Балашов Александр ВМК 210 группа
Теоретическая часть

1.Наивный Байесовский классификатор, модель Бернулли

a) P($v_i \in d | d \in c_i$) = $\frac{n}{m}$, n – количество документов класса c_i, в которых содержится слово d, m – количество документов класса c_i.

b) P($d = (k_1, k_2, k_3, \dots, k_M) | d \in c_j$) = $\prod_i P(v_i \in d | d \in c_i)^{k_i} (1 - P(v_i \in d | d \notin c_i))^{k_i}$

c) P($c_j | d$) = $\frac{P(d|c_j) * P(c_j)}{P(d)}$

d) c = $argmax_j(P(d|c_j) * P(c_j))$

2.Наивный Байесовский классификатор, мультиномиальная модель

a) P($v_i \in d | d \in c_i$) = $\frac{n}{m}$, n –количество слов v_i во всех документах класса c_i, m – количество всех слов во всех документов класса c_i.

b) P($d = (k_1, k_2, k_3, \dots, k_M) | d \in c_j$) = $\prod_i P(v_i \in d | d \in c_i)^{k_i}$

c) P($c_j | d$) = $\frac{P(d|c_j) * P(c_j)}{P(d)}$

d) c = $argmax_j(P(d|c_j) * P(c_j))$

Практическая часть

a) Для позитивных отзывов: минимальная длина = 70, максимальная длина = 10363, средняя длина = 1360.8, медиана = 996. Для негативных отзывов: минимальная длина = 52, максимальная длина = 8969, средняя длина = 1316.3б медиана = 981.

d)

30 самых частотных слов (позитивные отзывы)	Их частоты	30 самых частотных слов (негативные отзывы)	Их частоты
The	0.056667997835015034	The	0.05476501182869007
And	0.029285345341439308	A	0.026512066810828253
A	0.027340728752860085	And	0.024898260005767937
Of	0.02500643636916756	Of	0.023067326848640282
To	0.02180894491381715	To	0.023010973444286924
Is	0.018746361906154245	Br	0.017722593184773972
In	0.01650182932630161	Is	0.01677011015433098
Br	0.016100328886371074	It	0.015980610009027594
It	0.01566281702813352	I	0.015435860433611823
I	0.013237152389035115	In	0.0146325481793983
That	0.011622550887627713	This	0.01356570088717938
This	0.011328009734907588	That	0.012443605159319913
S	0.010797513169698165	S	0.010341954667553587
As	0.008708850907890853	Was	0.008688921473188487
With	0.007572072553870663	Movie	0.00820052530212607
For	0.007358691463305171	But	0.007252462146534321
Was	0.007143697921173693	For	0.007249147240395888
Film	0.006858831477849484	With	0.007003291701795464
But	0.006787836089460964	As	0.006872905393683778
Movie	0.006296623365175669	T	0.006712684930326198
His	0.005731727833225209	Film	0.006483956406774342
On	0.005575320031324559	You	0.0058104779763161005
He	0.005425362035687853	On	0.005683959058699253
You	0.005385050746538201	Not	0.005480644815542048
Are	0.004852404245907464	Have	0.004980646472995117
Not	0.004646010445461244	Be	0.004851365133596242
T	0.004510027030063085	Are	0.004811033775578645
One	0.004398230388154716	He	0.004692802123307879
Have	0.0040961644614599885	One	0.004430924538371696
Be	0.004022529173279958	They	0.004322085120159823

Макс наив. Байес. вес	Мин наив. Байес. вес	Макс веса	Мин веса	Макс поз частота	Макс нег частота	Мин поз частота	Мин нег частота
Sox	Boll	4.03	-4.44	0.00003117	0.00000055	0.00000054	0.00004586
Kolchak	Uwe	3.96	-4.07	0.00002902	0.00000055	0.00000054	0.00003149
Corbett	Dreck	3.90	-3.85	0.00002741	0.00000055	0.00000054	0.00002541
Adele	Ariel	3.78	-3.78	0.00002419	0.00000055	0.00000054	0.00002376
Trier	Seagal	3.66	-3.72	0.00002150	0.00000055	0.00000107	0.00004475
Mclaglen	Arquette	3.56	-3.58	0.00001935	0.00000055	0.00000054	0.00001934
Giovanna	Unwatchable	3.56	-3.57	0.00001935	0.00000055	0.00000107	0.00003812
Luzhin	Embarrassingly	3.53	-3.49	0.00001881	0.00000055	0.00000054	0.00001768
Clara	Stinker	3.47	-3.49	0.00001774	0.00000055	0.00000107	0.00003536
Haines	Wayans	3.41	-3.46	0.00001666	0.00000055	0.00000054	0.00001713
Philo	Hammerhead	3.41	-3.29	0.00001666	0.00000055	0.00000054	0.00001436
Anton	Varna	3.37	-3.29	0.00001612	0.00000055	0.00000054	0.00001436
Bourne	Welch	3.36	-3.29	0.00004784	0.00000166	0.00000054	0.00001436
Vertigo	Ripley	3.34	-3.29	0.00001559	0.00000055	0.00000054	0.00001436
Delightfully	Turgid	3.30	-3.21	0.00001505	0.00000055	0.00000054	0.00001326
Stardust	Gamera	3.30	-3.18	0.00001505	0.00000055	0.00000107	0.00002597
Anchors	Rangers	3.30	-3.18	0.00001505	0.00000055	0.00000107	0.00002597

Eustache	Awfulness	3.27	-3.16	0.00001451	0.00000055	0.00000054	0.00001271
Lindy	Kirkland	3.27	-3.16	0.00001451	0.00000055	0.00000054	0.00001271
Melbourne	Revolting	3.23	-3.16	0.00001397	0.00000055	0.00000054	0.00001271
Sho	Dreadfully	3.23	-3.16	0.00001397	0.00000055	0.00000054	0.00001271
Moonstruck	Blah	3.19	-3.13	0.00001344	0.00000055	0.00000269	0.00006133
Kazan	Paycheck	3.19	-3.12	0.00001344	0.00000055	0.00000054	0.00001215
Feinstone	Aag	3.19	-3.12	0.00001344	0.00000055	0.00000054	0.00001215
Dev	Thinner	3.19	-3.12	0.00001344	0.00000055	0.00000054	0.00001215
Astaire	Aztec	3.18	-3.12	0.00003977	0.00000166	0.00000054	0.00001215
Pecker	Atrocious	3.15	-3.11	0.00001290	0.00000055	0.00000269	0.00006022
Johansson	Yawn	3.15	-3.05	0.00001290	0.00000055	0.00000107	0.00002265
Alvin	Segal	3.15	-3.02	0.00002580	0.00000110	0.00000054	0.00001105
tigerland	interminable	3.15	-2.92	0.00001290	0.00000055	0.00000054	0.00000994

e) Классификатор обучается 4.16 секунд, на обработку тестовой выборки уходит 10 секунд. Точность классификатора на тренировочной выборке равна 99.67, на валидационной - 87.85, на тестовой – 73.40.

f) Добавление биграмм помогло улучшить результат примерно на два процента на валидационной выборке, а так же помогло сильно улучшить результат на тренировочной выборке до 99 процентов, добавление триграмм улучшило тренировочную выборку до 100 процентов, но ухудшило результат валидационной. оптимальным в моем случае оказалось использовать исключительно биграмм без использования униграмм.