Лекція 5. Наївний байєсівський метод: модель Бернуллі [1].

5.1. Алгоритм

Існує два види наївної байєсівської моделі. Одна з них — мультиноміальна — була описана вище. Вона генерує однин термін зі словника на кожній позиції документа.

Альтернативою мультиноміальній моделі є *багатомірна модель Бернуллі* (multivariate Bernoulli model), чи просто *модель Бернуллі*. Вона еквівалентна бінарній моделі незалежності, що генерує індикатор для кожного терміну словника: 1, якщо термін є присутнім у документі, і 0, якщо відсутнім. На рис. 5.1 показані алгоритми навчання і тестування для моделі Бернуллі. Часова складність моделі Бернуллі така ж, як і в мультиноміальній моделі.

```
TrainBernoulliNB(ℂ, 𝔻)
1   V ← ExtractVocabulary(𝔻)
2   N ← CountDocs(𝔻)
3   for each c∈ ℂ
4   do N_c ← CountDocsInClass(𝔻, c)
5       prior[c] ← N_c/N
6       for each t∈ V
7       do N_ct ← CountDocsInClassContainingTerm(𝔻, c, t)
8          condprob[t][c] ← (N_ct + 1)/(N_c + 2)
9   return V, prior, condprob
ApplyBernoulliNB(ℂ, V, prior, condprob, d)
1   V_d ← ExtractTermsFromDoc(V, d)
2   for each c∈ ℂ
3   do score[c] ← log prior[c]
4      for each t ∈ V
5      do if t ∈ V
6            then score[c] += log condprob[t][c]
7            Else score[c] += log (1 – condprob[t][c])
8   return arg max_{c∈ℂ} score[c]
```

Рис. 5.1. Наївний байєсівський алгоритм (модель Бернуллі): навчання і тестування. У рядку 8 застосоване додавання одиниці, аналогічно згладжуванню за Лапласом, де B = 2.

Різні моделі використовують різні стратегії оцінки і різні правила класифікації. У моделі Бернуллі імовірність $\hat{P}(t|c)$ оцінюється як *частка документів із класу c*, що містять термін *t* (рис. 5.1, алгоритм TrainBernoulliNB, рядок 8). На противагу йому в мультиноміальній моделі імовірність $\hat{P}(t|c)$ оцінюється як *частка лексем*, або *частка позицій* у документах з класу *c*, що містять термін *t*. При класифікації тестового документа на основі моделі Бернуллі використовується бінарна інформація про появу терміну, що ігнорує кількість входжень цього терміну, у той час як мультиноміальна модель відслідковує багаторазові появи терміну в документі. У результаті при класифікації довгих документів модель Бернуллі, як правило, допускає багато помилок. Наприклад, вона може віднести до класу *China* цілу книгу через єдине згадування терміну *China*.

Ці моделі розрізняються також тим, як використовуються терміни, що не з'являються в документі. У мультиноміальній моделі ця інформація ніяк не впливає на рішення, а в моделі Бернуллі імовірність відсутності терміну при обчисленні імовірності $P(c|d)$ факторизується (рис. 5.1, ApplyBernoulliNB, рядок 7). Це пояснюється тим, що только моделі Бернуллі враховують інформацію про відсутність терміну явно.

Приклад 5.1. Застосувавши модель Бернуллі до даних, наведених у табл. 4.1, ми одержали оцінки $\hat{P}(c)$ =3/4 і $\hat{P}(\bar{c})$ =1/4. Умовні імовірності набувають наступні значення. $\hat{P}$(Chinese|*c*) = (3 + 1)/(3 + 2) = 4/5,

$\hat{P}$(Japan|*c*) = $\hat{P}$(Tokio|*c*) = (0 + 1)/(3 + 2) = 1/5,

$\hat{P}$(Beijing|*c*) = $\hat{P}$(Macao|*c*) = $\hat{P}$(Shanghai|*c*) = (1 + 1)/(3 + 2) = 2/5,

$\hat{P}$(Chinese|$\bar{c}$) = (1 + 1)/(1 + 2) = 2/3,

$\hat{P}$(Japan|$\bar{c}$) = $\hat{P}$(Tokyo|$\bar{c}$) = (1 + 1)/(1 + 2) = 2/3,

$\hat{P}$(Beijing|$\bar{c}$) = $\hat{P}$(Macao|$\bar{c}$) = $\hat{P}$(Shanghai|$\bar{c}$) = (0 + 1)/(1 + 2) = 1/3.

Знаменники рівні (3 + 2) і (1 + 2), оскільки в класі *c* існують три документи, а в класі $\bar{c}$ — один документ, а константа *B* у рівності Лагранжа дорівнює двом: для кожного терміну існують два варіанти — вони або входять у документ, або ні.

Ранги тестового документа стосовно цих двох класів такі.

$$\hat{P}(c|d_5) \propto \hat{P}(c)\cdot\hat{P}(\text{Chinese}|c)\cdot\hat{P}(\text{Japan}|c)\cdot\hat{P}(\text{Tokyo}|c)\cdot$$
$$\cdot\left(1-\hat{P}(\text{Beijing}|c)\right)\cdot\left(1-\hat{P}(\text{Shanghai}|c)\right)\cdot\left(1-\hat{P}(\text{Makao}|c)\right)=$$
$$=3/4\cdot 4/5\cdot 1/5\cdot 1/5\cdot(1-2/5)\cdot(1-2/5)\cdot(1-2/5)\approx 0{,}005.$$

Аналогічно,

$$\hat{P}(\bar{c}|d_5) \propto 1/4\cdot 2/3\cdot 2/3\cdot 2/3\cdot(1-1/3)\cdot(1-1/3)\cdot(1-1/3)\approx 0{,}022.$$

Отже, класифікатор віднесе тестовий документ до класу $\bar{c}$ = *not-China*. Якщо враховувати тільки бінарний індикатор входження терміну в документа, а не його частоту, то індикатори термінів Japan і Tokyo є індикаторами класу $\bar{c}$ (2/3 > 1/5), а умовні імовірності терміну Chinese для класів c і $\bar{c}$ недостатньо сильно відрізняються друг від друга (4/5 від 2/3) і не впливають на остаточну класифікацію.

5.2. Властивості наївної байєсівської моделі

Для того щоб краще зрозуміти ці дві моделі, а також припущення, на яких вони засновані, повернемося назад і перевіримо, як ми вивели правила класифікації раніше. Нагадаємо, что рішення про належність до класу набувається шляхом визначення класу з максимальною апостеріорною імовірністю.

$$c_{map} = \arg\max_{c\in\mathbb{C}} \hat{P}(c|d) = \tag{5.1}$$
$$= \arg\max_{c\in\mathbb{C}} \frac{P(d|c)P(c)}{P(d)} =$$
$$= \arg\max_{c\in\mathbb{C}} P(d|c)P(c). \tag{5.2}$$

Застосовуючи до виразу (5.1) правило Байєса, можна відкинути знаменник, оскільки він є постійним для всіх класів і не впливає на величину argmax.

Вираз (5.2) можна інтерпретувати як опис процесу генерування, характерного для байєсівської класифікації текстів. Для того щоб згенерувати документ, ми спочатку вибираємо клас c з імовірністю $P(c)$ (верхні вузли на рис. 5.2 і 5.3). Ці дві моделі відрізняються способом формалізації другого етапу, тобто генерування документа по заданому класі відповідно до умовного розподілу $P(d|c)$.

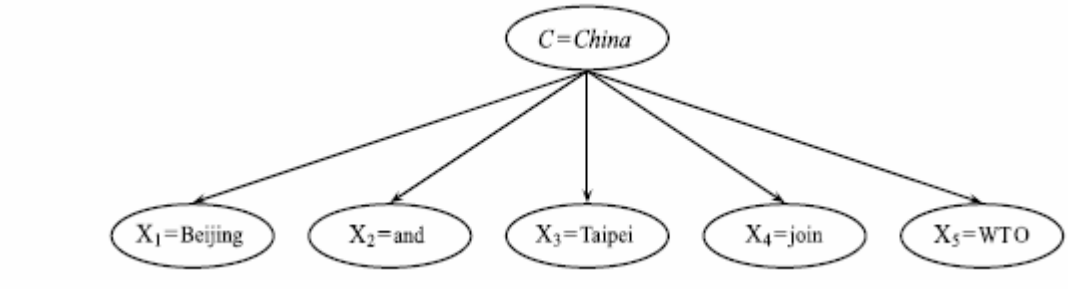

Рис. 5.2. Мультиноміальна наївна байєсівська модель

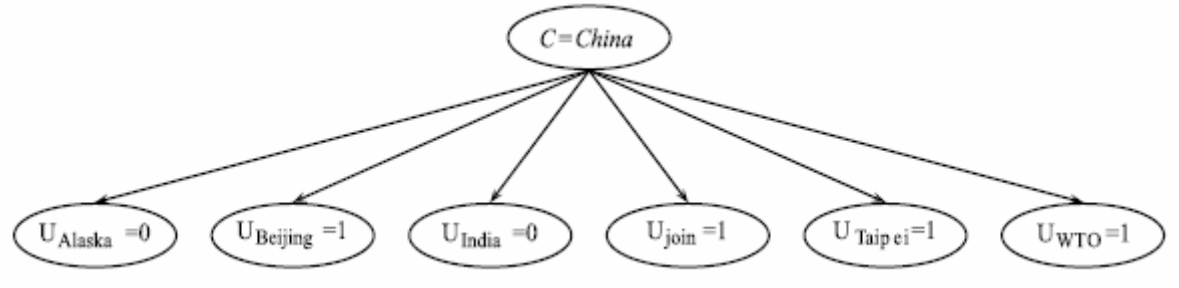

Рис. 5.3. Бернуллієвська наївна байєсівська модель

Мультиноміальна модель $P(d|c) = P\left(\langle t_1,...,t_k,...,t_{n_d}\rangle \middle| c\right)$ (5.3)

Бернуллієвська модель $P(d|c) = P\left(\langle e_1,...,e_i...,e_M\rangle \middle| c\right)$ (5.4)

Тут $\langle t_1,...,t_k,...,t_{n_d}\rangle$ — послідовність термінів, що з'являються в документі d (за винятком термінів, виключених зі словника), а $\langle e_1,...,e_i...,e_M\rangle$ — бінарний вектор, що складається з M індикаторів присутності кожного терміну в документі d.

Тепер зрозуміліше, чому в рівності (5.1), описуючи задачу класифікації текстів, ми ввели простір документів $\mathbb{X}$. Критичний крок у розв'язуванні задачі класифікації текстів — вибір представлення документа. Наприклад, такими представленнями документа є послідовності $\langle t_1,...,t_k,...,t_{n_d} \rangle$ і $\langle e_1,...,e_i...,e_M \rangle$. У першому варіанті простір $\mathbb{X}$ є множиною всіх послідовностей термінів (чи, точніше, послідовностей лексем термінів). В другому варіанті простір $\mathbb{X}$ являє собою множину $[0,1]^M$.

Рівності (5.3) і (5.4) неможливо безпосередньо застосувати для класифікації текстів. Наприклад, використовуючи модель Бернуллі, ми були б змушені оцінити $2^M|\mathbb{C}|$ різних параметрів, по одному для кожного можливих з M сполучень значень e_i і класу. Кількість параметрів у мультиноміальній моделі приблизно така ж. [1] Оскільки кількість параметрів в обох випадках дуже велика, їх оцінка стає практично нерозв'язною проблемою.

Для того щоб зменшити кількість параметрів, сформулюємо *припущення про умовну незалежність* (conditional independence assumption). Будемо вважати, что, якщо клас заданий, то значення атрибутів не залежать одна від одної.

$$\textbf{Мультиноміальна модель} \quad P(d|c) = P\left(\langle t_1,...,t_{n_d} \rangle | c\right) = \prod_{1 \le k \le n_d} P(X_k = t_k|c) \tag{5.5}$$

$$\textbf{Бернуллієвська модель} \quad P(d|c) = P\left(\langle e_1,...,e_M \rangle | c\right) = \prod_{1 \le i \le M} P(U_i = e_i|c) \tag{5.6}$$

Тут ми ввели дві випадкові величини, щоб явно розрізняти різні моделі генерування ознак. Змінна X_k — це випадкова величина для позиції k у документі, що набуває значення термінів зі словника, а $P(X_k = t|c)$ — імовірність того, що в документі з класу c термін t з'явиться в позиції k. Змінна U_i — це випадкова величина для словникового терміну i, що набуває значення 0 (термін є присутнім) і 1 (термін відсутній). Імовірність $P(U_i = 1|c)$ — це імовірність того, що в документі з класу c з'явиться термін t_i — у будь-якій позиції і, можливо, багаторазово.

Припущення про умовну незалежність проілюстроване на рис. 5.2 і 5.3. На цих малюнках показано, что клас *China* породжує значення для кожного з п'яти (у мультиноміальній моделі) чи шести (у бернуллієвській моделі) атрибутів терміну з визначеною імовірністю, що не залежить від значень інших атрибутів. Інакше кажучи, імовірність того, що документ у класі *China* містить термін Taipei ніяк не залежить від того, чи міститься в ньому термін Beijing.

На практиці припущення про умовну незалежність для текстових даних не виконується. Терміни *залежать* друг від друга. Однак, як буде незабаром показано, незважаючи на це наївні байєсівські моделі працюють досить добре.

Навіть якщо припущення про умовну незалежність виконується, то за умови, що кожна позиція k у документі має власний розподіл імовірностей, у мультиноміальній моделі як і раніше залишається занадто багато параметрів. Позиція терміну в документі сама по собі не містить інформації про клас. Незважаючи на різницю між термінами *China sues France* і *France sues China*, поява терміну *China* на першій, а не на третій позиції в документі нічого не дає для класифікації в рамках наївної байєсівської моделі, оскільки в ній усі терміни розглядаються ізольовано один від одного. Припущення про умовну незалежність змушує нас обробляти їх саме так.

Крім того, якщо розподіли термінів для кожної позиції k відрізняються один від одного, ми повинні оцінювати різні множини параметрів для кожної позиції k. Імовірність того, що термін bean з'явиться як перший термін у документі *coffie* повинна відрізнятися від імовірності того, що він виявиться другим і т.д. Це знову породжує проблеми з оцінками даних, що рідко зустрічаються.

З цих причин ми висунемо друге припущення про незалежність у мультиноміальній моделі — *припущення про позиційну незалежність* (positional independence): умовні імовірності появи терміну однакові незалежні від його позиції в документі, тобто

$$P(X_{k_1} = t|c) = P(X_{k_2} = t|c) \ .$$

для всіх позицій k_1 і k_2, термінів t і класів c. Таким чином, у моделі виникає єдиний розподіл термінів для всіх позицій k_i. Будемо позначати цей розподіл буквою X. [2] Припущення про позиційну незалежність еквівалентне моделі «мішка слів».

Прийнявши припущення про умовну і позиційну незалежність, задачу можна звести до оцінки усього $\Theta(M|\mathbb{C}|)$ параметрів $P(t_k|c)$ у мультиноміальній чи моделі $P(e_i|c)$ у моделі Бернуллі по одному для кожної комбінації термін-клас, а не до оцінки величезної кількості параметрів, пропорційної степеню розміру словника M.

---

[1] Фактично, якщо довжини документів не обмежені, то кількість параметрів у мультиноміальній моделі є нескінченою.

[2] Це нестандартна термінологія. Випадкова величина X є категоріальною, а не мультиноміальною, тому відповідну наївну баєсівську модель варто було б назвати *моделлю послідовностей* (sequence model). Ми ототожнили модель послідовностей і мультиноміальну модель тому, що з обчислювальної точки зору вони є ідентичними.

Підіб'ємо підсумок. У мультиноміальній моделі (рис. 5.2) ми спочатку генеруємо документ, вибираючи клас $C = c$ з імовірністю $P(c)$, де C — випадкова величина, що набуває значення в просторі $\mathbb{C}$. Потім ми генеруємо термін t_k у позиції k з імовірністю $P(X_k = t_k|c)$ для кожної з n_d позицій документа. Усі величини X_k мають однаковий розподіл над термінами у фіксованому класі c. У прикладі, продемонстрованому на рис. 5.2, ми показали генерування послідовності $<t_1, t_2, t_3, t_4, t_5>$ = <Beijing, and, Taipei, join, WTO>, що відповідає документу, що містить одну пропозицію *Beijing and Taipei join WTO*.

Для того щоб модель генерування документа стала цілком визначеною, необхідно визначити розподіл імовірностей $P(n_d|c)$ по довжинах. Без цього ми одержимо модель генерування лексем, а не документа.

У моделі Бернуллі (рис. 5.3) ми спочатку генеруємо документ, вибираючи клас $C = c$ з імовірністю $P(c)$, а потім ми генеруємо бінарний індикатор e_i для кожного терміну t_i зі словника ($1 \le i \le M$). У прикладі, продемонстрованому на рис. 5.3, показаний процес генерування послідовності бінарних індикаторів $<e_1, e_2, e_3, e_4, e_5, e_6>$ = <0, 1, 0, 1, 1, 1, 1>, що відповідає, як і раніше, документу, що містить одне речення *Beijing and Taipei join WTO* у припущенні, що слово and є забороненим.

Результати порівняння цих двох моделей приведені в табл. 5.1, включаючи формули для обчислення оцінок і вирішальних правил.

Таблиця 5.1. Результати порівняння мультиноміальної моделі і моделі Бернуллі

	Мультиноміальна модель	*Модель Бернуллі*
Подія	Генерування терміну	Генерування документа
Випадкові змінні	$X = t$, *якщо і только якщо термін t зустрівся в заданій позиції*	$U_t = 1$, *якщо і только якщо термін t зустрівся в документі*
Представлення документа	$d = \langle t_1, t_2, ..., t_k, ..., t_{n_d} \rangle, t_k \in V$	$d = \langle e_1, e_2, ..., e_i, ..., e_M \rangle, e_i \in [0,1]$
Оцінка параметра	$\hat{P}(X = t\|c)$	$\hat{P}(U_i = e\|c)$
Вирішальне правило: максимум	$\hat{P}(c)\prod_{1\le k\le n_d}\hat{P}(X = t_k\|c)$	$\hat{P}(c)\prod_{t_i\in V}\hat{P}(U_i = e\|c)$
Кратність	Враховується	Ігнорується
Довжина документа	Справляється з довгими документами	Найкраще працює з короткими документами
Кількість ознак	Справляється з великою кількістю ознак	Краще працює з невеликою кількістю ознак
Оцінка терміну the	$\hat{P}(X = the\|c) \approx 0,05$	$\hat{P}(U_{the} = 1\|c) \approx 1,0$

Розглянута модель називається наївною байєсівської, оскільки зроблені нами припущення про незалежність є дійсно наївними для моделі природної мови. Припущення про умовну незалежність стверджує, що для заданого класу ознаки є незалежними друг від друга. Це припущення рідко виконується для термінів у документах. У багатьох випадках верним є протилежне твердження. Прикладами сильно залежних термінів є пары hong і kong чи london і english, зазначені на рис. 5.3. Крім того, мультиноміальна модель заснована на припущенні про позиційну незалежність. Модель Бернуллі ігнорує позиції в документі, оскільки в ній враховуються тільки присутність чи відсутність терміну. Ця модель «мішка слів» відкидає всю інформацію, що породжує порядок слів. Як же пояснити ефективність наївної байєсівської моделі, якщо модель природної мови так сильно спрощена?

Відповідь полягає в тім, что незважаючи на низьку якість *оцінок імовірності* в наївній байєсівській моделі, *рішення про класифікацію* на диво точні. Розглянемо документ d із істинними імовірностями $P(c_1|d) = 0,6$ і $P(c_2|d) = 0,4$, що наведені в табл. 4.1. Припустимо, що в документі d термінів, що свідчать на користь класу c_1, більше, ніж термінів, що свідчать на користь класу c_2. Таким чином, у мультиноміальній моделі (5.5) оцінка $\hat{P}(c_1)\prod_{1\le k\le n_d}\hat{P}(t_k|c_1)$ буде набагато більшою, ніж $\hat{P}(c_2)\prod_{1\le k\le n_d}\hat{P}(t_k|c_2)$ (у табл. 5.2: 0,00099 > 0,00001). Після ділення на 0,001, дійдемо висновку, что одна оцінка близька до 1,0, а друга — до 0,0. Узагальнимо сказане: результуючий клас у наївній байєсівській моделі звичайно має набагато більшу імовірність, ніж інші класи, а оцінки сильно розрізняються від справжніх значень. Однак остаточне рішення засноване на тому, який із класів одержує негативний ранг. При цьому зовсім не важливо, наскільки точними є оцінки. Незважаючи на неточність, наївні байєсівські оцінки приписують класу c_1 велику імовірність і, отже, відносять документ d до правильного класу. *Правильна оцінка означає точний прогноз, але точний прогноз не означає правильну оцінку.* Наївні байесівські класифікатори спираються на неточні оцінки, але дозволяють досягти високої точності класифікації.

Таблиця 5.2. Правильна оцінка означає правильний прогноз, але правильний прогноз не завжди означає правильну оцінку

	c_1	c_2	*Обраний клас*
Справжня імовірність $P(c\|d)$	0,6	0,4	c_1
$\hat{P}(c)\prod_{1\le k\le n_d}\hat{P}(t_k\|c)$ (формула (5.1)	0,00099	0,00001	
Оцінка NB $\hat{P}(c\|d)$	0,99	0,01	c_1

Навіть якщо наївний байєсівський метод не має найбільшу точність класифікації текстів, він має багато переваг. Якщо існує багато однаково важливих ознак, що впливають на остаточний висновок, то наївний байєсівський метод виявляється краще інших. Крім того, іноді він виявляється стійким до перешкод і *зсуву понять* (concept drift), тобто до поступової зміни поняття, що лежить в основі класу, наприклад, поняття *US president* зміщається від Білла Клінтона до Джорджа Бушу. Класифікатори, засновані на методі kNN, можна точно настроїти на особливості конкретного періоду часу. Однак ці настроювання можуть виявитися невірними, коли з часом поняття злегка зміняться.

Модель Бернуллі особливо стійка до зсуву понять. Крім того, вона може досягати пристойної точності, використовуючи менше дюжини термінів. Зміна найбільш важливих індикаторів класу малоймовірні. Отже, імовірність того, що модель, яка спирається тільки на такі індикатори, буде зберігати високий рівень точності при зсуві понять, досить велика.

Сила наївного байєсівського методу полягає в його ефективності: навчання і класифікацію можна провести за один прохід. Оскільки цей метод має як високу ефективність, так і гарну точність, його часто використовують як основний метод класифікації текстів. Його вибирають у першу чергу в ситуаціях, коли 1) утрата декількох відсотків точності не викликає занепокоєння, 2) існує великий обсяг навчальних даних і, завдяки йому, високу точність можна досягти за рахунок великої кількості даних, а не за рахунок більш якісного класифікатора, що використовує невеликий обсяг даних і 3) спостерігається зсув понять.

Наївний байєсівський метод вважається основним при класифікації текстів. Припущення про незалежність для текстів не виконується. Однак можна показати, что цей метод є *оптимальним класифікатором* (завдяки мінімальному рівню помилок, що спостерігається на нових даних) для даних, щодо яких ці припущення виконуються.

5.3. Варіант мультиноміальної моделі

Альтернативна формалізація мультиноміальної моделі передбачає представлення кожного документа *d* як *M-вимірного* вектора частот $\langle tf_{t_1,d},...,tf_{t_M,d}\rangle$, де $tf_{t_M,d}$ — частота терміну t_i у документі *d*. У такому випадку імовірність $P(d|c)$ обчислюється в такий спосіб.

$$P(d|c)=P\left(\langle tf_{t_1,d},...,tf_{t_M,d}\rangle\middle|c\right)\propto\prod_{1\le i\le M}P\left(X=t_i|c\right)^{tf_{t_i,d}}\ . \qquad (5.7)$$

Зверніть увагу на те, що ми пропустили мультиноміальний множник.

Формула (5.7) еквівалентна моделі послідовностей (5.5), оскільки для термінів, що не зустрічаються в документі *d*, тобто при $tf_{t_i,d}=0$, виконується рівність $P\left(X=t_i|c\right)^{tf_{t_i,d}}=1$, а терміни, що зустрічаються в документі, тобто для який виконується нерівність $tf_{t_i,d}\ge 1$, породжують той самий множник $tf_{t_i,d}$ як у формулі (5.5), так і у формулі (5.7).

Вправа 5.1. **[*]** Які з документів, приведених у табл. 4.1 мають представлення, що співпадає з представленням «мішка слів» 1) у моделі Бернуллі і 2) у мультиноміальній моделі? Якщо розходження є, то опишіть їх.

Таблиця 5.3. Сукупність документів, для яких виконання припущень про незалежність є проблематичним.

1	He moved from London, Ontario, to London, England.
2	He moved from London, England, to London, Ontario.
3	He moved from England to London, Ontario.

Вправа 5.2. Факт, що термін з'являється в документі в позиції *k*, не несе корисної інформації. Саме це дозволяє висунути припущення про незалежність термінів. Знайдіть виключення з цього правила. Розглянете формулярні документи з фіксованою структурою.

Вправа 5.3. У табл. 5.1 приведені оцінки для слова the, отримані в моделі Бернуллі і мультиноміальній моделі. Поясните різницю між ними.

Література

1. Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск.— М.: Изд-во «Вильямс», 2011.