Naïve Bayes

Prelucrarea documentelor

Documentele (fisier/linie) au fost prelucrate pentru a se obtine doar grupurile formate din caractere alfabetice sau cifre (a-zA-Z0-9)

Am retinut cuvintele dintr-un document sub forma de **bag of words.**

Test 1

- Training set 70%
- Test set 30%

SpamAssasim

	Ham	Spam	Total
Ham	**2057**	10	**2067**
Spam	48	**670**	**718**

	Ham	Spam
Ham	**99.51 %**	0.49 %
Spam	6.69 %	**93.31 %**

Rotten Tomatoes

	Class 0	Class 1	Class 2	Class 3	Class 4	Total
Class 0	**1166**	783	113	50	32	**2144**
Class 1	2059	**4016**	1227	576	264	**8142**
Class 2	3099	5151	**8202**	4662	2737	**23851**
Class 3	405	831	1367	**4430**	2598	**9639**
Class 4	40	83	126	860	**1621**	**2730**

	Class 0	Class 1	Class 2	Class 3	Class 4
Class 0	**54.38 %**	36.52 %	5.27 %	2.33 %	1.49 %
Class 1	25.29 %	**49.32 %**	15.07 %	7.07 %	3.24 %
Class 2	12.99 %	21.6 %	**34.39 %**	19.55 %	11.48 %
Class 3	4.2 %	8.62 %	14.18 %	**45.96 %**	26.95 %
Class 4	1.47 %	3.04 %	4.62 %	31.5 %	**59.38 %**

Test 2

- Training set 50%
- Test set 50%

SpamAssasim

	Ham	Spam	Total
Ham	**3491**	26	**3445**
Spam	93	**1089**	**1182**

	Ham	Spam
Ham	**99.24 %**	0.49 %
Spam	7.87 %	**92.13 %**

Rotten Tomatoes

	Class 0	Class 1	Class 2	Class 3	Class 4	Total
Class 0	***1767***	1389	240	81	54	**3531**
Class 1	*3231*	**6709**	2227	1000	487	**13654**
Class 2	*5125*	8475	**13873**	7992	4483	**39948**
Class 3	*708*	1410	2474	**7699**	4142	**16433**
Class 4	*77*	145	241	1623	**2548**	**4634**

	Class 0	Class 1	Class 2	Class 3	Class 4
Class 0	**50.04 %**	39.34 %	6.8 %	2.29 %	1.53 %
Class 1	23.66 %	**49.14 %**	16.31 %	7.32 %	3.57 %
Class 2	12.83 %	21.22 %	**34.73 %**	20.01 %	11.22 %
Class 3	4.31 %	8.58 %	15.06 %	**46.85 %**	25.21 %
Class 4	1.66 %	3.13 %	5.2 %	35.02 %	**54.98 %**

Test 3

- Training set 30%
- Test set 70%

SpamAssasim

	Ham	Spam	Total
Ham	**4832**	33	**4865**
Spam	195	**1491**	**1686**

	Ham	Spam
Ham	**99.32 %**	0.68 %
Spam	11.57 %	**88.43 %**

Rotten Tomatoes

	Class 0	Class 1	Class 2	Class 3	Class 4	Total
Class 0	***1887***	2257	535	189	115	**4983**
Class 1	*3683*	**9217**	3836	1648	678	**19062**
Class 2	*7620*	11614	**19815**	11358	5526	**55933**
Class 3	*1027*	2003	4050	**11281**	4649	**23010**
Class 4	*87*	226	405	2761	**2905**	**6384**

	Class 0	Class 1	Class 2	Class 3	Class 4
Class 0	**37.87 %**	45.29 %	10.74 %	3.79 %	2.31 %
Class 1	19.32 %	**48.35 %**	20.12 %	8.65 %	3.56 %
Class 2	13.62 %	20.76 %	**35.43 %**	20.31 %	9.88 %
Class 3	4.46 %	8.7 %	17.6 %	**49.03 %**	20.2 %
Class 4	1.36 %	3.54 %	6.34 %	43.25 %	**45.5 %**

Test 4

- Training set 10%
- Test set 90%

SpamAssasim

	Ham	Spam	Total
Ham	**6243**	34	**6277**
Spam	857	**1318**	**2175**

	Ham	Spam
Ham	**99.46 %**	0.54 %
Spam	39.4 %	**60.6 %**

Rotten Tomatoes

	Class 0	Class 1	Class 2	Class 3	Class 4	Total
Class 0	***1182***	2943	1661	457	108	**6351**
Class 1	*2760*	**10569**	7874	2624	719	**24546**
Class 2	*8226*	13711	**30216**	14868	4601	**71622**
Class 3	*1132*	2694	8045	**14460**	3294	**29625**
Class 4	*89*	347	1343	4509	**1966**	**8254**

	Class 0	Class 1	Class 2	Class 3	Class 4
Class 0	**18.61 %**	46.34 %	26.15 %	7.2 %	1.7 %
Class 1	11.24 %	**43.06 %**	32.08 %	10.69 %	2.93 %
Class 2	11.49 %	19.14 %	**42.19 %**	20.76 %	6.42 %
Class 3	3.82 %	9.09 %	27.16 %	**48.81 %**	11.12 %
Class 4	1.08 %	4.2 %	16.27 %	54.63 %	**23.82 %**

Rotten Tomatoes

Prelucrarea documentelor

Documentele (fisier/linie) au fost prelucrate pentru a se obtine doar grupurile formate din caractere alfabetice sau cifre (a-zA-Z0-9)

Am retinut cuvintele dintr-un document sub forma de **bag of words.**

Distanta dintre 2 documente este reprezentata de numarul de cuvinte ce se regasesc in ambele documente (dictionare). Am incercat mai multe variante folsind dictionare dar rezultatele obtinute erau foarte slabe.

Alternative: Diferenta absoluta dintre aparitiile aceluias cuvand in ambele dictionare

Diferenta absoluta / numarul de cuvinte ce se repeta

Test 1

- Training set 99%
- Test set 1%

SpamAssasim

```
==============================
CLS 0:    [62, 0]
CLS 1:    [3, 14]
==============================
CLS 0:    [58, 4]
CLS 1:    [2, 15]
==============================
CLS 0:    [60, 2]
CLS 1:    [3, 14]
==============================
CLS 0:    [60, 2]
CLS 1:    [1, 16]
==============================
CLS 0:    [61, 1]
CLS 1:    [1, 16]
==============================
CLS 0:    [60, 2]
CLS 1:    [1, 16]
==============================
```

CLS 0: [62, 0]
CLS 1: [1, 16]
==========
CLS 0: [62, 0]
CLS 1: [1, 16]
==========
CLS 0: [62, 0]
CLS 1: [1, 16]

Rotten Tomatoes

K = 1

	Class 0	Class 1	Class 2	Class 3	Class 4	**Total**
Class 0	***15***	13	11	14	22	**75**
Class 1	*22*	**66**	56	47	22	**213**
Class 2	*26*	112	**167**	155	296	**756**
Class 3	*5*	23	65	**92**	156	**341**
Class 4	*0*	4	9	29	**52**	**94**

	Class 0	Class 1	Class 2	Class 3	Class 4
Class 0	**20 %**	17.33 %	14.67 %	18.67 %	29.33 %
Class 1	10.33 %	**30.99 %**	26.29 %	22.07 %	10.33 %
Class 2	3.44 %	14.81 %	**22.09 %**	20.5 %	39.15 %
Class 3	1.47 %	6.74 %	19.06 %	**26.98 %**	45.75 %
Class 4	0 %	4.26 %	9.57 %	30.85 %	**55.32 %**

CLS 0: [12, 11, 11, 16, 25]
CLS 1: [17, 48, 67, 53, 102]
CLS 2: [12, 85, 180, 158, 321]
CLS 3: [0, 21, 50, 95, 175]
CLS 4: [0, 1, 8, 24, 61]
==========
CLS 0: [21, 10, 9, 15, 20]
CLS 1: [21, 62, 65, 47, 92]
CLS 2: [25, 106, 188, 155, 282]
CLS 3: [3, 24, 56, 104, 154]
CLS 4: [3, 2, 8, 24, 57]

==========
CLS 0: [22, 13, 7, 14, 19]
CLS 1: [23, 66, 69, 42, 87]
CLS 2: [26, 105, 203, 157, 265]
CLS 3: [2, 24, 57, 105, 153]
CLS 4: [4, 1, 3, 22, 64]
==========
CLS 0: [20, 17, 6, 16, 16]
CLS 1: [23, 70, 67, 44, 83]
CLS 2: [18, 104, 242, 159, 233]
CLS 3: [1, 22, 60, 120, 138]
CLS 4: [3, 1, 5, 26, 59]
==========
CLS 0: [21, 16, 5, 17, 16]
CLS 1: [23, 82, 70, 40, 72]
CLS 2: [16, 108, 246, 156, 230]
CLS 3: [3, 18, 64, 130, 126]
CLS 4: [3, 1, 5, 22, 63]
==========
CLS 0: [23, 19, 4, 17, 12]
CLS 1: [22, 98, 66, 37, 64]
CLS 2: [15, 115, 261, 162, 203]
CLS 3: [3, 20, 69, 136, 113]
CLS 4: [3, 1, 5, 29, 56]
==========
CLS 0: [21, 22, 4, 16, 12]
CLS 1: [18, 109, 62, 42, 56]
CLS 2: [11, 99, 270, 173, 203]
CLS 3: [2, 17, 64, 139, 119]
CLS 4: [2, 0, 5, 26, 61]
==========
CLS 0: [21, 21, 5, 15, 13]
CLS 1: [16, 112, 62, 42, 55]
CLS 2: [11, 94, 269, 179, 203]
CLS 3: [1, 12, 57, 149, 122]
CLS 4: [1, 0, 4, 27, 62]

Test 2

Training set:

- CLS: 0 -> 6378
- CLS: 1 -> 24526
- CLS: 2 -> 71491
- CLS: 3 -> 29700
- CLS: 4 -> 8276

Test set:

- CLS: 0 -> 694
- CLS: 1 -> 2747
- CLS: 2 -> 8091
- CLS: 3 -> 3227
- CLS: 4 -> 930