УДК 025.4.03
МРНТИ 20.23.21

ЖАСЫРЫН ДИРИХЛЕ ҮЛЕСТІРІЛІМІН МӘТІН ҚҰЖАТТАРЫН СЕГМЕНТТЕУ ҮШІН ҚОЛДАНУДЫ ЖҮЗЕГЕ АСЫРУ

М.Е. МАНСУРОВА, Ә.Ғ. ОСПАН, МАХМУТ ЕРЛАН

әл-Фараби атындағы Қазақ Ұлттық университеті

Аңдатпа: *Бұл жұмыста белгілі TextTiling алгоритмінде негізделген, мәтіндік құжаттарды сегменттеудің алгоритмінің жүзеге асуы көрсетілген. Сегменттеудің жұмысы жасырын Дирихле үлестірілуі (LDA) тақырыптық моделінің қолданылуымен орындалады. Сегменттелудің алгоритмі сызықты уақытта іске асқан, яғни, ол LDA-де негізделген сегменттелудің басқа әдістерімен салыстырғанда арзанырақ болып келеді. TextTiling алгоритмінен айырмашылығы, берілген алгоритмде сөздер қарастырылмайды. Бұл жерде LDA–дің байестік шешім шығару әдісімен көрсетілген, тақырыптар идентификаторлары қарастырылған. Тақырыптық айырмашылықтарды анықтау үшін, мазмұны бойынша тексерілетін құжаттарға ұқсас келетін құжаттардағы тақырыптық модельдеріне оқыту жүргізіледі. Орындалған сегменттелудің сапасын бағалау үшін перплексияның ішкі критерийі қолданылады. Құрылған алгоритм Википедиядағы 40 құжатты сегменттеуге пайдаланылды.*

Түйінді сөздер: *Жасырын Дирихле үлестірілімі, TextTiling алгоритмі, мәтін сегменттелуі, Гиббс бойынша үлгі жию, перплексия*

REALIZATION OF THE USE OF LATENT DIRICHLET ALLOCATION FOR SEGMENTATION OF TEXT DOCUMENTS

Abstract: *In this paper, we present an implementation of the text segmentation algorithm based on the TextTiling algorithm. The segmentation problem is carried out using the Latent Dirichlet Allocation topic model (LDA). The presented segmentation algorithm is linear in time, therefore, less expensive, compared to other LDA-based segmentation methods. In contrast to the TextTiling algorithm, this algorithm is based not on words, but on the topic identifiers specified by the Bayesian LDA method. To identify topic differences, the topic model is trained on documents that are similar in content to test documents. To assess the quality of the segmentation, an internal perplexity criterion is used. The developed algorithm is applied to segmentation of 40 documents from Wikipedia.*

Key words: *latent Dirichlet allocation, algorithm TextTiling, text segmentation, Gibs sampling, perplexity*

РЕАЛИЗАЦИЯ ИСПОЛЬЗОВАНИЯ ЛАТЕНТНОГО ДИРИХЛЕ РАСПРЕДЕЛЕНИЯ ДЛЯ СЕГМЕНТАЦИИ ТЕКСТОВЫХ ДОКУМЕНТОВ

Аннотация: *В этой работе представлена реализация алгоритма сегментации текстовых документов, основанного на хорошо известном алгоритме TextTiling. Задача сегментации выполняется с использованием тематической модели латентного размещения Дирихле (LDA). Алгоритм сегментации реализован за линейное время, следовательно, является менее дорогостоящим по сравнению с другими методами сегментации на основе LDA. В отличие от алгоритма TextTiling, в данном алгоритме рассматриваются не слова, а идентификаторы тем, заданные методом байесовского вывода LDA. Для выявления тематических различий проводится обучение тематической модели на документах, аналогичных по содержанию тестовым документам. Для оценки качества выполненной сегментации используется внутренний критерий перплексии. Разработанный алгоритм применен для сегментации 40 документов из Википедии.*

Ключевые слова: *Латентное Дирихле распределение, алгоритм TextTiling, сегментация текста, семплирование по Гиббсу, перплексия*

Кіріспе

Бұл мақалада қарастырылған тапсырма – ол мәтіндердің тақырыптық ұқсас бірліктерге жіктелуінен тұратын мәтіндердің тақырыптық сегменттелуі. Мәтінді сегменттеудің мәні мәтіннің тақырыпшалар құрылымын анықтаудан тұрады. Берілген тапсырма келесі бірнеше тәсілдер арқылы шешілуі мүмкін:

- мәтін ішіндегі лексикалық байланыстар талдауында негізделген тәсіл (терминдер үлестірілуі, көршілес сөйлемдер байланысы және т.б.) [1-3];
- мәтіндік құжаттар коллекциясындағы терминдердің жиілік сипаттамаларын ескеретін статистикалық тәсіл [4, 5];
- мәтіндік құжаттар коллекциясындағы ықтималды тақырыптық модельдерде негізделген ықтималды тәсіл [6].

Тақырыптық сегменттелу келесідей кіші салаларға бөлінеді:

- тақырыптық өзгерулердің жүйелі талдауларымен байланысты сызықтық сегменттелу;
- мәтін ішіндегі майда тақырыпшаларды іздеумен байланысты иерархиялық сегменттелу.

Ұстазсыз тақырыптық сегменттелудің ең алғашқы сызықты алгоритмі TextTilling [1] алгоритмі болған. Берілген алгоритмнің негізгі идеясы болып, қолданылатын лексикон жылдам өзгеретін мәтіндегі нүктелерді іздеу болып табылады. Бұл үшін мәтін алдымен *терезе* деп аталатын терминдердің кішігірім жүйелеріне бөлінеді. Сосын әрбір шекара нүктесінде терезелердің m –нен қатар-қатар солға және оңға жүретін екі тізбек арасындағы косинустық ұқсастық шамасы есептелінеді. Егер қандай да бір нүктеде көршілес нүктелермен салыстырғанда шаманың мәні біраз кемісе, бұл осы нүктеде сегменттелу жүргізу керек дегенді білдіреді. TextTilling – та негізделген LcSeg [6] алгоритмінде tf-idf терминдерінің салмақты коэффициенттері қолданылған, бұл өз кезегінде тақырыптық сегменттелудің нәтижесінің едәуір жақсаруына алып келді. Dotplotting [2] – бұл мәтін жазықтықта нүктелердің геометриялық орны ретінде ұсынылатын тақырыптық сегменттелудің графикалық әдісі. Жазықтықта мәтіннің геометриялық таңбалары орындалғаннан кейін тақырыптық сегменттерді анықтайтын «ұйымалар» ерекшеленеді. Link Set Median [3] алгоритмі сөйлемдер арасындағы тікелей лексикалық байланыстар талдауында негізделген. [7]–де терминдерді кластерлеуде және сегменттелу кезеңінде алынған кластерлерді пайдалануда негізделген тақырыпшаларды ерекшелеу әдісі ұсынылған.

Жасырын айнымалыларымен тақырыптық модельдер мәтіндік коллекцияларда жасырын құрылымдарды анықтау үшін тиімді болып шықты. Негізгі тақырыптық модельдерге келесілер жатады:

- Ықтималды латентті семантикалық талдау (probabilistic latent semantic analysis, PLSA) [8], автоматты құжаттарды индекстеу талдауының статистикалық моделі;
- Латентті Дирихле үлестірілуі моделі (latent Dirichlet allocation, LDA) [9] әрбір құжат әртүрлі тақырыптар қоспасы ретінде қарастырылатын тудырушы модель.

Берілген жұмыста [10]-шы жұмыста ұсынылған мәтіннің сызықты сегменттелуі үшін ықтималды әдіс жүзеге асқан. TopicTiling деп аталатын берілген алгоритм TextTiling алгоритмінде негізделген және тақырыптық LDA моделін қолданады. Оның ішінде сөздердің орнына LDA қорытындысы кезінде алынған, тақырыптар идентификаторы қарастырылады. Бұл идентификаторларды тақырыптың әрбір идентификаторының жиілігінен тұратын, екі вектор түрінде ұсынылған, сөйлемдердің екі көршілес блоктарының арасындағы косинустық ұқсастықтарды есептеу үшін қолданады.

[11]-де көрсетілгендей, тақырыптық сегменттелудің сапасын бағалау айтарлықтай күрделі проблема болады. Орташа ішкі кластерлер және кластерлер арасындағы қашықтықтар сияқты кластерлердің сапасының стандартты критерийлері құжаттар мен терминдердің бірлескен кластерлерін бағалау үшін нашар сәйкес келеді. Тақырыптық мо-

дельдердің сапа критерийі екі негізгі топқа бөлінеді: ішкі және сыртқы. Берілген жұмыста тақырыптық сегменттелудің сапасын бағалауға құжаттар коллекциясы моделінің сапасын бағалайтын, ішкі критерий – перплексия қолданылды. Перплексия шындық үйлестік логарифмі арқылы анықталатын, D коллекциясындағы d құжаттарында бақыланатын p(w|d) моделінің w терминдеріне сәйкес келмеуін өлшейтін шама.

Мәтіндерді сегменттеудің әдісі

Мәтіндік құжаттардан білім алудың негізгі тәсілдерінің бірі – ол тақырыптық модельдеу, яғни, әрбір құжат қандай тақырыпқа жататынын анықтайтын мәтіндік құжаттардың коллекциясы моделінің құрылу тәсілі [1]. Ықтималды тақырыптық модель әрбір тақырыпты терминдер жиымында дискретті үлестірілумен сипаттайды, ал әрбір құжатты – тақырыптар жиымында дискретті үлестірілумен сипаттайды. Алдын-ала құжаттар коллекциясы кездейсоқ таңдалған және осы сияқты үлестірілулер жиымынан тәуелсіз терминдердің тізбегі деп болжамдайды. Тақырыптық модельдеудің мақсаты – таңдау бойынша жиымның құрамдас бірліктерін қайта құру.

Берілген мақалада тақырыптық модельдеудің ішіндегі бір моделін қолданған кездегі нәтижелері сипатталған, атап айтатын болсақ, мәтіндік құжаттарды сегменттеу үшін LDA моделі. Құжаттар коллекциясының ықтималды моделі. D – мәтіндік құжаттардың коллекциясы, W – соның ішінде қолданылатын барлық терминдер сөздігі (сөз немесе сөз тіркестері). d∈D –ның әрбір құжаты W сөздігіндегі $(w_1, \ldots, w_{nd})$ тізбегінің n_d терминдерін білдіреді. Термин құжатта көп рет қайталануы мүмкін.

Ақырғы T тақырыптар жиымы бар деп және w терминін әрбір d құжатында әр кезде қолданылуы қандай-да бір белгісіз болатын t ∈T тақырыбымен байланысты. Құжаттар коллекциясы кездейсоқ таңдалған және D×W×T ақырғы жиымында берілген, p(d,w, t) дискретті үлестірілуінен тәуелсіз, (d, w, t) үштік жиымы ретінде қарастырылады. d∈D құжаттары және w∈W терминдері бақыланатын айнымалылар болады, t ∈T тақырыбы жасырын (латентті) айнымалы болады.

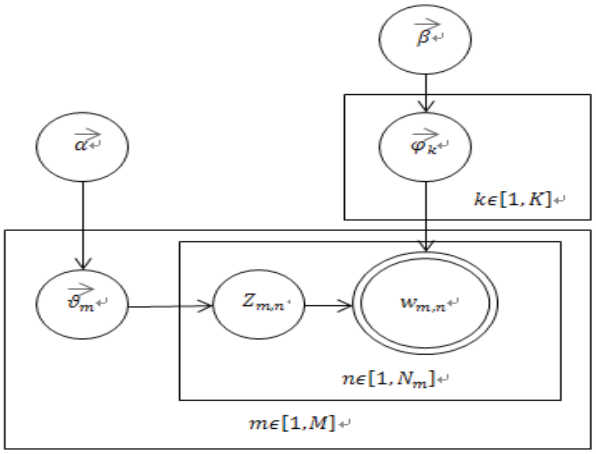
D құжаттар коллекциясының тақырыптық моделін құру – ол барлық t ∈T тақырыптары үшін p(w|t) үлестірілулерін табу және барлық d ∈D құжаттары үшін p(t|d) үлестірілулерін табу дегенді білдіреді. Табылған үлестірілулер сосын қосымша тапсырмаларды шешуге қолданылады. p(t|d) үлестірілуі ақпаратты іздеу, құжаттарды классификациялау мен категорияларға жіктеу тапсырмаларындағы құжаттардың ыңғайлы белгі сипаттамасы болып табылады.

Жасырын (латентті) айнымалылары бар модельдер ғылыми мақалалардағы және жаңалықтар ағымындағы трендтерді анықтау үшін [3], ұсыну жүйелеріндегі [5] және басқа да қосымшалардағы құжаттарды классификациялау мен категорияларға жіктеу [4] үшін қолданылатын мәтін коллекцияларының [2] жасырын құрылымын анықтау үшін ерекше тиімді болып шықты.

Жасырын Дирихле үлестірілімі моделі

Жасырын Дирихле үлестірілімі тақырыптық модель ретінде, мәтіндер жиынтығындағы (corpus) әрбір құжаттың тақырыптарын ықтималды үлестірілу ретінде көрсетіп береді. Ол құжаттарға талдау жасау арқылы олардың тақырыптарын (topic) суырып алып шығады және осы тақырыптарға негізделе отырып, тақырыптарды топтарға бөледі немесе мәтіндерді түрлерге бөлу қызметін жүргізеді. Жалпы алғанда жасырын Дирихле үлестіруі бір құжаттың қалай қалыптасқандығын бейнелейді.

Бастапқыда, біраз жасырын айнымалыларды жорамалдап алады, бұл айнымалылар біз айтып отырған тақырыптар, z әрпімен өрнектеледі. Әдетте 1-суретте көрсетілгендей әрбір мәтін тақырыптардың үлестіруіне сәйкес келеді, ал әрбір тақырып ішіндегі сөздерде белгілі бір үлестіруге сәйкес келеді.



LDA - ықтималдық график моделы

1 – сурет. LDA-ықтималдық графикалық моделі

LDA моделінде, бір құжаттың қалыптасу барысы төмендегідей жүргізіледі:

1. Жасырын Дирихле үлестірілімі $\alpha\alpha$– дан үлгі ала отырып і-інші құжаттың тақырыптық үлестіруі $\theta_i\theta_i$ –ды құрайды.

2. Тақырыптық көпмүшелі үлестіруі $\theta_i\theta_i$ – дан үлгі ала отырып і-інші құжат j-інші сөзінің тақырыбы $z_{i,j}z_{i,j}$–ды құрайды.

3. Жасырын Дирихле үлестірілімі $\beta\beta$–дан үлгі ала отырып $z_{i,j}z_{i,j}$ –дың сөздер үлестіруі $\phi_{z_{i,j}}\phi_{z_{i,j}}$ –ды құрайды.

4. Сөздердің көпмүшелі үлестіруі $\phi_{z_{i,j}}$ $\phi_{z_{i,j}}$ –нен үлгі ала отырып ең соңғы нәтижесі $w_{i,j}w_{i,j}$ –ды құрайды.

Берілген мәтіндер жиынындa$w_{m,n}w_{m,n}$ бақылауға болатын берілген айнымалы, $\vec{\alpha}$ мен $\vec{\beta}\vec{\alpha}$ мен $\vec{\beta}$күрделенген параметрлер деп аталады да, екеуінің мәні тәжірибеге негізделе отырып беріледі. Ал басқа$z_{m,n}z_{m,n}$, $\vec{\vartheta}_m\vec{\vartheta}_m$ және$\vec{\varphi}_k\vec{\varphi}_k$ қатарлы айнымалылар белгісіз жасырын айнымалылар болып, олардың мәнін бақылауға болатын айнымалылар арқылы оқып бағалайды.

Жасырын Дирихле үлестіруі моделіндегі барлық көрінетін және көрінбейтін айнымалылардың бірлесіп үлестірілуі формула (1) түрінде өрнектеледі.

мұнда: M- Мәтіндер жиыны, K- тақырып саны,N_mN_m-m-інші мәтіндегі сөздердің жал-

$$P(\vec{w}_m\,\vec{z}_m,,\vec{\vartheta}_m,\Phi|\vec{\alpha},\vec{\beta}) = \prod_{n=1}^{N_m} P(w_{m,n}|\vec{\varphi}_{z_{m,n}})P(z_{m,n}|\vec{\vartheta}_m)\;P(\vec{\vartheta}_m|\vec{\alpha})P(\Phi|\vec{\beta}), \qquad (1)$$

пы саны,$\vec{\alpha}$ –$\vec{\alpha}$ – әрбір мәтіндегі тақырыптардың көп мүшелі үлестіруінің Dirichlet априорлы параметрі; $\vec{\beta}$ –$\vec{\beta}$ – әрбір тақырыптағы сөздердің көп мүшелі үлестіруінің Dirichlet априорлы параметрі; $z_{m,n}z_{m,n}$- m-інші мәтіндегі n-інші сөздің тақырыбы,$w_{m,n}w_{m,n}$- m-інші мәтіндегі n-інші сөз; $\theta_m - \theta_m$ – m-інші мәтіндегі тақырып үлестіруі; $\varphi_k\varphi_k$: k-інші тақырыптағы сөздің үлестіруі.

Онда белгілі $w_{m,n}$ $w_{m,n}$ сөзінің мәтіндер қорындағы ықтималдығы формула (2) түрінде өрнектеледі.

Түсінікті болу үшін, жоғарыдағы формуланы келесідей түрлендіріп өрнектеуге болады. Мысалы, T сандағы тақырып бар деп

$$p(w_{m,n} = t|\vec{\vartheta}_m,\Phi) = \sum_{k=1}^{K} p(w_{m,n} = t|\vec{\varphi}_k)p(z_{m,n} = k|\vec{\vartheta}_m), \qquad (2)$$

қарасақ, онда берілген мәтіндегі і-сөз w_iw_i -дың ықтималдығы формула (3) түрінде көрсетілген.

мұнда, Z_iZ_i-жасырын айнымалы немесе табылатын тақырыптар, і-нөмірленген w_iw_i -сөзі осы тақырыптардан алынған дегенді

$$P(w_i) = \sum_{j=1}^{T} P(w_i|Z_i = j)P(Z_i = j), \quad (3)$$

білдіреді; $P(w_i|Z_i = j)P(w_i|Z_i = j)$-нөмірленген сөз$w_iw_i$-дың тақырып j-ге тәуелділігінің ықтималдық шамасы. T сандағы тақырып W-сандағы талданған барлық сөздерді қамтитын D сандағы мәтінді құрайтын болсын деп қарайық. Белгілеуге қолайлы болу үшін, $\varphi_w^{z=j}(z = j) = P(w|z = j)$ $\varphi_w^{z=j}(z = j) = P(w|z = j)$-ды тақырып j-ге қарата W сандағы сөздердегі көп мүшелі үлестіру ретінде қарайық. Мұнда, w-болса W сандағы сараланған барлық сөздердің бірі; $\vartheta_{z=j}^{(d)} = P(z = j)\vartheta_{z=j}^{(d)} = P(z = j)$-ды мәтін d-ге қарата айтқанда, саны T болған тақырыптардағы көп мүшелі үлестіру ретінде қарайық. Сонымен мәтін d-дегі сөздер w-дың ықтималдығы формула (4) түрінде болады.

$$P(w|d) = \sum_{j=1}^{T} \varphi_w^{z=j} * \vartheta_{z=j}^{d} (4)$$

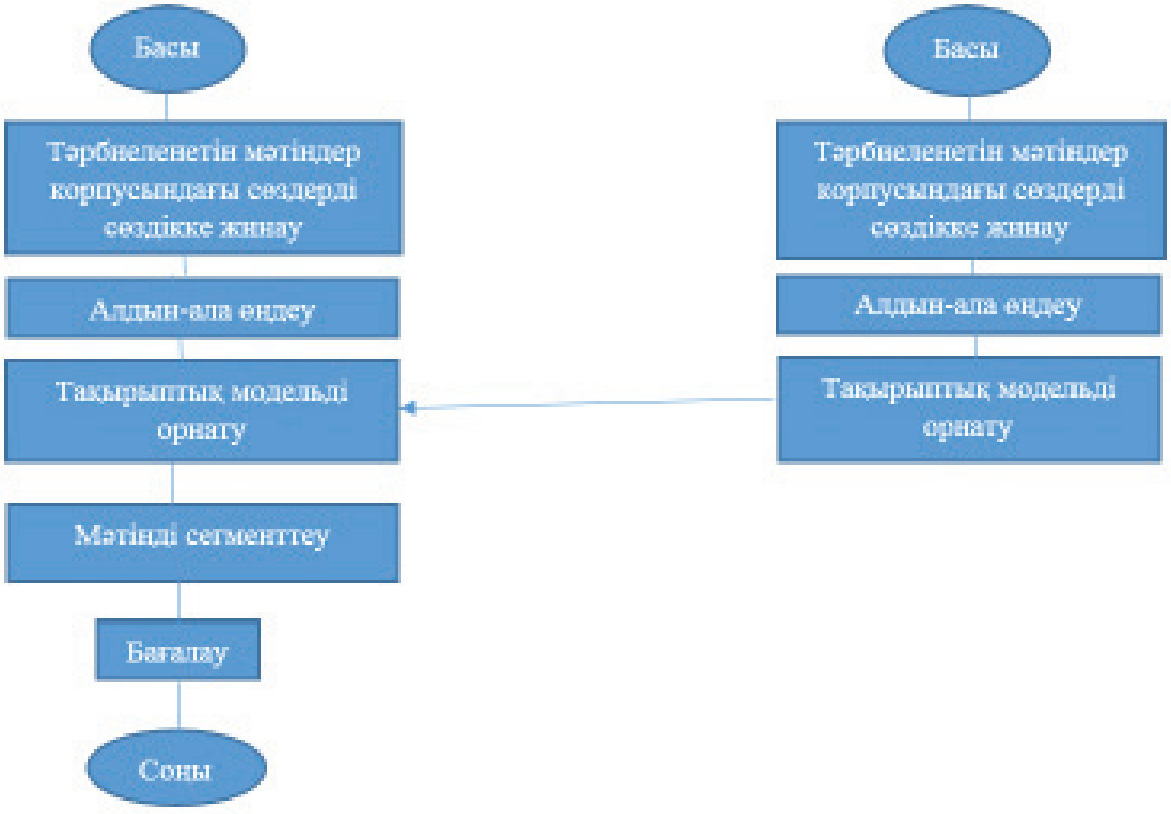
Жасырын Дирихле үлестіріміне негіз делген сегменттелу алгоритмі

Тақырыптық модельді пайдаланып жасырын Дирихле үлестіріміне негізделген мәтінді сегменттеудің алгоритмі төменгі сұлбада көрсетілген:

Жасырын Дирихле үлестірімі арқылы мәтіндерді сегменттеу үшін алдымен келесідей қадамдарды орындау қажет:

1. Бірінші қадамда – деректер қорын алдын ала өңдеп, қайталанатын сөздерді шығарып тастап, мағыналары сәйкес сөздерді векторлық кеңістікте, яғни сөз–индекс векторы және индекс-сөз векторына орналастырып аламыз.

2. Екінші қадамда – үш матрица құрамыз, сосын LDA–моделін тұрғызамыз, яғни деректер қорындағы сөздердің қай тақырыпқа тәуелді екенін білдіретін матрицаны Z(m,n) арқылы өрнектейік: m-деректер базасындағы мәтін саны, n-әр мәтіндегі сөздердің саны; мәтін-тақырып матрицаны nmk(m,k)



1-сұлба. Мәтінді сегменттеу алгоритмі

арқылы өрнектейік: K- тақырып санын біл-діреді; сондай-ақ, тақырып – сөз матрица-сы nkt(k,t) арқылы өрнектейік: t-тақырыпта туылған сөздерді білдіреді.

3. Үшінші қадамда, деректер қорының LDA–моделін тұрғызамыз, сондай-ақ, осы модельдің көмегімен сегменттелетін мәтін-дердің LDA–моделін құрып шығамыз, яғни екі үлестірудің мәнін есептеп аламыз. Оны есептеу үшін жеке-жеке мәтін-тақырып үлестіруі ϑ-ді (5) қолданамыз.

4.

$$\vartheta_{z=j}^{d} = \frac{n_m^{(j)}}{\sum_{j=1}^{K} n_m^{(j)} + \alpha_j} \qquad (5)$$

Осы қадамда деректер қорының LDA–моделін тұрғызамыз, сондай-ақ, осы модельдің көмегімен сегменттелетін мәтіндердің LDA–моделін құрып шығамыз. Нақтылап айтар болсақ, алдымен Гиббс бойынша үлгі жию алгоритмін пайдаланып, деректер қорының LDA–моделін құрамыз, яғни екі үлестірудің мәнін есептеп аламыз. Оны есептеу үшін жеке-жеке тақырып-сөз үлестіруі φ-ді (6) қолданамыз.

$$\varphi_t^{z=j} = \frac{n_j^{(t)} + \beta_t}{\sum_{t=1}^{V} n_j^{(t)} + \beta_t} \qquad (6)$$

Төртінші қадамда, сегменттелетін мәтіндердің LDA–моделінде есептелген φ (5) мен ϑ-дің (6) мәнін пайдаланып, берілген тақырыпқа сәйкесті мазмұндарды анықтаймыз.

5. Бесінші қадамда, сегменттелетін мәтіндерді сөйлемдерге ажыратып, әрбір сөйлемді бір мәтін ретінде қарап, келесі әрбір сөйлемдерге LDA–моделін тұрғызып, сәйкесті φ(3) мен ϑ-дің (4) мәнін есептеп аламыз.

6. Алтыншы қадамда, әрбір сөйлемдегі сөздердің ықтималдығын (7) формула бойынша есептеп шығарамыз.

$$p(s) = \sum_{k=1}^{K} \varphi_w^{(z=k)} * \vartheta_{(z=k)}^{s} \qquad (7)$$

7. Жетінші қадамда, белгілі жуықтықты есептеу арқылы сөйлемдер арасында сақталған жуықтық мөлшерін косинусоидалық өлшеу (8) тәсілі арқылы есептеп шығарамыз:

$$Sim_{cos} = \frac{\sum_{\omega \in W} P(\omega|d_1)P(\omega|d_2)}{\sqrt{\sum_{\omega \in W} P(\omega|d_1)^2} * \sqrt{\sum_{\omega = W} P(\omega|d_2)^2}} \qquad (8)$$

8. Соңында, белгілі шекараны тану алгоритмін пайдаланып, осы сөйлемдердің шекарасын ажыратамыз. Мұнда біз динамикалық тұрақты сандар алгоритмін (9) пайдаландық. Динамикалық тұрақты сандар формуласын төменде көрсеттік:

SimTab={Sim_1, Sim_2, $Sim_{3,\dots,}$ Sim_{n-1}} ;

мұнда, Sim_iSim_i=Sim($d_i, d_{i+1}d_i, d_{i+1}$) ,
1≤i≤n-1

avgSim=Sim1+Sim2+…+Simn-1n-1 ,
1≤i≤n-1; (9)

(9) формуланы келесі түрге түрлендіреміз:

$$avgSim = \frac{(Sim_2 - Sim_1) + \cdots + (Sim_{n-1} + Sim_{n-2})}{n-2}$$

Бұл жағдайда, егер avgmSim≤ ≤Sim(d_1,d_2)≤avgSim болса, онда d_1 мен d_2 ұқсамаған бөліктерге тәуелді болады деп қарайміз.

Мәтіннің тақырып санын анықтау үшін перплекстік шаманы қолдану

Жасырын Дирихле үлестірімінің өнімділігі тақырып санының әсеріне ұшырайды, сондықтан ең алдымен тақырыптардың саны K-ны тұрақтандырып алу керек. Тақырыптардың санын тұрақтандыратын әдістердің түрлері өте көп, мәселен параметрленбеген тақырыптық модель (hierarchical dirichlet process[10]) әдісін пайдалану, иерархиядық топтастырып ажырату әдісін [8] қолдану және модельдің перплекстік шамасын талдау әдісін [1] пайдалану және т.б. Бұл жұмыста модельдің перплекстік шамасын талдау әдісінен пайдаланып тақырыптардың санын тұрақтандыруды негіз етті.

Бұл шындық үйлестік логарифмі арқылы анықталатын, D коллекциясындағы d құжаттарында бақыланатын, w терминдеріне p(w|d) моделінің сәйкес келмеуі немесе «алшақтық» шамасы. Осы шама неғұрлым аз болған сайын, соғұрлым p моделі D коллекциясындағы

d құжаттарында w терминдерінің пайда болуын жақсы болжайды [1].

$$\text{perplexity(D)} = \exp\left\{\frac{\sum_{s=1}^{N} lgp(s)}{\sum_{s=1}^{N} N_s}\right\}$$

$$\text{perplexity(D)} = \exp\left\{\frac{\sum_{s=1}^{N} lgp(s)}{\sum_{s=1}^{N} N_s}\right\} (10)$$

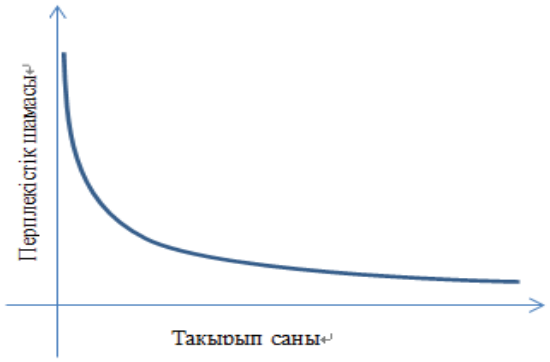
Мұнда, N-мәтіндер жиынындағы сөйлемдердің саны; N_s- сөйлем S- тегі сөздердің саны; p(s)- сөйлем S- тің жуықтық шамасы; Жасырын Дирихле Үлестіруі моделінде, сөйлемдердің жуықтық шамасы – сөйлемнің тақырыптық үлестіруі мен тақырыптың сөздерге үлестіруі арқылы есептелінеді, оны формула (11) де өрнектеліп көрсеттік.

$$\lg\left(p(s)\right) = \sum_{n=1}^{N} n(w,s)\,\varphi_w^{(z=j)} * \vartheta_{z=j}^{(s)}$$

$$\lg\left(p(s)\right) = \sum_{n=1}^{N} n(w,s)\,\varphi_w^{(z=j)} * \vartheta_{z=j}^{(s)} (11)$$

мұнда, n(w,s) болса, сөйлем S- тегі сөздердің туылу санын білдіреді.

Қорытып айтқанда, модельдің перплекстік шамасы сөйлемдердің жуықтық шамасының жоғарылауына байланысты кемиді.



2 – сурет. Перплекстік шаманың мәні тақырып санының өзгерісіне байланысты өзгеру схемасы

Деректерді оқыту

Жасырын Дирихле үлестірімі моделінде, тақырыптарды тауып шығу, яғни тақырыптардың сәйкесті ықтималдығын есептеу деректер базасын оқытудың негізінде жүргізіледі. Модельді оқыту жалпы алғанда екі түрлі үлестіруді бағалайды: тақырып – сөз үлестіруі $\varphi_t^{z=j}\varphi_t^{z=j}$ мен құжат-тақырып үлестіруі $\vartheta_{z=j}^{d}\vartheta_{z=j}^{d}$.

$\varphi_t^{z=j}$ мен $\vartheta_{z=j}^{d}\varphi_t^{z=j}$ мен $\vartheta_{z=j}^{d}$-ның мәнін есептеп шығару үш түрлі матрица құру керек, олар жеке-жеке тақырып-құжат матрицасы $n_m^{(j)}n_m^{(j)}$, тақырып – сөз матрицасы $n_j^{(t)}n_j^{(t)}$ және құжат-сөз матрицасы $z_m^n z_m^n$.

Тақырып-құжат матрицасы $n_m^{(j)}n_m^{(j)}$- берілген құжат m-дегі тақырып j-дың туылу санын есептейді; тақырып – сөз матрицасы $n_j^{(t)}n_j^{(t)}$- берілген тақырып j-ге байланысты сөз t-дың туылу санын есептейді, ал құжат-сөз матрицасы $z_m^n z_m^n$- мәтіндегі әрбір сөздің қай тақырыпқа тәуелді екенін есептейді.

Жасырын Дирихле үлестірімі моделін оқытудан бұрын, әсіресе Гиббс бойынша үлгі жию алгоритмінің сәтті орындалуына кепілдік ету үшін $n_m^{(j)}n_m^{(j)}$, $n_j^{(t)}n_j^{(t)}$ және $z_m^n z_m^n$ - дердің алғашқы мәндері есептелініп алынады. Содан кейін, осы мәндердің негізінде $n_m^{(j)}$ $n_m^{(j)}$, $n_j^{(t)}n_j^{(t)}$ және $z_m^n z_m^n$- дердің жаңа мәндері есептелінеді, соңында $\varphi_t^{z=j}$ мен $\vartheta_{z=j}^{d}$ $\varphi_t^{z=j}$ мен $\vartheta_{z=j}^{d}$- мәні шығарылады. Бұл барыс көп рет қайталанады, яғни $\varphi_t^{z=j}$ мен $\vartheta_{z=j}^{d}$ $\varphi_t^{z=j}$ мен $\vartheta_{z=j}^{d}$- дың мәні қанықканға дейін.

Қорыта келгенде, кез келген жаңадан енген сегменттеуге тиісті мәтіндердің тақырыптық моделін оқыту барысы дербес жүргізіледі, оның оқытылған тақырып моделінің соңғы нәтижесі деректер базасының оқытылған тақырып моделінің негізінде есептелінеді, сегменттелетін мәтіннің тақырып моделінің соңғы нәтижесі φ мен $\vartheta\varphi$ мен ϑ- келесідей есептелінеді:

мұнда, $trnModel.n_k^{(t)}trnModel.n_k^{(t)}$- оқылатын деректер базасының тақырып-сөз матрицасы; $segModel.n_k^{(t)}segModel.n_k^{(t)}$

$$\varphi_k^{(t)} = \frac{trnModel.n_k^{(t)} + segModel.n_k^{(t)} + segModel.\beta}{\sum_{i=1}^{V} trnModel.n_k^{(t)} + \sum_{i=1}^{V} segModel.n_k^{(t)} + trmModel.V * segModel.\beta} (12)$$

$$\vartheta_m^{(k)} = \frac{segModel.n_m^{(k)} + segModel.\alpha}{\sum_{k=1}^{K} segModel.n_m^{(k)} + segModel.\alpha} \quad (13)$$

- сегменттелетін мәтіннің тақырып-сөз матрицасы; $segModel.\beta$ $-segModel.\beta$ – сегменттелетін мәтіннің күрделенген параметрі; $trmModel.V trmModel.V$- оқылатын деректер базасындағы барлық сөздер; $segModel.\alpha$ $segModel.\alpha$- сегменттелетін мәтіннің күрделенген параметрі;

Жоғырыдағы формулада, белгісіз айнымалы $\vartheta_m^{(k)}\vartheta_m^{(k)}$- дың есептеу жолы, оқылатын деректер базасындағы $\vartheta_m^{(k)}$ $\vartheta_m^{(k)}$- дың есептеу жолымен негізінен ұқсас. Бірақ, $\varphi_k^{(t)}\varphi_k^{(t)}$- дың есептелу барысында сәл айырмашылдық бар, яғни сегменттелетін мәтіндердегі $\varphi_k^{(t)}\varphi_k^{(t)}$ –дың соңғы мәні оқылатын деректер базасындағы тақырып-сөз матрицасының соңғы нәтижесі мен деректер қорындағы барлық сөздердің санын параметр етіп пайдаланады.

Тәжірибе барысы

Тәжірибе барысында тек ағылшын тілінде жазылған мәтіндерге жүгіндік. Тәжірибенің тепе-теңдік қасиетіне кепілдік ету үшін, әртүрлі мазмұнды қамтыған мәтіндерге жүргізілді. Мәтіндер Википедия сайтынан алынды. Атап айтқанда, мәтіндерді сегменттеу технологиясы, имигранттар (immigration) саясаты, IBM компаниясы, үй жануарлары, құстар туралы мазмұндарға сегменттеу жүргізілді. Бұл жерде мәтіндерді шексіз алуға болады. Нәтижесінде бізде алдымен осы барлық мәтін ішіндегі сөздер толығымен суырылып алынады. Олардың жалпы саны есептелінеді. Оны бағдарламада сөздік деп алдық (vocabulary). Сөздердің барлығы суырылып алынғаннан кейін олар енді бір-бірімен салыстырылады. Салыстырылған сөздердің қай тақырыпқа жақын екендігінің ықтималдықтары көрсетіледі.

```
M=1  V=13
Test File parameters is saved!!!
occurs  1.7283097131005876E-4   nature  1.7283097131005876E-4
occurs  2.1949078138718174E-4   nature  0.002414398595258999
occurs  2.0678246484698098E-4   nature  2.0678246484698098E-4
```

3-сурет. Сөздердің әр тақырыпқа қатысты ықтималдықтары

Жасырын Дирихле үлестірімі арқылы мәтіндерді сегменттеу алгоритмін іске асырамыз:

1. Алдымен осы Z(m,n) - деректер қорындағы сөздердің қай тақырыпқа тәуелді екенін білдіретін матрицаны, nmk(m,k) - мәтін-тақырып матрицаны, nkt(k,t) -тақырып-сөз матрицасын құрып аламыз.

```
men training corpus:::

3  11  1  2  6  11  11  9  7  11  0  0  8  3  7  11  3  7  9  1  0  5  5  7  1  10  7  4  10
6  8  1  8  9  9  2  6  4  2  4  9  8  0  8  9  8  7  1  7  3  6  10  6  1  4  3  9  2  1  1
9  5  3  6  0  9  10  6  7  6  11  9  3  8  5  8  7  6  1  11  7  5  8  0  8  9  8  2  11  7
8  2  1  6  10  5  10  10  9  6  0  6  8  10  5  3  3  6  0  5  3  10  1  5  6  11  11  2  3
4  0  7  11  2  3  2  5  9  9  6  8  9  8  9  11  2  7  9  6  2  7  5  8  4  0  9  1  8  1  8
8  11  1  1  3  4  1  3  2  3  2  7  0  2  2  1  2  0  6  2  3  10  10  7  4  5  9  1  0  5  4
0  4  5  7  9  7  11  2  7  5  10  8  5  10  4  1  3  5  3  8  1  0  6  10  7  1  5  9  6  8
11  11  5  1  7  4  3  8  6  2  11  9  2  7  10  1  11  6  4  8  8  10  9  8  10  6  0  4  7

trmModel.V=2413
```

4-сурет. Деректер қорындағы сөздердің қай тақырыпқа тәуелді екенін білдіретін матрица

1-3. Деректер қорының LDA–моделін тұрғызамыз. Нақтылап айтар болсақ, алдымен Гиббс бойынша үлгі жию алгоритмін пайдаланып, екі үлестірудің мәнін есептеп аламыз. Оны есептеу үшін жеке-жеке мәтін-тақырып үлестіруі $\vartheta - di\vartheta - di$ (5) және тақырып-сөз үлестіруі $\varphi\varphi$ –ді (6) қолданамыз.

1.7143836790673754E-4	1.7143836790673754E-4	1.7143836790673754E-4
1.7143836790673754E-4	1.7143836790673754E-4	0.001885822046974113
1.7143836790673754E-4	1.7143836790673754E-4	1.7143836790673754E-4
1.7143836790673754E-4	0.001885822046974113	1.7143836790673754E-4
1.7143836790673754E-4	1.7143836790673754E-4	1.7143836790673754E-4

5-сурет. Мәтін-тақырып үлестіруі $\vartheta\vartheta$

0.05957767722473605	0.036953242835595784	0.02790346907993967
0.032428355957767725	0.06862745098039216	0.07767722473604827
0.027964205816554812	0.054809843400447436	0.04809843400447428
0.08165548098434006	0.09507829977628636	0.10178970917225952
0.04368108566581849	0.015691263782866838	0.048770144189991524

6-сурет. Тақырып-сөз үлестіруі $\varphi\varphi$

1-5. Сегменттелетін мәтіндердің LDA–моделінде есептелген φ φ мен ϑ ϑ-нің мәнін пайдалана отырып, берілген тақырыпқа сәйкесті мазмұндарды анықтаймыз.

topic0 :	data	allows	time	applications	rate,	rank	processing,	number
topic1 :	provide	products		high	choanoflagellates.		plan	undergo dinosaur
topic2 :	model	example,		words	document		dirichlet	topics latent
topic3 :	apoikozoa		point	characterised	perching		observations	undergo
topic4 :	although		eukaryotic		creation		oxygen	construction provide
topic5 :	heterotrophs:		avian	beaked	lightweight		fixed	metamorphosis ingest
topic6 :	time	standard		temperature		construction	define	multicellular,
topic7 :	atmospheric		approximately		transparent		fixed	analysis define
topic8 :	birds	living	known	group	birds.	great	laying	eggs,
topic9 :	water	ice	liquid	state	chemical		earth's	lakes, oceans,
topic10 :	animals	animal	kingdom	meaning	body		organisms	lives. move
topic11 :	design	although		eventually		process	number	applications creation

7-сурет. Берілген тақырыпқа мазмұндары сәйкес сөздер топтастырылуы

6. Әрбір сөйлемдегі сөздердің ықтималдығын (7) формула бойынша есептеп шығарамыз.

occurs 0.00242664901831017 nature 2.2060445621001546E-4 snow, 0.00242664901
water 2.2060445621001546E-4 strictly 2.2060445621001546E-4 refers 2.2060445621001546E-4
chemical 2.2060445621001546E-4 formula 2.2060445621001546E-4 h2o, 2.2060445621001546E-4
water 2.2060445621001546E-4 transparent 2.2060445621001546E-4 colorless 2.2060445621001546E-4
natural 2.2060445621001546E-4 language 2.2060445621001546E-4 processing, 2.2060445621001546E-4
subgroup 2.2060445621001546E-4 reptilia, 2.2060445621001546E-4 birds 2.2060445621001546E-4
birds 2.2060445621001546E-4 live 2.2060445621001546E-4 worldwide 2.2060445621001546E-4

8-сурет. Сөйлемдегі әрбір сөздің әр тақырыпқа қатысты ықтималдығы

7. Сөйлемдер арасында сақталған жуықтық мөлшерін есептеп шығарамыз. Ол үшін біз косинусоидалық өлшеу тәсілін (8) қолданамыз.

0.9694790001336098	0.990960991610636	0.9900562939374687	0.9082754160423033
0.998739015493811	0.8791127180024486	0.991678236971226	0.9862739899349768
0.9468214875682653	0.8567312675078897	0.9327778234166832	0.8995152574778105
0.9229742008443774	0.9846711769027562	0.9796228892943339	0.9521810049061797

9-сурет. Сөйлемдердің жуықтық шамалары

8. Осы сөйлемдердің шекарасын ажыратамыз. Мұнда біз динамикалық тұрақты сандар алгоритмін (9) пайдаландық.

Тәжірибенің нәтижесі

Біз бұл тәжірибеде K-ның мәні, сондай-ақ әрбір тақырыпқа үлестірілетін сөздердің санының өзгерісіне байланысты, үлестірілген сөздердің аз-көптігінің ұқсас тақырыптарды шынайы бейнелеу дәрежесі мен ұқсамаған тақырыптардың аз-көптігінің мәтін мазмұнын бейнелеу шынайылығына көз жеткіздік. Төменде, K=5 және m=10 болған жағдайлардағы әрбір тақырыптарда үлестірілген сөздердің жағдайлары көрсетілген. Мұнда, K-тақырып саны, m- әрбір тақырыптарда үлестірілген сөздердің саны. Кестелерге қарап, Жасырын Дирихле үлестірілімінің мәтіндерді сегменттеуде ең жуық мәндерге ие болатынын көруге болады.

Нәтижеге қарап тақырып санының артуы және кемуі, сондай-ақ, әрбір тақырыптарда үлестірілген сөздер санының аз-көптігі мәтіндердің не туралы жазылғандығын шынайы бейнеленуіне әсер ететін факторлардың бірі екенін көруге болады.

1 – кесте. K=5, m=10 болған уақытта

topic0	topic	model	document	displaystyle	words
	variables	dirichlet	word	topics	latent
topic1	cat	local	cats	well	veterinarian
	including	infrastructure	make	leads	expectations
topic2	including	immigrants	government	undocumented	people
	country	recent	years	mexico	guidelines
topic3	northern	equity	market	private	industry
	management	high	network	designed	regulatory
topic4	birds	pets	level		considered
	classes	vary	social	animals	rewarding
topic5	trust	ibm	blockchain	private	solution
	financial	services	fund	innovation	security

Қорытынды

Жасырын Дирихле үлестірілімі типтік тақырыптық модель болғандықтан, сөздердің семантикалық мәнін дұрыс бейнелеп бере алды, яғни мағыналас сөздер, жуық мәнді сөздер және көп мағыналы сөздер арасындағы байланыстар мен айырмашылықтарды анық-ашық бейнелеп берді, сол себепті қазіргі таңда, табиғи тілдерді өңдеу саласында Жасырын Дирихле үлестірілімінің қолдану аумағы барған сайын күшейе түсуде.

Тақырыптық айырмашылықтарды анықтау үшін, мазмұны бойынша тексерілетін құжаттарға ұқсас келетін құжаттардағы тақырыптық модельдеріне оқыту жүргізілді. Құрылған алгоритм Википедиядағы 40 құжатты сегменттеуге қолданылды.

Бұл жұмыс нәтижесінен көріп отырғанымыздай, алынған мәтіндік құжаттар көп болған сайын шығатын нәтиженің дәлдік дәрежесі де сонша жоғары болады.

REFERENCES

1. Воронцов К.В. Вероятностное тематическое моделирование [Электронный ресурс] / К.В. Воронцов // MachineLearning.ru. – Режим доступа: ttp://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf (Дата обращения: 03.03.2018).
2. Daud, A. Knowledge discovery through directed probabilistic topic models: a survey / A. Daud, J. Li, L. Zhou, F. Muhammad // Frontiers of Computer Science in China. - 2010. -Vol. 4, Iss. 2. - PP. 280-301.

3. TextFlow: Towards better understanding of evolving topics in text. / W. Cui, S. Liu,L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, X. Tong // IEEE transactions on visualization and computer graphics. 2011. Vol. 17, no. 12. Pp. 2412–2421.
4. Rubin T.N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // Machine Learning. 2012. Vol. 88, no. 1-2. Pp. 157–208.
5. Yeh J.-h., Wu M.-l. Recommendation based on latent topics and social network analysis // Proceedings of the 2010 Second International Conference on Computer Engineering and Applications. Vol. 1. IEEE Computer Society, 2010. Pp. 209–213.
6. Sardinha T. B. Segmenting corpora of texts // DELTA. 2002. Vol. 18. P. 273–286.
7. Li Hang, Yamanishi Kenji. Topic analysis using a finite mixture model // Information Processing and Management. 2003. Vol. 39. P. 521–541.
8. D. M. Blei, A. Y Ng, and M. I. Jordan. 2003. Latent Dirichlet Allocation. JMLR ’03, 3:993–1022.
9. Riedl M., Biemann Ch. TopicTiling: A Text Segmentation Algorithm based on LDA // Student Research Workshop of the 50th Meeting of the Association for Computational Linguistics.