

# ЭЛЕКТРОН РЕСУРСЛАРНИ СИНФЛАШ УСУЛИ ВА АЛГОРИТМЛАРИ

Мўминов Б.Б.

*Ушбу мақолада маълумотларни излашда электрон ресурсларни синфлаш муаммолари ва ечимлари, усуллари келтирилган. Электрон ресурсларни излашда Синфлаштириш соҳасида кўп қўлланиладиган моделлардан Матнларни синфлаш, Матнларни содда Байес усули орқали синфлаш усуллари ва алгоритмларини жорий қилиш зояси илгари сурилган.*

**Калим сўзлар:** *Электрон ресурс, излаш, сўров, синфлаштириш, ўқитиш усули, Байес усули, мултиномиал алгоритми, Бернулли модели.*

## METHODS AND ALGORITHMS OF CLASSIFICATION OF ELECTRONIC RESOURCES

Muminov B.B.

*This article explores the problems, solutions and methods for classification of electronic resources. Searching for Electronic Resources The classification of texts from the most commonly used models in classification, the introduction of methods and algorithms for classifying texts through simple Bayes method.*

**Key Words:** *E-resource, search, query, classification, teaching method, Bayes method, multinomial algorithm, Bernoulli model*

Кўп фойдаланувчилар доимий ахборот эҳтиёжига эга. Масалан, улар тежамкор техника ва автомашиналарнинг ишлаб чиқарилишига боғлиқ янгиликларни кузатиб боришлари мумкин. Бундай эҳтиёжни қондириш учун ҳар куни эрталаб [*тежамкор and техника and автомашина*] сўрови бўйича янгиликларни излаши мумкин. Бунинг учун кўп тизимлар излаш сўровига фойдаланувчи ёзилишин қўллашади (доимий сўровлар) (standing queries). Доимий сўров ҳар қандай бошқа сўровга ўхшаш, лекин у доимий равишда янги ҳужжатлар билан тўладиган тўпламда даврий бажарилади.

Агар доимий сўров [*тежамкор and техника and автомашина*] кўринишига эга бўлса, унда фойдаланувчи бошқа терминлар, масалан, [*енгил машина*] қўлланиладиган кўпгина бошқа янги долзарб мақолаларни ўтказиб юбориш ҳавфи бўлади. Излашнинг етарлича тўлиқлигини таъминлаш учун доимий сўровлар вақт ўтиши билан аниқлаштириб борилиши лозим ва шунинг учун улар тобора мураккаблашиб боради. Ушбу мисолда, стемминг билан мантиқий излаш тизимидан фойдаланиб, яқунда тахминан қуйидаги сўровга келиш мумкин: [*тежамкор and техника and (енгил and (автомашина or машина))*].

Умумийлик даражасини таърифлаш ва доимий сўровлар қайта ишланиши тегишли бўлган предмет соҳаси чегараларини белгилаш учун синфлаш масаласини аниқлаймиз (classification problem): берилган синфлар тўпламига эга бўлган ҳолда, берилган объект уларнинг қайси бирига (ёки қайси бирларига) тегишли эканлигини аниқлаш лозим. Ушбу мисолда доимий сўровнинг қайта ишланиши янгиликларнинг икки синфи бўлган синфлаш масаласига олиб келади: кўп ядроли компьютер микросхемалари ҳақидаги ҳужжатлар ва кўп ядроли бўлмаган компьютер микросхемалари ҳақидаги ҳужжатлар. Бундай масала бинар синфланиш (two-class classification) деб аталади. Доимий сўровлар асосидаги синфланиш шунингдек, маршрутизация (routing) ёки филтрация (filtering) деб аталади.

Синф [*тежамкор техника автомашина*] доимий сўровдаги каби тор бўлиши шарт эмас. Кўпинча у етарлича кенг предмет соҳасини қамраб олади, масалан, [*спорт*] ёки [*техника*]. Бундай кенг синфлар одатда мавзулар (topics), синфланиш масаласи эса матнлар синфланиши (text classification), матнлар категорияланиши (text categorization), тематик синфланиш (topic classification) ёки тематикани аниқлаш (topic spotting) деб аталади. Доимий сўров ва мавзулар аниқлаштириш даражаси билан фарқланади, лекин матнларни маршрутизация, фильтрация ва синфлаш усуллари аслида мос келади.

Синфлаш тушунчаси жуда умумий тушунча ва маълумот излаш соҳасида, ҳам ундан ташқарида кўп дастурларга эга. Масалан, компьютер соҳасида синфловчи образларни landscape, portrait ва neither каби синфларга ажратиш учун ишлатилиши мумкин. Бу параграфда эътиборимизни маълумотларни излашда синфлаш мисолларига қаратамиз.

Синфлаштириш соҳасида кўп қўлланиладиган моделлар:

- Хужжат кодировкасининг идентификацияси (ASCII, Unicode UTF-8 ва бошқалар), сўзлар сегментацияси (икки ҳарф орасидаги бўш жой сўз чегараси), сўзнинг ҳақиқий регистрини, шунингдек, хужжат тилини аниқлаш.

- Веб-спамни автоматик аниқлаш (бундай саҳифалар излаш тизимининг индексига киритилмайди).

- Таркибни автоматик аниқлаш (бундай материаллар излаш натижаларига кўшилади, агар фойдаланувчи ҳавфсиз излаш соҳасини ўчирган бўлса).

- Фикрларни ифодалаш (sentiment detection), ёки кино ёки товар ҳақида мулоҳазаларнинг автоматик синфланиши. Бундай дастурга фойдаланувчи фотоаппарат сотиб олишдан олдин, унда яширинча камчиликлар ва сифат билан боғлиқ муаммолар йўқлигига ишонч ҳосил қилиш учун салбий муҳокамаларни излаши мисол бўла олади.

- Электрон почтани саралаш. Фойдаланувчи бир нечта жилдлар яратиши мумкин, масалан, эълонлар, ҳисоблар, дўстлардан хатлар ва ҳоказо, келадиган электрон хабарларни синфлаш ва уларни автоматик тарзда мос келувчи жилдларга жойлаштириш. Хабарни кирувчи хабарларнинг жуда катта жилдидан топишдан кўра, хабарларнинг тартибланган жилдидан топиш анча осон. Кўпинча бу имкониятдан спамни фильтрациялаш учун фойдаланилади.

- Мавзули излашнинг излаш тизимлари излашни маълум мавзу билан чеклайди. Масалан, *машина* сўровига жавобан *Ўзбекистон* мавзуси бўйича вертикал излаш тизими оддий излаш тизими *Ўзбекистон машина* сўровига қайтарадиган жавобига қараганда, *Ўзбекистон* машиналари тўғрисидаги маълумотларни анча аниқ ва тўлиқ рўйхатини чиқаради. Бу вертикал излаш тизими ўзининг индексига *Ўзбекистон* терминини бошқа маъноларига эга веб-саҳифаларни қўшмаслиги мумкин, лекин долзарб саҳифаларни, ҳатто агар *Ўзбекистон* терминни уларда айтиб ўтилмаган бўлсада кўшиши билан тушунтирилади.

- Ихтиёрий маълумотларни излашда аҳамиятига қараб тартиблаш функцияси синфловчига асосланиши мумкин.

Бу рўйхатда маълумот излаш соҳасидаги синфлашнинг асосий дастурлари санаб ўтилган. Кўпгина замонавий излаш тизимлари синфловчилардан у ёки бу кўринишда фойдаланадиган бир нечта компонентларга эга. Синфлашнинг асосий мисоли сифатида матнлар синфланишидан фойдаланамиз. Синфлашни компьютерларсиз ҳам ўтказиш мумкин.

Синфлашнинг кўп масалалари анъанавий равишда қўлда ечилган. Масалан, кутубхоначилар китобларга мос бўлимни белгилашади.

Қоида синфга хос бўлган бир нечта калит сўзлардан ташкил топган. Қўлда ташкил этилган қоида масштаблаш учун яхши имкониятларга эга, лекин вақт ўтиши билан уларнинг яратилиши ва кузатилиши мураккаб бўлиб бораверади. Тажрибали инсонлар (масалан, доимий ифодалар билан яхши таниш бўлган предмет соҳасидаги мутахассислар) синфловчилар билан автоматик генерациялаш билан рақобатлашадиган ва ҳатто улардан ўтадиган қоидалар тўпламини яратишлари мумкин, лекин одатда бундай инсонларни топиш қийин.

Қўлда ёзилган, қўли синфланиш ва қоидалардан ташқари матнларни синфлашга учинчи ёндашув мавжуд, бу: машинали ўқув асосида матнларни синфлаш. Машинали ўқувда қоидалар тўплами, ёки умумий ҳолатда – қарор қабул қилиш омили ўқув танлов асосида автоматик чиқарилади. Агар ўқув усули статистик ҳисобланса, бу ёндашув матнларни статистик синфлаш деб аталади (statistical text classification). Матнларни статистик синфлаш учун ҳар бир синф учун бир нечта яхши мисоллар (ёки ўқув ҳужжатлар) талаб этилади. Қўлда синфлашнинг зарурияти шунингдек ҳисобга олинади, чунки ўқув ҳужжатлар уларни белгилайдиган инсондан келади. Белгилаш (Labeling) – бу ҳар бир ҳужжатга синфини кўрсатиш жараёни. Ҳужжатларнинг тематик белгиланиши – синфлаш қоидаларини тузишдан кўра анча содда жараён.

Деярли ҳар қандай инсон, ҳужжатга қараб, унда донолик ҳақида гапириляптими ёки йўқ айта олади. Баъзан бундай белгилаш мавжуд технологик жараённинг бир қисми ҳисобланади. Масалан, сиз доимий сўровга жавобан ҳар куни эрталаб олинандиган янгиликларни кўриб чиқишингиз ва махсус жилд, масалан, техника ва автомашинага долзарб мақолаларни жойлаштирган ҳолда долзарблик бўйича тесқари алоқани таъминлашингиз мумкин.

**Матнларни синфлаш.** Матнларни синфлаш масаласига  $d \in X$  ҳужжатни таърифлаш берилган, бу ерда  $X$  – ҳужжатлар тўплами (document space) ва  $C = \{c_1, c_2, \dots, c_n\}$  синфларнинг белгиланган тўплами. Синфлар шунингдек, категория (category) ва белгилар (labels) деб аталади. Қоидага кўра,  $X$  ҳужжатлар тўплами катта ўлчамга эга, синфлар эса дастурга қараб мутахассислар томонидан аниқланади. Бундан ташқари, белгиланган ҳужжатларнинг  $D$  (training set) ўқув тўплами берилган, бу ерда  $\langle d, c \rangle \in X \times C$ . Масалан, қуйидаги жуфтлик

$\langle d, c \rangle = \langle \text{Шанхай ҳамкорлик ташкилоти}, \text{Ўзбекистон} \rangle$

ўқув тўпламга тегишли, бу ерда  $d$  - *Шанхай ҳамкорлик ташкилоти* бир гапдан ташкил топган ҳужжат,  $c$  эса – *Ўзбекистон* синфи (ёки белгиси).

Ўқитиш усули (learning method), ёки ўқиш алгоритмидан (learning algorithm) фойдаланган ҳолда ҳужжатларни синфларга акс эттирадиган синфлаш функцияси (classification function) ёки синфловчини оламиз.

$$y: X \rightarrow C \quad (1)$$

Бу усул назорат билан ўқиш (supervised learning) деб аталади, чунки контролёр (синфларни аниқловчи ва ўқув ҳужжатларни тайёрловчи инсон) ўқиш жараёнини бошқарувчи ўқитувчи ролини ўйнайди. Ўқитувчи билан ўқиш усулини  $\Gamma$  ҳарфи билан белгилаймиз ва  $\Gamma(D) = y$  деб ёзамиз.  $\Gamma$  ўқиш усули киришига  $D$  тўплами олади ва  $y$  синфлаш функциясини қайтаради.  $y$  синфловчи ва  $\Gamma$  ўқитиш усулини кўпинча ажратишмайди.

Матнларни синфлашдан мақсад – матнли ёки янги маълумотларда юқори аниқликни таъминлаш, масалан, эрта оладиган тежамкор техникалар ҳақида хабарлар. Ўқув танловда юқори аниқликка эришиш етарлича осон (масалан, шунчаки белгиларни эслаб қолишимиз мумкин). Бироқ ўқув тўпламда эришилган юқори аниқлик синфловчи янги маълумотларда яхши ишлайди дегани эмас. Ўқув синфловчи учун ўқув тўпламдан фойдаланган ҳолда ўқув ва тест маълумотлари ўхшаш ёки бир тақсимланишга эга деб тахмин қилишимиз мумкин.

**Матнларни содда Байес усули орқали синфлаш.** Кўриб чиқадиган уқувли ўқитишнинг биринчи усули – Байеснинг мултиномиал содда усули (Naïve Baes - NB). Бу усулда  $d$  ҳужжат  $s$  синфга тегишлилик эҳтимоли қуйидагича ҳисобланади.

$$P(c|d) = P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \quad (2)$$

Бу ерда  $P(t_k|c)$  –  $t_k$  термин  $c$  синфдаги ҳужжатда пайдо бўлишининг шартли эҳтимоли.  $P(t_k|c)$  – бу  $t_k$  терминини ҳужжат  $c$  синфга тегишли эканлигининг баҳоси,  $P(c)$  эса – ҳужжат  $c$  синфга тегишли эканлигининг априор эҳтимоли. Агар ҳужжат термини бир синфни бошқасидан аниқ ажратиш имконини бермаса, унда улардан анча юқори априор эҳтимолга эга эканлигини танлаш керак. Синфлаш учун фойдаланилаётган  $\langle t_1, t_2, \dots, t_{n_d} \rangle$  кетма-кетлик лексиконнинг бир қисми ҳисобланган  $d$  ҳужжат лексемаларидан ташкил топган,  $n_d$  эса –  $d$  ҳужжатда бундай лексемаларнинг сони. Масалан, бир гапдан иборат бўлган *Бухоро ва Навоий вилоятлари чегарадош* сўров учун  $\langle t_1, t_2, \dots, t_{n_d} \rangle$  кетма-кетлик < Бухоро, Навоий, вилоят, чегарадош > кўринишига эга бўлиши мумкин, агар стоп-сўзни ўчириб ташланса, бу ерда  $n_d - 4$ .

Матнларни синфлашдан мақсад – ҳужжат учун энг яхши синфни топиш. ВМ усулда максимал апостериор эҳтимолга эга  $c_{map}$  синф (maximum a posteriori - MAP) ёки энг эҳтимоли катта синф энг яхши ҳисобланади.

$$c_{map} = \underset{c \in C}{argmax} \hat{P} = \underset{c \in C}{argmax} \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c) \quad (3)$$

$P$  эмас  $\hat{P}$  деб ёзамиз, чунки  $P(c)$  ва  $P(t_k|c)$  параметрларнинг ҳақиқий қийматларини билмаймиз, фақатгина уларни уқув тўпламлар ёрдамида баҳолашимиз мумкин.

(2) тенглик  $1 < k \leq n_d$  бир қиймат учун биттадан бир неча шартли эҳтимолларга кўпаяди. Бу аҳамиятли разрядлар йўқолишига олиб келиши мумкин. Мос равишда, эҳтимолларнинг кўпайтирилишини уларнинг логарифмлари кўшилиши билан алмаштириш яхшироқ. Эҳтимол логарифмининг қиймати кичик бўлган синф энг эҳтимолли бўлиб қолади, чунки  $\log(xy) = \log(x) + \log(y)$  ва логарифмик функция монотон ҳисобланади. Шундан келиб чиқиб, Байеснинг содда усулида аслида кейинги функциянинг максимум нуқтасини топиш талаб қилинади.

$$c_n = \underset{c \in C}{argmax} \left[ \log \hat{P}(c) + \sum_{1 \leq k \leq n_d} \log \hat{P}(t_k|c) \right] \quad (3)$$

(3) тенглик оддий интерполяцияга йўл қўяди.  $\log \hat{P}(t_k|c)$  шартли эҳтимолнинг ҳар бир логарифми – бу с синф учун  $t_k$  термин қанчалик муҳимлигини кўрсатувчи вазн. Худди шундай  $\log \hat{P}(c)$  априор эҳтимол с синфнинг нисбатан частотасини характерлайдиган вазн. Анча кўпроқ учрайдиган синфлар, кўпинча, кам учрайдиганларга қараганда анча тўғри ҳисобланади. Шундай қилиб, априор эҳтимоллар логарифми ва термин вазнларининг йиғиндиси ҳужжат синфга тегишли эканлигига гувоҳлик қилади, (3) тенглик эса гувоҳликларнинг энг кўп сонига мос келувчи синфни индентификациялайди.

$\hat{P}(c)$  ва  $\hat{P}(t_k|c)$  эҳтимолларни қандай баҳолаш мумкин? Олдин нисбатан частотани ўзида намоён этадиган ва берилган ўқув маълумотларда ҳар бир параметрнинг энг эҳтимолли катталигига мос келадиган максимал ҳақиқийлик баҳосини олишга уриниб кўрамиз. Априор эҳтимоллар учун баҳо қуйидаги кўринишга эга

$$\hat{P}(c) = \frac{N_c}{N} \quad (4)$$

Бу ерда  $N_c$  – с синфдаги ҳужжатлар сони,  $N$  эса – ҳужжатларнинг умумий сони.

$\hat{P}(t|c)$  шартли эҳтимолни с синфга тегишли ҳужжатларда  $t$  терминнинг нисбатан частотаси сифатида баҳолаймиз.

$$\hat{P}(t|c) = \frac{T_{at}}{\sum_{t' \in V} T_{at'}} \quad (5)$$

Бу ерда  $T_{at}$  ҳужжатда терминнинг кўп бора пайдо бўлишини ҳисобга олган ҳолда с синфдаги ўқув маълумотларда  $t$  терминнинг пайдо бўлиш сони. Бу баҳо позицион мустақиллик ҳақидаги тахминга (positional independence assumption) асосланган.  $T_{at}$  бу уқув тўплам ҳужжатларидаги барча  $k$  координаталарда терминнинг пайдо бўлиш сони. Шундай қилиб, турли координаталар учун турли баҳоларни ҳисоблаймиз, масалан, агар сўз ҳужжатда  $k_1$  ва  $k_2$  координаталарда икки марта учраса, унда  $\hat{P}(t_{k_1}|c)$  ва  $\hat{P}(t_{k_2}|c)$

Максимал ҳақиқийлик билан икки муаммо боғлиқ: агар “термин-синф” жуфтлиги ўқитиладиган ўқув маълумотларда учрамаса, унда MLE баҳо нолга тенг. Масалан, агар ўқув маълумотлардаги ШХТ термин фақат Ўзбекистон синф ҳужжатларида учраса, унда бошқа синфлар, масалан, АҚШ синф учун MLE баҳоси нолга тенг бўлади. Яъни,

$$\hat{P}(\text{ШХТ}|\text{АҚШ}) = 0$$

Энди бир гапдан иборат бўлган *Хитой ШХТининг аъзоси* ҳужжатнинг *China* синфга тегишли эканлигининг шартли белгиси нолга тенг бўлади, чунки (3) тенгликда барча терминлар учун шартли эҳтимолни кўпайтирамиз. Аёнки, модел *China* синфга катта эҳтимол бериши керак, чунки гапда *Хитой* термини учрамоқда. Шунга қарамасдан бошқа омиллар билан таъминланган, *China* синфи фойдасига қанчалик кўп гувоҳлик бўлмасин, ШХТ термини учун ноль эҳтимолини шунчаки ташлаб юбориш мумкин эмас. Бу баҳо терминнинг кам учраши сабабли нолга тенг.

Нолдан кутулиш учун ҳар бир частотага шунчаки бир қўшган ҳолда Лаплас силлиқлашидан (Laplace smoothing) фойдаланамиз.

$$\hat{P}(t|c) = \frac{T_{\alpha} + 1}{\sum_{t' \in V} (T_{\alpha} + 1)} = \frac{T_{\alpha} + 1}{\sum_{t' \in V} T_{\alpha} + B} \quad (6)$$

Бу ерда  $B = |V|$  - тўплам лексиконидаги терминлар сони. Лаплас силлиқлашидан кейин келадиган ўқув маълумотлар асосида аниқлаштириладиган априор раван тақсимланиш сифатида интерпретацияланиши мумкин (ҳар бир термин ҳар бир синфда бир мартадан учрайди). Шунини айтиб ўтиш керакки, бу – ҳужжат даражасида (4) формула билан баҳоланадиган синф эмас, термин пайдо бўлишининг априор эҳтимоли.

Энди ўқиш ва содда Байес синфловчисини қўллаш учун зарур бўлган барча элементларни киритамиз. Тўлиқ алгоритм қуйидагича бўлади.

### 2.1-алгоритм. Ўқиш ва содда Байес синфловчи учун мултиномиал алгоритми

```

TrainNB(C,D){
V←ExtractVacabulery(D)
N←CountDocs(D)
Foreach c∈ C do
    Nc←CountDocsInClass(D,c)
    priorc←Nc/N
    textc←ConcatenateTextOfAllDocsInClass(D,c)
    Foreach t∈ V do
        Tct←CountTkensOfTerm(textc,t)
    Foreach t∈ V do
        Condprob[t][c]← $\frac{T_c+1}{\sum(T_c+1)}$ 
Return V, prior, condprob
}
ApplyNB(C,V,priormcondprob,d){
W←ExtractTokensFromDoc(V,d)
Foreach c∈ C do
    score[c]←log prior[c]
Foreach t∈ W do
    Score[c]+=log condprob[t][c]
Return  $\underset{c \in C}{\operatorname{argmax}} \text{score}[c]$ 
}

```

**Бернулли модели.** Содда Байес моделини қуришнинг икки усули мавжуд. Мультиномиал модел юқорида таърифлаб ўтилди. У ҳужжат ичида ҳар бир координатада лексикондаги бир термин бўйича генерациялайди. Мультиномиал моделнинг устунлиги Бернуллининг кўп ўлчамли модели (multivariate Bernoulli model) ёки шунчаки Бернулли модели ҳисобланади. У лексиконнинг ҳар бир термини учун индекатор генерациялайдиган бинар мустақиллик моделига эквивалент, яъни, агар, термин ҳужжатда бўлса 1 ва агар, бўлмаса 0. Қуйида келтирилган алгоритмда Бернулли модели учун ўқиш ва тестлаш алгоритмлари кўрсатилган. Бернулли моделининг вақт мураккаблиги мультиномиал моделники каби ҳисобланади.

## 2.2-алгоритм. Содда Байес алгоритми (Бернулли модели): ўқиш ва тестлаш

```
TBernoulliNB(C,D){
  V←ExtractVocabulary(D)
  N←countDocs(D)
  foreach c ∈ C do
    Nc←CountDocsInClass(D,c)
    Prior[c] ←Nc /N
  foreach t∈ V do
    Nct←CountDocsInClassContaininTerm(D, c, t)
    Condprob[t][c]←(Nct+1)/(Nct+2)
  return V, prior, condprop }
ApplyBernouLLinb (C, V, prior, Condprob. D){
  Vd←EtractTermsFromDoc(V,d)
  foreach c ∈ C do
    Score[c]←log prior[c]
  foreach t ∈ V do
    if t ∈ V then
      score[c]+=log condprob[t][c]
    else
      score[c]+=log(1-condprob[t][c])
  return arg max score[c]
      c∈C
}
```

2.2-алгоритмнинг 8-қаторида (5) формулага ўхшаб, бирни қўшган ҳолда силлиқлаш қўлланилган, у ерда  $V = 2$ .

Турли моделлар баҳонинг турли стратегиялари ва синфлашнинг турли қоидаларини кўзда тутати. Бернулли моделида  $\hat{P}(t|c)$  эҳтимол  $t$  терминга эга  $c$  синфлардан хужжатлар ҳиссаси сифатида баҳоланади (2.2- алгоритмнинг 8-қатори). Унга қарама-қарши тарзда мультиномиал моделда  $\hat{P}(t|c)$  эҳтимол лексемлар ҳиссаси, ёки  $t$  терминга эга  $c$  синф хужжатларида сўзларнинг ҳиссаси сифатида баҳоланади ((5) тенглик). Бернулли модели асосида тест хужжатни синфлаштиришда бу терминнинг кириш сонини эътиборга олмайдиган термин пайдо бўлиши ҳақидаги бинар ахборотдан фойдаланади, бу вақтда мультиномиал модел хужжатда терминнинг кўп бор ишлатилиши кузатилади. Натижада хужжат маълумотларини синфлашда Бернулли модели, қоидага кўра, кўп хатоликларга йўл қўяди. Масалан, у Чина гуруҳига чина терминининг бир марта ишлатилиши эвазига бутун китобни киритиши мумкин.

Бу моделлар шунингдек, хужжатда пайдо бўлмайдиган терминлар қандай ишлатилишига қараб фарқланади. Мультиномиал моделда бу ахборот ечимга ҳеч қандай таъсир қилмайди, Бернулли моделида эса  $\hat{P}(t|c)$  эҳтимолни ҳисоблашда терминнинг йўқлик эҳтимоли ҳисобга олинади. Бу фақат Бернулли модели термин йўқлиги ҳақидаги ахборотни аён равишда эътиборга олиши билан тушунтирилади.

### Фойдаланилган адабиётлар

1. Шокин Ю.И. Проблемы поиска информации / Ю.И. Шокин, А.М. Федотов, В.Б. Барахнин. Новосибирск: Наука, 2010. – 220 с. ISBN 918-5-02-018969-0
2. Мўминов Б.Б. Маълумотларни излаш тизими. –Т.: Фан ва технология. 2016. -210 б.
3. Мўминов Б.Б. Маълумотларни излаш усуллари. –Т.: Фан ва технология. 2016. -276 б.