Lucas 5/2/2014

Results for first 43 news when difference < 15

(first two columns are the id of news and the third column is their difference value computed. Result is a list of pairs considered to be similar/duplicate. Highlighted ones are those that are not actually duplicates but considered as duplicates.)

Duplicate/Similar news:

7 4 11
4 7 11
42 41 8
14 2 11
23 22 13
35 38 14
40 41 7
41 42 8
17 42 8
42 17 8
2 14 11
14 13 14
26 27 8
40 42 11
40 17 9
17 41 12
41 27 13
22 23 13
38 34 13
33 9 12
26 24 14
2 1 13
2 19 10
40 18 14
7 19 14
41 40 7
19 2 10
41 17 12
17 40 9
41 26 13
9 33 12
1 2 13
4 9 14
27 41 13
32 2 11
13 32 12
38 35 14
35 34 11
18 17 11
9 4 14
3 6 10
26 41 13
34 35 11

2 32 11
17 18 11
18 42 11
24 26 14
14 32 12
32 13 12
42 18 11
27 26 8
13 14 14
19 7 14
34 38 13
42 40 11
32 14 12
20 2 12
18 40 14
2 20 12
6 3 10

Total # of duplicates exist: 36

Changing the threshold of difference gives:

Duplicate/Similar news:

diff<15 : precision 1- 24/60 ~ 60% Recall: (60-24)/36 = 100%

diff<14: precision: 1- 20/48 ~ 58.3% Recall: (48-20)/36= 77.78%

diff<13 precision: 1- 12/38 ~ 68.4% Recall: (38-12)/36 = 72.2%

diff<12 precision: 1- 10/32 ~ 68.8% Recall: (32-10)/36 = 61.1%

diff<11 precision: 1- 2/18 ~ 88.9% Recall: (18-2)/36 = 44.4%

diff<10 precision: 1- 0/14 ~ 100% Recall: (14-0)/36 = 38.9%