Отчет по заданию 1 “Сглаживание”

Основные задачи

- Взять отрывок из стихотворения “Дом, который построил Джек”
- Разбить стихотворение на обучающую и тестовую части
- Убрать из текста знаки препинания, заменив при этом дефис пробелом
- Разбить обучающую выборку на униграммы и биграммы, подсчитать их вероятности
- Провести для тестовой выборки аддитивное сглаживание и сглаживание с подбором параметра λ согласно закону Лидстоуна
- Подсчитать перплексию для униграмм и биграмм

Униграммы

Обучение на обучающем множестве

Исходная вероятность униграммы:

$$P(w_i) = \frac{c_i}{N}$$

Исходные вероятности

#	Aa Words	# Probabilities
0	вот	0.02702702702702703
1	дом	0.02702702702702703
2	который	0.08108108108108109
3	построил	0.08108108108108109
4	джек	0.08108108108108109
5	а	0.05405405405405406
6	это	0.05405405405405406
7	пшеница	0.02702702702702703

#	Aa Words	# Probabilities
8	которая	0.08108108108108109
9	в	0.10810810810810811
10	темном	0.05405405405405406
11	чулане	0.05405405405405406
12	хранится	0.05405405405405406
13	доме	0.05405405405405406
14	веселая	0.02702702702702703
15	птица	0.02702702702702703
16	синица	0.02702702702702703
17	часто	0.02702702702702703
18	ворует	0.02702702702702703
19	пшеницу	0.02702702702702703

Вероятности при аддитивном сглаживании (сглаживание Лапласа)

$$P_{Lap}(w_i) = \frac{c_i + 1}{N + V}$$

Сглаживание Лапласа для униграмм

#	Aa Words	# Probabilities
0	вот	0.03508771929824561
1	кот	0.017543859649122806
2	который	0.07017543859649122
3	пугает	0.017543859649122806
4	и	0.017543859649122806
5	ловит	0.017543859649122806
6	синицу	0.017543859649122806
7	которая	0.07017543859649122
8	часто	0.03508771929824561
9	ворует	0.03508771929824561
10	пшеницу	0.03508771929824561

#	Aa Words	# Probabilities
11	в	0.08771929824561403
12	темном	0.05263157894736842
13	чулане	0.05263157894736842
14	хранится	0.05263157894736842
15	доме	0.05263157894736842
16	построил	0.07017543859649122
17	джек	0.07017543859649122

Сглаживание с подбором λ (закон Лидстоуна)

$$P_{Lid}(w_i) = \frac{c_i + \lambda}{N + B\lambda}$$

$\lambda = 0.25$

Закон Линдстоуна с параметром 0.25

#	Aa Words	# Probabilities
0	вот	0.02976190476190476
1	кот	0.005952380952380952
2	который	0.07738095238095238
3	пугает	0.005952380952380952
4	и	0.005952380952380952
5	ловит	0.005952380952380952
6	синицу	0.005952380952380952
7	которая	0.07738095238095238
8	часто	0.02976190476190476
9	ворует	0.02976190476190476
10	пшеницу	0.02976190476190476
11	в	0.10119047619047619
12	темном	0.05357142857142857
13	чулане	0.05357142857142857
14	хранится	0.05357142857142857
15	доме	0.05357142857142857

#	Aa Words	# Probabilities
16	построил	0.07738095238095238
17	джек	0.07738095238095238

$\lambda = 0.5$

Закон Линдстоуна с параметром 0.5

#	Aa Words	# Probabilities
0	вот	0.031914893617021274
1	кот	0.010638297872340425
2	который	0.07446808510638298
3	пугает	0.010638297872340425
4	и	0.010638297872340425
5	ловит	0.010638297872340425
6	синицу	0.010638297872340425
7	которая	0.07446808510638298
8	часто	0.031914893617021274
9	ворует	0.031914893617021274
10	пшеницу	0.031914893617021274
11	в	0.09574468085106383
12	темном	0.05319148936170213
13	чулане	0.05319148936170213
14	хранится	0.05319148936170213
15	доме	0.05319148936170213
16	построил	0.07446808510638298
17	джек	0.07446808510638298

$\lambda = 0.75$

Закон Линдстоуна с параметром 0.75

#	Aa Words	# Probabilities
0	вот	0.03365384615384615
1	кот	0.014423076923076924
2	который	0.07211538461538461

#	Aa Words	# Probabilities
3	пугает	0.014423076923076924
4	и	0.014423076923076924
5	ловит	0.014423076923076924
6	синицу	0.014423076923076924
7	которая	0.07211538461538461
8	часто	0.03365384615384615
9	ворует	0.03365384615384615
10	пшеницу	0.03365384615384615
11	в	0.09134615384615384
12	темном	0.052884615384615384
13	чулане	0.052884615384615384
14	хранится	0.052884615384615384
15	доме	0.052884615384615384
16	построил	0.07211538461538461
17	джек	0.07211538461538461

Перплексия

$+1 : 16.175358888003007$

$\lambda = 0.25 : 20.976732603415496$

$\lambda = 0.5 : 18.23156241301383$

$\lambda = 0.75 : 16.946076133700398$

Биграммы

Обучение на обучающем множестве

$$P(w_n|w_{n-1}) = \frac{c(w_n w_{n-1})}{c(w_{n-1})}$$

Исходные вероятности

#	Aa Words	# Probabilities
0	('вот', 'дом')	1

#	Aa Words	# Probabilities
1	('дом', 'который')	1
2	('который', 'построил')	1
3	('построил', 'джек')	1
4	('джек', 'а')	0.6666666666666666
5	('а', 'это')	1
6	('это', 'пшеница')	0.5
7	('пшеница', 'которая')	1
8	('которая', 'в')	0.6666666666666666
9	('в', 'темном')	0.5
10	('темном', 'чулане')	1
11	('чулане', 'хранится')	1
12	('хранится', 'в')	1
13	('в', 'доме')	0.5
14	('доме', 'который')	1
15	('это', 'веселая')	0.5
16	('веселая', 'птица')	1
17	('птица', 'синица')	1
18	('синица', 'которая')	1
19	('которая', 'часто')	0.3333333333333333
20	('часто', 'ворует')	1
21	('ворует', 'пшеницу')	1
22	('пшеницу', 'которая')	1

Аддитивное сглаживание

$$P_{Lap}(w_n|w_{n-1}) = \frac{c(w_n w_{n-1}) + 1}{c(w_{n-1}) + V^2}$$

Сглаживание Лапласа для биграмм

#	Aa Words	# Probabilities
0	('вот', 'кот')	0.001890359168241966
1	('кот', 'который')	0.001890359168241966

#	Aa Words	# Probabilities
2	('который', 'пугает')	0.001890359168241966
3	('пугает', 'и')	0.001890359168241966
4	('и', 'ловит')	0.001890359168241966
5	('ловит', 'синицу')	0.001890359168241966
6	('синицу', 'которая')	0.001890359168241966
7	('которая', 'часто')	0.0037593984962406013
8	('часто', 'ворует')	0.0037735849056603774
9	('ворует', 'пшеницу')	0.0037735849056603774
10	('пшеницу', 'которая')	0.0037735849056603774
11	('которая', 'в')	0.005639097744360902
12	('в', 'темном')	0.005628517823639775
13	('темном', 'чулане')	0.005649717514124294
14	('чулане', 'хранится')	0.005649717514124294
15	('хранится', 'в')	0.005649717514124294
16	('в', 'доме')	0.005628517823639775
17	('доме', 'который')	0.005649717514124294
18	('который', 'построил')	0.007518796992481203
19	('построил', 'джек')	0.007518796992481203

Сглаживание с подбором λ

Общая формула подсчета вероятности:

$$P_{Lid}(w_1...w_n) = \frac{C(w_1...w_n) + \lambda}{N + B\lambda}, \qquad B = V$$

где V— словарь n-грамм.

$\lambda = 0.25$

Закон Линдстоуна при безусловном подсчете частотности биграмм с параметром 0.25

#	Aa Words	# Probabilities
0	('вот', 'кот')	0.005988023952095809
1	('кот', 'который')	0.005988023952095809

#	Aa Words	# Probabilities
2	('который', 'пугает')	0.005988023952095809
3	('пугает', 'и')	0.005988023952095809
4	('и', 'ловит')	0.005988023952095809
5	('ловит', 'синицу')	0.005988023952095809
6	('синицу', 'которая')	0.005988023952095809
7	('которая', 'часто')	0.029940119760479042
8	('часто', 'ворует')	0.029940119760479042
9	('ворует', 'пшеницу')	0.029940119760479042
10	('пшеницу', 'которая')	0.029940119760479042
11	('которая', 'в')	0.05389221556886228
12	('в', 'темном')	0.05389221556886228
13	('темном', 'чулане')	0.05389221556886228
14	('чулане', 'хранится')	0.05389221556886228
15	('хранится', 'в')	0.05389221556886228
16	('в', 'доме')	0.05389221556886228
17	('доме', 'который')	0.05389221556886228
18	('который', 'построил')	0.07784431137724551
19	('построил', 'джек')	0.07784431137724551

$\lambda = 0.5$

Закон Линдстоуна при безусловном подсчете частотности биграмм с параметром 0.5

#	Aa Words	# Probabilities
0	('вот', 'кот')	0.010526315789473684
1	('кот', 'который')	0.010526315789473684
2	('который', 'пугает')	0.010526315789473684
3	('пугает', 'и')	0.010526315789473684
4	('и', 'ловит')	0.010526315789473684
5	('ловит', 'синицу')	0.010526315789473684
6	('синицу', 'которая')	0.010526315789473684
7	('которая', 'часто')	0.031578947368421054

#	Aa Words	# Probabilities
8	('часто', 'ворует')	0.031578947368421054
9	('ворует', 'пшеницу')	0.031578947368421054
10	('пшеницу', 'которая')	0.031578947368421054
11	('которая', 'в')	0.05263157894736842
12	('в', 'темном')	0.05263157894736842
13	('темном', 'чулане')	0.05263157894736842
14	('чулане', 'хранится')	0.05263157894736842
15	('хранится', 'в')	0.05263157894736842
16	('в', 'доме')	0.05263157894736842
17	('доме', 'который')	0.05263157894736842
18	('который', 'построил')	0.07368421052631578
19	('построил', 'джек')	0.07368421052631578

$\lambda = 0.75$

Закон Линдстоуна при безусловном подсчете частотности биграмм с параметром 0.75

#	Aa Words	# Probabilities
0	('вот', 'кот')	0.014084507042253521
1	('кот', 'который')	0.014084507042253521
2	('который', 'пугает')	0.014084507042253521
3	('пугает', 'и')	0.014084507042253521
4	('и', 'ловит')	0.014084507042253521
5	('ловит', 'синицу')	0.014084507042253521
6	('синицу', 'которая')	0.014084507042253521
7	('которая', 'часто')	0.03286384976525822
8	('часто', 'ворует')	0.03286384976525822
9	('ворует', 'пшеницу')	0.03286384976525822
10	('пшеницу', 'которая')	0.03286384976525822
11	('которая', 'в')	0.051643192488262914
12	('в', 'темном')	0.051643192488262914
13	('темном', 'чулане')	0.051643192488262914

#	Aa Words	# Probabilities
14	('чулане', 'хранится')	0.051643192488262914
15	('хранится', 'в')	0.051643192488262914
16	('в', 'доме')	0.051643192488262914
17	('доме', 'который')	0.051643192488262914
18	('который', 'построил')	0.07042253521126761
19	('построил', 'джек')	0.07042253521126761

Перплексия

$+1 : 273.7287283373542$

$\lambda = 0.25 : 139.09749617838835$

$\lambda = 0.5 : 200.17679598819893$

$\lambda = 0.75 : 242.19156910489474$

Сглаживание с подбором λ. Подсчет частотности биграмм в зависимости от первого слова в биграмме

В ходе выполнения задания было выяснено, что существует еще один подход к сглаживанию с параметром λ, когда мы не просто подсчитываем вероятность появления биграммы как независимого объекта, а определяем частотность появления в биграмме слова после конкретного слова (формула почти аналогична формуле для сглаживания +1)

Общая формула подсчета вероятности:

$$P_{Lid}(w_n|w_1...w_{n-1}) = \frac{C(w_1...w_n) + \lambda}{C(w_1...w_{n-1}) + B\lambda}, \qquad B = V^2$$

где V— словарь n-грамм.

$\lambda = 0.25$

Закон Линдстоуна с параметром 0.25

#	Aa Words	# Probabilities
0	('вот', 'кот')	0.001890359168241966
1	('кот', 'который')	0.001890359168241966

#	Aa Words	# Probabilities
2	('который', 'пугает')	0.001890359168241966
3	('пугает', 'и')	0.001890359168241966
4	('и', 'ловит')	0.001890359168241966
5	('ловит', 'синицу')	0.001890359168241966
6	('синицу', 'которая')	0.001890359168241966
7	('которая', 'часто')	0.009242144177449169
8	('часто', 'ворует')	0.009380863039399626
9	('ворует', 'пшеницу')	0.009380863039399626
10	('пшеницу', 'которая')	0.009380863039399626
11	('которая', 'в')	0.0166358595194085
12	('в', 'темном')	0.01651376146788991
13	('темном', 'чулане')	0.01675977653631285
14	('чулане', 'хранится')	0.01675977653631285
15	('хранится', 'в')	0.01675977653631285
16	('в', 'доме')	0.01651376146788991
17	('доме', 'который')	0.01675977653631285
18	('который', 'построил')	0.024029574861367836
19	('построил', 'джек')	0.024029574861367836

$\lambda = 0.5$

Закон Линдстоуна с параметром 0.5

#	Aa Words	# Probabilities
0	('вот', 'кот')	0.001890359168241966
1	('кот', 'который')	0.001890359168241966
2	('который', 'пугает')	0.001890359168241966
3	('пугает', 'и')	0.001890359168241966
4	('и', 'ловит')	0.001890359168241966
5	('ловит', 'синицу')	0.001890359168241966
6	('синицу', 'которая')	0.001890359168241966
7	('которая', 'часто')	0.005607476635514018
8	('часто', 'ворует')	0.005649717514124294

#	Aa Words	# Probabilities
9	('ворует', 'пшеницу')	0.005649717514124294
10	('пшеницу', 'которая')	0.005649717514124294
11	('которая', 'в')	0.009345794392523364
12	('в', 'темном')	0.00931098696461825
13	('темном', 'чулане')	0.009380863039399626
14	('чулане', 'хранится')	0.009380863039399626
15	('хранится', 'в')	0.009380863039399626
16	('в', 'доме')	0.00931098696461825
17	('доме', 'который')	0.009380863039399626
18	('который', 'построил')	0.013084112149532711
19	('построил', 'джек')	0.013084112149532711

$\lambda = 0.75$

Закон Линдстоуна с параметром 0.75

#	Aa Words	# Probabilities
0	('вот', 'кот')	0.001890359168241966
1	('кот', 'который')	0.001890359168241966
2	('который', 'пугает')	0.001890359168241966
3	('пугает', 'и')	0.001890359168241966
4	('и', 'ловит')	0.001890359168241966
5	('ловит', 'синицу')	0.001890359168241966
6	('синицу', 'которая')	0.001890359168241966
7	('которая', 'часто')	0.004377736085053158
8	('часто', 'ворует')	0.0043997485857950975
9	('ворует', 'пшеницу')	0.0043997485857950975
10	('пшеницу', 'которая')	0.0043997485857950975
11	('которая', 'в')	0.006879295622263915
12	('в', 'темном')	0.006862133499688085
13	('темном', 'чулане')	0.006896551724137931
14	('чулане', 'хранится')	0.006896551724137931
15	('хранится', 'в')	0.006896551724137931

#	Aa Words	# Probabilities
16	('в', 'доме')	0.006862133499688085
17	('доме', 'который')	0.006896551724137931
18	('который', 'построил')	0.009380863039399626
19	('построил', 'джек')	0.009380863039399626

Перплексия

$+1 : 273.7287283373542$

$\lambda = 0.25 : 139.09749617838835$

$\lambda = 0.5 : 200.17679598819893$

$\lambda = 0.75 : 242.19156910489474$

Заметим, что перплексии двух различных вариаций закона Лидстоуна совпадают при одинаковых λ.