**Information Retrieval HW03**

Team: **Crawling in My Skin**

Team members:

* Dmytri Kislitsyn 323349738
* Liel Pasman 323068502
* Maxim Lebedinsky 318171485

[GitHub Link](https://github.com/MaximL98/CrawlingInMyProtein.github.io)

[Colab Link 1](https://colab.research.google.com/drive/1yqArFeSeDVwN0JxcNIhm6J95wcrL9RHj?usp=sharing) | [Colab Link 2](https://colab.research.google.com/drive/1YXGn_4YNiHQ5gXQKtLcbIrtVUHLMlCmK?usp=sharing) | [Colab Link 3](https://colab.research.google.com/drive/1954WQ2ew7y6GjxjcivEDj3_MWctbaYxJ?usp=sharing)

**Answer to question 1:**

We will use a simple one-dimensional example with two sets of learning examples - one positive and one negative to show that the Rocchio classifier can associate a wrong tag with a learning example.

Positive group: (1,2,3)

Negative group (Negative): (6,7,8)

Calculate the center of each group, which is the average of the values ​​in the group.

- Center of the positive group:

=**2**

- Center of the negative group:

**7**=

Let's say we want to classify the new sample X=5

Calculate the distance of X=5 from each of the centers of the groups.

- The distance from the center of the positive group:

|5-2|=**3**

- The distance from the center of the negative group:

|5-7|=**2**

In this case the distance of x=5 from the center of the negative group is smaller and therefore the classifier will identify the sample as belonging to the negative group.

But if we assume that the sample belongs to the positive group because there are no more details on it, it means that the classifier classified it incorrectly.

**Answer to question 2:**

**Section A**

A black line with a circle and a arrow

Description automatically generated

**A:**

In-Degree score: 0 (0 incoming edges), authority score: 0.

Out-Degree score: 1 (1 outgoing edges), hub score: 1.

**B:**

In-Degree score: 1 (1 incoming edges), authority score: 1.

Out-Degree score: 1 (1 outgoing edges), hub score: 1.

**C:**

In-Degree score: 1 (1 incoming edges), authority score: 1.

Out-Degree score: 1 (1 outgoing edges), hub score: 1.

**D:**

In-Degree score: 1 (1 incoming edges), authority score: 1.

Out-Degree score: 0 (0 outgoing edges), hub score: 0.

**Section B**

**Initial stage:**

|  |  |  |
| --- | --- | --- |
| Page | Hub\_0 (Out\_deg) | Auth\_0 (In\_deg) |
| A | 1 | 1 |
| B | 1 | 1 |
| C | 1 | 1 |

**Iteration 1:**

Authority update:

a(A) = h(A) + h(C) = 2

a(B) = h(A) + h(C) = 2

a(C) = h(A) + h(B) = 2

Normalize: Sum = 6

a(A) = 2/6, a(B) = 2/6, a(C) = 2/6

Hub Update:

h(A) = a(A) + a(B) + a(C) = 1

h(B) = a(C) = 1/3

h(C) = a(A) + a(B) = 2/3

Normalize: Sum = 2

h(A) = 0.5, h(B) = 0.166, h(C) = 0.333

|  |  |  |
| --- | --- | --- |
| Page | Hub\_1 (Out\_deg) | Auth\_1 (In\_deg) |
| A | 0.5 | 0.333 |
| B | 0.166 | 0.333 |
| C | 0.333 | 0.333 |

**Iteration 2:**

Authority update:

a(A) = h(A) + h(C) = 0.833

a(B) = h(A) + h(C) = 0.833

a(C) = h(A) + h(B) = 0.666

Normalize: Sum = 2.332

a(A) = 0.357, a(B) = 0.357, a(C) = 0.285

Hub Update:

h(A) = a(A) + a(B) + a(C) = 0.999

h(B) = a(C) = 0.285

h(C) = a(A) + a(B) = 0.714

Normalize: Sum = 1.998

h(A) = 0.5, h(B) = 0.142, h(C) = 0.357

|  |  |  |
| --- | --- | --- |
| Page | Hub\_2 (Out\_deg) | Auth\_2 (In\_deg) |
| A | 0.5 | 0.357 |
| B | 0.142 | 0.357 |
| C | 0.357 | 0.285 |

**Iteration 3:**

Authority update:

a(A) = h(A) + h(C) = 0.857

a(B) = h(A) + h(C) = 0.857

a(C) = h(A) + h(B) = 0.642

Normalize: Sum = 2.356

a(A) = 0.363, a(B) = 0.363, a(C) = 0.272

Hub Update:

h(A) = a(A) + a(B) + a(C) = 0.998

h(B) = a(C) = 0.272

h(C) = a(A) + a(B) = 0.726

Normalize: Sum = 1.996

h(A) = 0.5, h(B) = 0.136, h(C) = 0.363

|  |  |  |
| --- | --- | --- |
| Page | Hub\_3 (Out\_deg) | Auth\_3 (In\_deg) |
| A | 0.5 | 0.363 |
| B | 0.136 | 0.363 |
| C | 0.363 | 0.272 |

Note that after performing the third iteration, the scores converge and barely change between the 2nd and 3rd iterations. Also, most changes vary because the rounding of the scores. We could perform a 4th iteration to get more accurate scores, but if we only want a rough estimate of the hub and auth scores, 3 iterations are enough in this case.