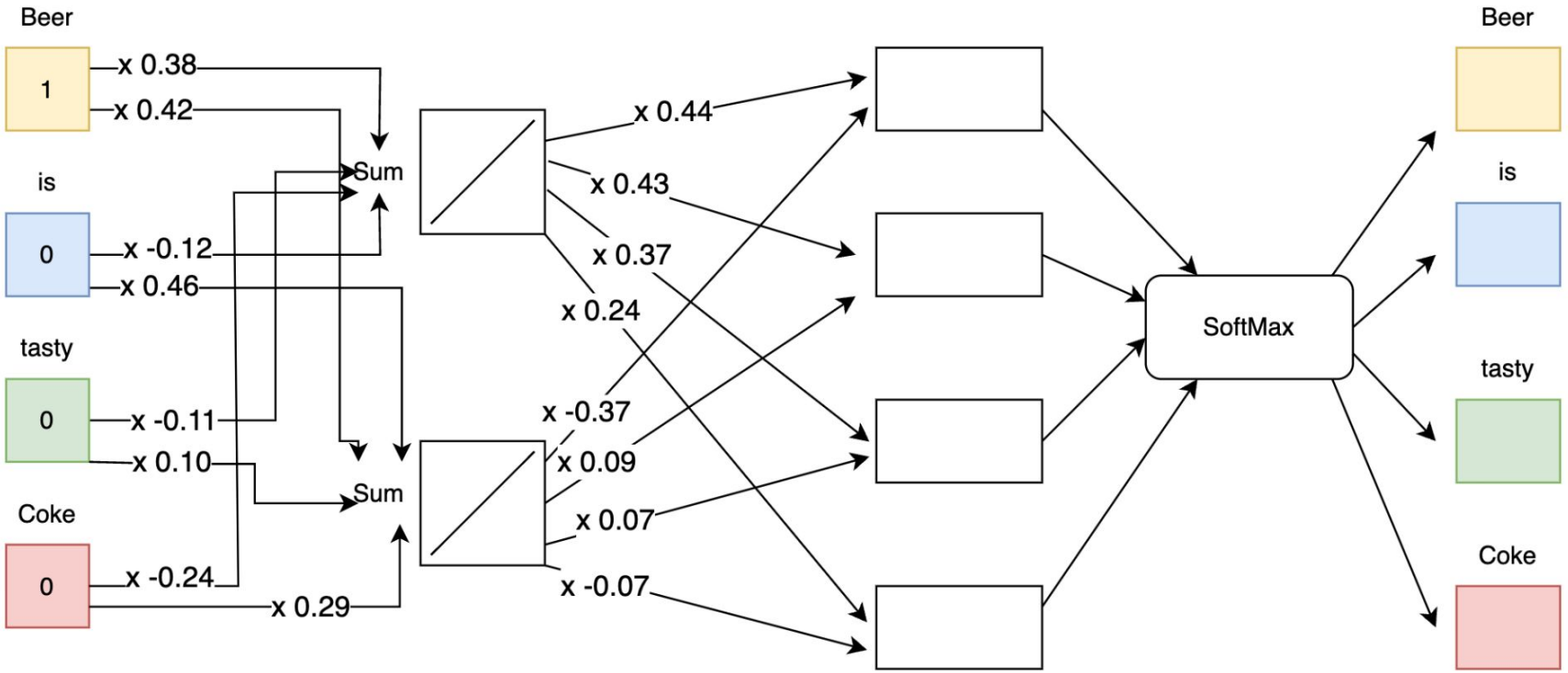
- Появява се през 2013- Естествено от Гугъл
- Continuous bag of words:
  - Дадено ни е изречението: Пайтън е много ___ език.
  - Искаме да попълним празното място с подходяща дума, например- “як” или “готин” (може и “труден”).
- Skip-Gram:
  - Като CBOW, ама наобратно. Дадена ни е думата “бира” .
  - Искаме да предскажем възможните думи наоколо - “крафт”, “лагер”, “пия”.

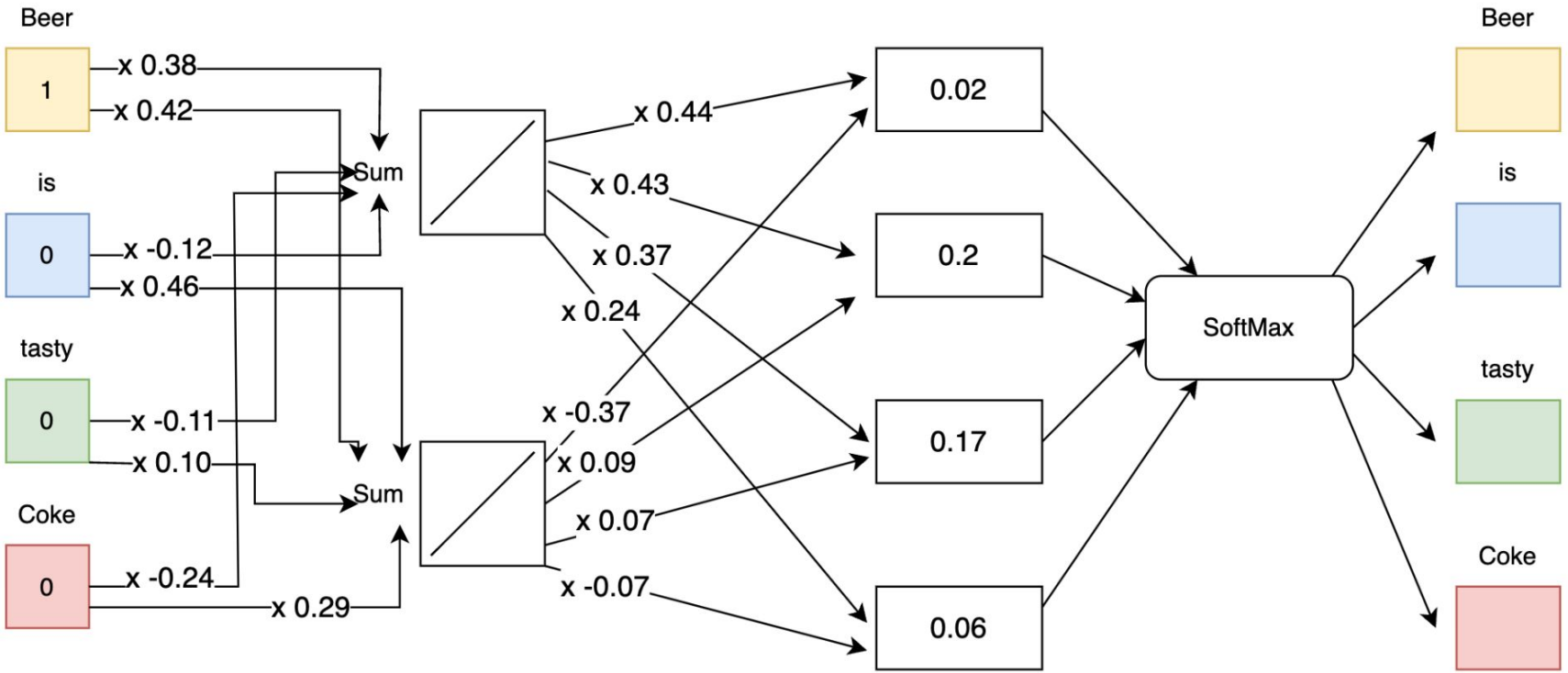
Нека направим един много прост Embedding

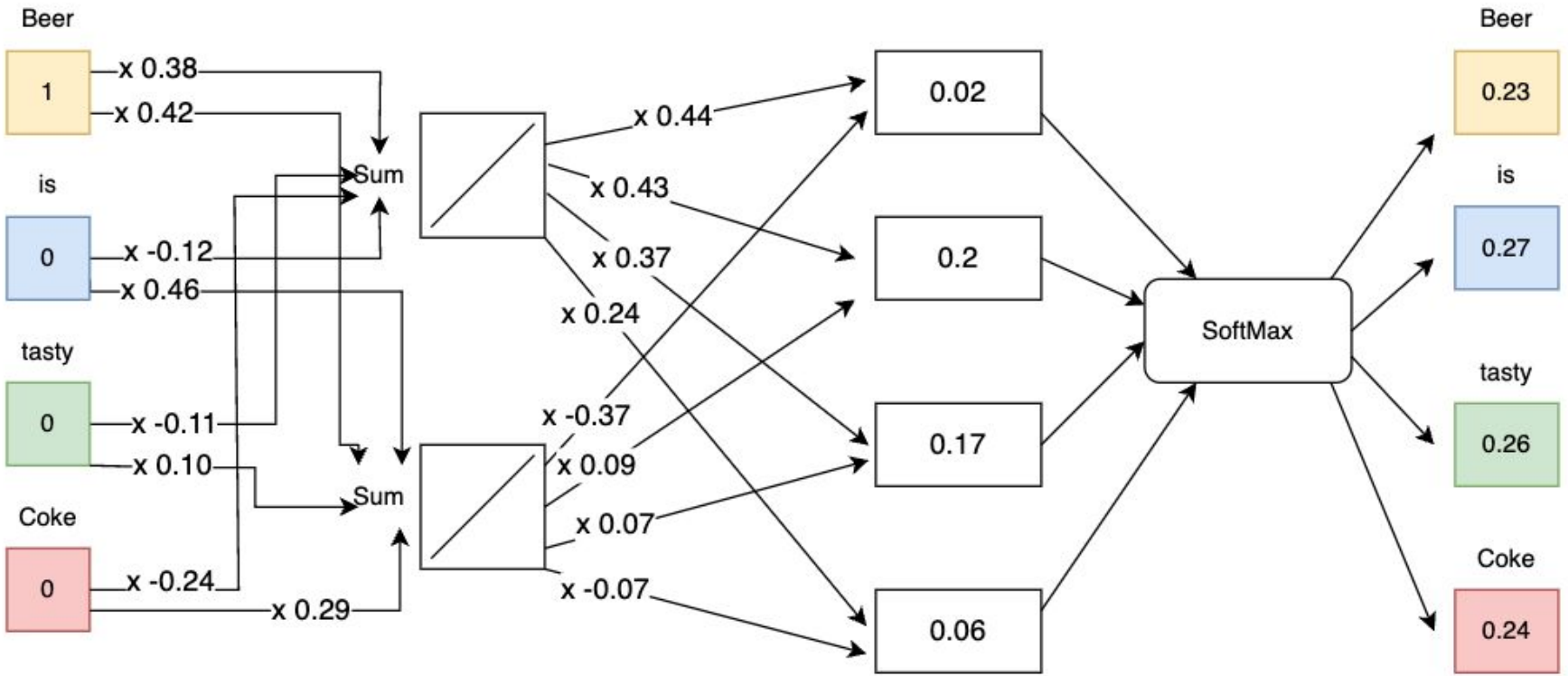
- Взимаме следните 2 изречения:
  - Beer is tasty
  - Coke is tasty
- Tokenization (Много важна стъпка на обработване):
  - `["Beer", "is", "tasty"]`
  - `["Coke", "is", "tasty"]`
- Нека сега направим един наш Embeddings Модел
- Извинявам се за следващите слайдове

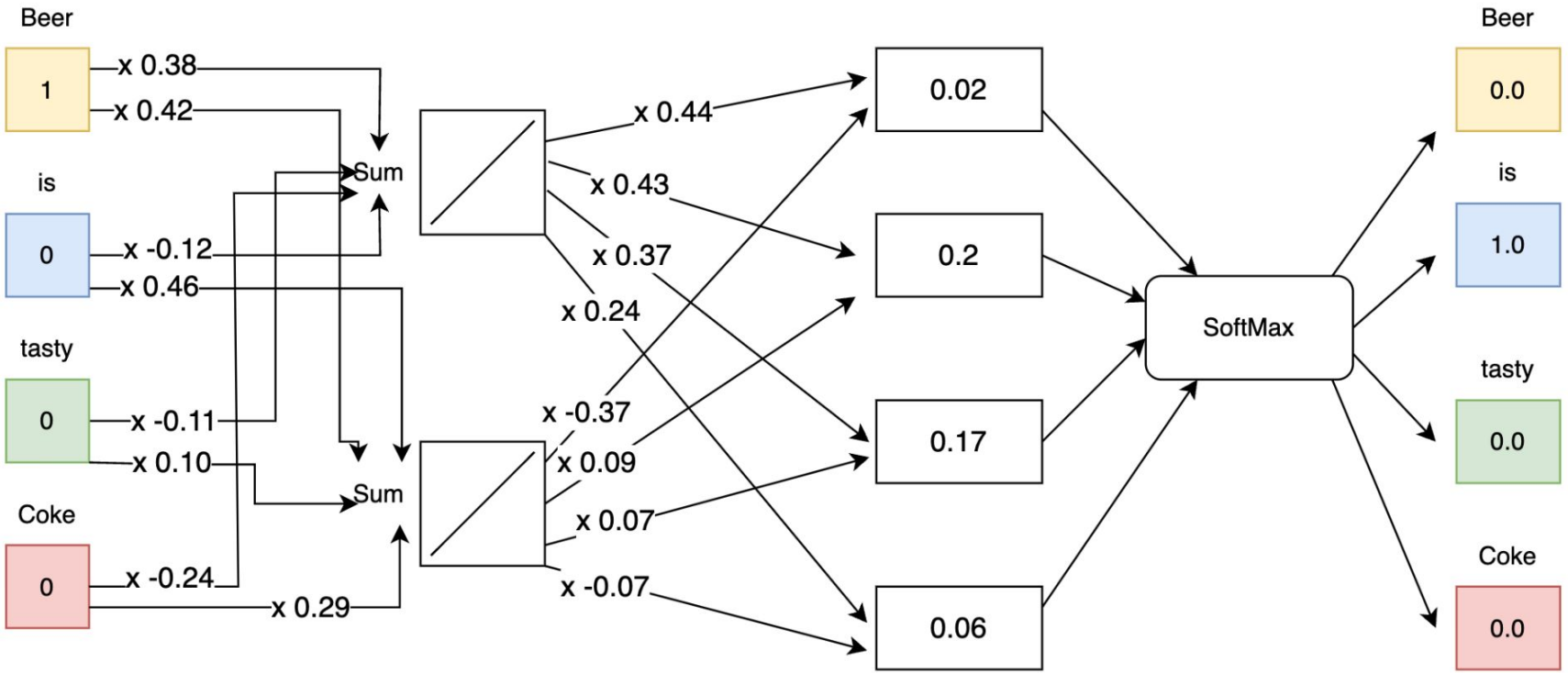

Beer
is
tasty
Coke
Sum
Sum
SoftMax
Beer
is
tasty
Coke


Beer
1
is
0
tasty
0
Coke
0
Sum
Sum
SoftMax
Beer
0
is
0
tasty
1
Coke
0


Beer
1
x 0.38
x 0.42
is
0
x -0.12
x 0.46
tasty
0
x -0.11
x 0.10
Coke
0
x -0.24
x 0.29
Sum
Sum
x 0.44
x 0.43
x 0.37
x 0.24
x -0.37
x 0.09
x 0.07
x -0.07
SoftMax
Beer
is
tasty
Coke


Beer
1
x 0.38
x 0.42
is
0
x -0.12
x 0.46
tasty
0
x -0.11
x 0.10
Coke
0
x -0.24
x 0.29
Sum
Sum
x 0.44
x 0.43
x 0.37
x 0.24
x -0.37
x 0.09
x 0.07
x -0.07
0.02
0.2
0.17
0.06
SoftMax
Beer
is
tasty
Coke


Beer
1
x 0.38
x 0.42
is
0
x -0.12
x 0.46
tasty
0
x -0.11
x 0.10
Coke
0
x -0.24
x 0.29
Sum
Sum
x 0.44
x 0.43
x 0.37
x 0.24
x -0.37
x 0.09
x 0.07
x -0.07
0.02
0.2
0.17
0.06
SoftMax
Beer
0.23
is
0.27
tasty
0.26
Coke
0.24


Beer
1
x 0.38
x 0.42
is
0
x -0.12
x 0.46
tasty
0
x -0.11
x 0.10
Coke
0
x -0.24
x 0.29
Sum
Sum
x 0.44
x 0.43
x 0.37
x 0.24
x -0.37
x 0.09
x 0.07
x -0.07
0.02
0.2
0.17
0.06
SoftMax
Beer
0.0
is
1.0
tasty
0.0
Coke
0.0


pmitf.com

Въпроси?