ҒТАХР 20.19.27 DOI: 10.52512/2306-5079-2021-85-1-84-90

МӘТІНДЕР ЖАНРЫН АНЫҚТАУҒА АРНАЛҒАН БЕЛГІЛЕРДІ ЗЕРТТЕУ ЖӘНЕ ТАЛДАУ

С.Ш. Шаншар*, И.М. Уалиева

Әл-Фараби атындағы Қазақ Ұлттық университеті, Алматы, Қазақстан
e-mail*: saniyashanshar@gmail.com

Бұл мақалада мәтіндер жанрын анықтауға арналған белгілерді зерттеу және талдаудағы қолдануға болатын алгоритмдер қарастырылады. Мәтіндер жанрын анықтаудағы ерекшеліктер көрсетілген. Сонымен қатар мәтіндерді сипаттай отырып, тек мәтіндерді қалдырып, керек емес символдарды алып тастайды. Соның ішінен мәліметтер қорымен салыстырмалы түрде анықталады. Мақалада барлық параметрлерге негізделген мәтіндік жанрды автоматты түрде танудың тәжірибелік әдісі сипатталған. Логистикалық регрессия, шешім ағашы, кездейсоқ орман, MLPClassifier, AdaBoostClassifier, GaussianNB алгоритмдерін салыстыра келе мәтіндер үшін ең маңызды параметрлерді таңдау қарастырылды. Мәтіндер жанрын анықтау қазірге кезде ақпараттық қоғамда барлық салада өзекті болып отыр.

Түйін сөздер: мәтіндер жанры, белгілерді зерттеу, машиналық оқыту алгоритмдері, логистикалық регрессия, шешім ағашы, кездейсоқ орман, MLPClassifier, AdaBoostClassifier.

Кіріспе

Мәтінді автоматты түрде жіктеу цифрлық құжаттар пайда болғаннан бері әрдайым маңызды қолданбалы және зерттеу тақырыбы болды. Бүгінгі күні күнделікті қажеттілікке байланысты өте үлкен мәтіндік құжаттардың негізінде мәтінді жіктеу қажетті болып табылады. Жалпы мәтінді жіктеу тақырыптық мәтіндік жіктеуді және мәтіндік жанрға негізделген жіктеуді қамтиды. Тақырыпқа негізделген мәтінді санаттарға бөлу құжаттарды тақырыптарына қарай жіктейді. Мәтіндерді көптеген жанрларда жазуға болады, мысалы: ғылыми мақалалар, жаңалықтар туралы есептер, фильмдерге шолу және жарнама. Солардың негізінде осы зерттеуде қарастырылған мәндер: автомобиль, экономика және бизнес, шоу-бизнес және ойын-сауық, отбасы, сән, компьютер ойындары, денсаулық және медицина, саясат, меншік, ғылым мен технология, спорт, туризм, саяхат, тамақтар. Бұл сыныптардың алыну себебі, қазіргі өмірде қолданылып жатқан барлық мәндерді осы сыныптарға негіздесе болады.

Мәтіндер жанрын анықтаудың тағы бір маңызды бөлігі - жіктеу дәлдігін жақсартуға бағытталған алгоритмдерді дұрыс таңдай білу. Мәтіндер жанрын анықтау бойынша қарастырылатын алгоритмдер: логистикалық регрессия, шешім ағашы, кездейсоқ орман, MLPClassifier, AdaBoostClassifier, GaussianNB.

Әр түрлі мәтін жанрларындағы очерктерді зерттеуде қытай ғалымдары SVM және кездейсоқ орман алгоритмдеріне негізделе отырып, 80% дәлдікке қол жеткізді [1;15]. Ал, мәтінді автоматты түрде жіктеу бойынша машиналық оқыту әдістерінің алгоритміне негізделіп, сол бойынша зерттеу жүргізілген [2;3]. Сол зерттеудің нәтижесі машиналық оқыту бойынша мәтінді автоматты түрде жіктеу жақсы нәтижелер көрсеткен.

Мәтіндер жанрын анықтауға арналған осы мақалада машиналық оқыту бойынша алогритмдер қарастырылған. Сол алгоритмдердің жұмыс істеу принципіне негізделе отырып, мәліметтер қорындағы сыныптарды салыстыра келе қолайлы алгоритмдерді анықтады.

Негізгі бөлім

Жанрды тану белгілі бір параметрлер негізінде жүзеге асырылады, олар салмақтың коэффициенттері берілген мәтіннің лингвистикалық сипаттық актеристикасы деп түсініледі, яғни берілген мәтін үшін сипаттамалардың маңыздылығын көрсететін сандық мәндер. Сәйкесінше, жанрды тану процедурасы екі негізгі мәселені шешуді көздейді: тілдік сипаттамаларды ажырату және оларды өлшеу алгоритмін құру [3;4].

Автоматты түрде веб-құжат жанрын тану жөніндегі жұмыстардан айырмашылығы, бұл жобаның екі ерекшелікті қамтиды:

1. Келесі қорытындылауға арналған кіріс құжаты TXT форматындағы мәтіндік файл болып табылады. Сондықтан HTML-тегтер немесе веб-құжатты форматтаудың кез келген басқа қасиеттері параметрлер ретінде қолданыла алмайды.

2. Басты мақсат адаптивті қорытындылау алгоритмдерін құру болғандықтан, параметрлер ретінде тек дәнекер ғана емес, сонымен қатар мазмұндық сөздер де ескерілуі керек. Дәнекер сөздер жинақтау теориясы мен практикасында мәтінді талдау процесінде ескерілмейтін стоп-сөз ретінде қарастырылады.

Деректер қоры. Мәтіндер жанрын анықтауға арналған белгілерді зерттеу және талдауда табиғаты жағынан ұқсас пәндер (мысалы, экономика және бизнес) оқыту моделіндегі сабақтардың бір-бірімен қабаттасып, дәлдіктің төмендеуіне әкеледі. Мұндай жағдайларда, әдетте, мұндай сыныптар бір топқа біріктіріледі, содан кейін субклассинг немесе сынып ішіндегі құжаттарды қайта жіктеу қолданылады. Бұл автоматты жіктеу жүйесі TFiDF өлшемімен танымал қолдау векторлық машинасын қолданады [4;5]. Осы сыныптарға сәйкес әрбір кіріс құжаты оның сол немесе басқа сыныпқа жақындығының дәрежесін ескере отырып жіктеледі. Егер құжат екі тақырыпқа жақын болса, онда ол сәйкес екі сыныпқа жатады. Егер құжат бірден бірнеше тақырыпқа ұқсас болса, онда бұл шу болуы мүмкін [5;3].

Осы нұсқадағы модель алдын-ала анықталған бірнеше сыныптарда оқытылады: автомобиль, экономика және бизнес, шоу-бизнес және ойын-сауық, отбасы, сән, компьютер ойындары, денсаулық және медицина, саясат, меншік, ғылым мен технология, спорт, туризм, саяхат, тамақтар.

Осы жобада қолданылып жатқан датасет Github жобада (проектіде) https://github.com/zamgi/lingvo--classify сілтемесі бойынша орысша жиналған 2013-2014 жылдан алынды [6; 1]. Алғашқы нұсқа ретінде әртүрлі бағыттағы 13 сынып алынды, сонымен қатар, осы сыныптарды және датасетті қазақ тілінде қолдану үшін Google аудармашы бағдарламасы көмегімен аударылды. Алғашқы нұсқа ретінде олар бөлек 13 түрлі файлда орналасқан. Сол файлдарды python тілінде оқытып әр сыныптағы құжаттарды "\n" арқылы бөліп алынды [7;106]. Соның арқасында орташа есеппен алған 2-3мың құжат әр сыныпта пайда болды. Ал барлық құжаттар саны 22 мыңға жуық. Осы құжаттардың ішінен кездейсоқ түрде оқытуға 80%-ы, және тестілеуге 20%-ы алынады. Осыдан кейін оны тазалау немесе препроцессинг тәсіліне көшті. Бұл процесте артық символдарды, html, тағы да басқа ақпараттың үйретуін қиындатып жіберетін символдар, элементтер алынып тасталды. Сонымен қатар 13-ші "Тамақ жасау" сыныбы алынып тасталды, себебі ондағы объекттер саны 200-ге жуық болды және ол оқыту барысында керісінше әсер көрсетіп, дәлдікті азайтты.

Әдістерге сипаттама

Мақала барысында машиналық оқытуға арналған бірнеше алгоритмдерді жанр бойынша алынған датасеттің көмегімен оқытып, болжам жасалынады. Ол алгоритмдер қатарына логистикалық регрессия, шешім ағашы, кездейсоқ орман, MLPClassifier, AdaBoostClassifier жатады. Осы алгоритмдердің нәтижелерін алып, алдағы уақытта салыстырылады.

Логистикалық регрессия. Негізгі сандық әдіс ретінде логистикалық регрессия (LR) қарастырылды. Екі бейресми пилоттық зерттеулер оның сызықтық дискриминация мен сызықтық регрессияға қарағанда жақсы нәтиже бергендігін көрсетті [8;63]. Логистикалық регрессия - логикалық сілтеме функциясын қолдана отырып, бір немесе бірнеше болжағыш айнымалылардың сызықтық комбинациясы арқылы екілік жауап айнымалысын модельдеуге арналған статистикалық әдіс

$$g(\pi) = log(\pi/(1 - \pi)) \qquad (1)$$

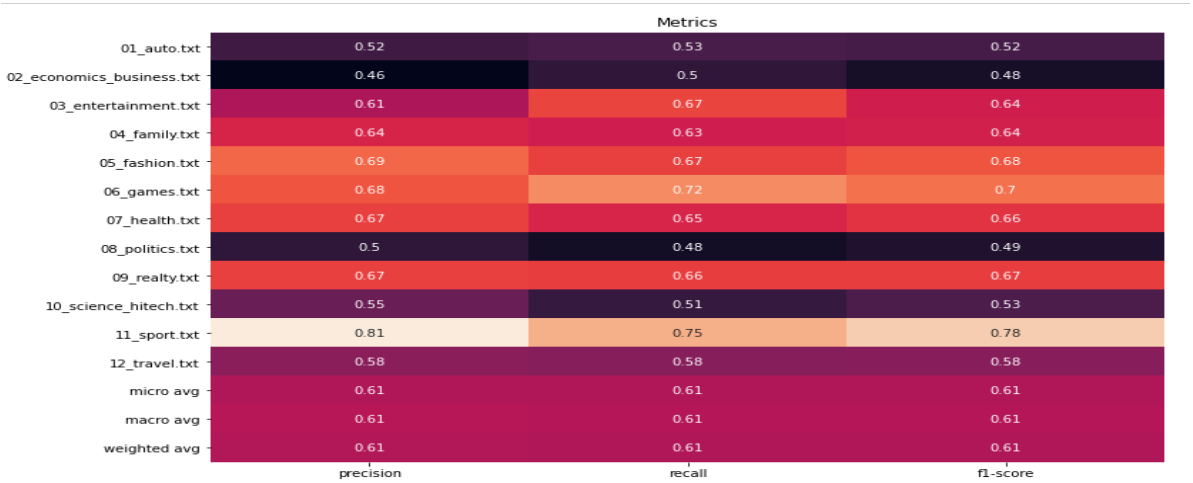
және биномдық кездейсоқ шамамен дисперсияны модельдеу, яғни тәуелді айнымалы журнал (π / (1 - π) тәуелсіз айнымалылардың сызықтық комбинациясы ретінде модельденеді. Модельде g (π) = хіβ формасы бар, мұндағы π - жауаптың болжамды ықтималдығы (біздің жағдайда белгілі бір фасондық мәннің ықтималдығы), хі - бұл мәтін үшін ерекшелік векторы, ал β - салмақ векторы функционалды векторлардың матрицасы. В оңтайлы мәні максималды ықтималдылықты бағалау арқылы шығарылады [9;7].

Екілік шешімдер үшін LR қолдану қарапайым болды. Полимотозды қырлар жанры мен қасы үшін әр қырдың әр деңгейіне болжау функциясын дербес есептеді және ең жоғары болжаммен санатты таңдады.

55 айнымалылардың ішіндегі ең дискриминациясы AIC критерийі бойынша біртіндеп артқа таңдау арқылы таңдалды. Әрбір екілік дискриминация тапсырмасы үшін айнымалылардың жеке жиынтығы таңдалды. Мәтіндер жанрын анықтауға арналған белгілерді зерттеу және талдаудағы логистикалық регрессия классификаторының нәтижесін келесі суреттен көре аламыз(1-сурет):

1-сурет. Мәтіндер жанрын анықтауға арналған белгілерді зерттеу және талдаудағы логистикалық регрессия классификаторының нәтижесі

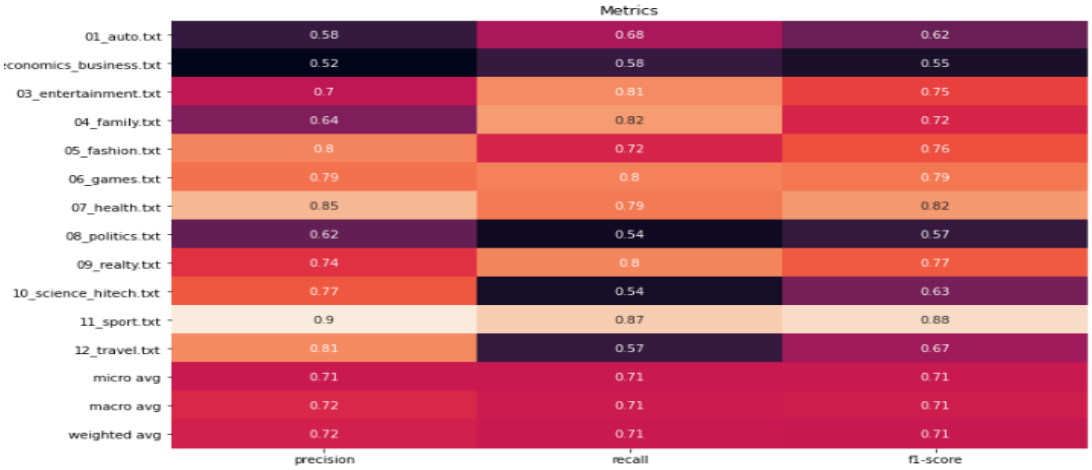
Шешім ағашы алгоритмі. Шешім ағашы - бұл тамыр түйінін, бұтақтар мен жапырақ түйіндерін қамтитын құрылым. Әрбір ішкі түйін атрибут бойынша тестіні, әр тармақ тест нәтижесін білдіреді және әрбір жапырақ түйінінде класс белгісі болады [10;34]. Шешімдер ағашының мәтіндік классификаторы ағаштан тұрады, онда ішкі түйіндер сөздермен белгіленеді, олардан тарайтын бұтақтар тестілеу құжатын ұсынудағы сөздер салмағы бойынша тестермен белгіленеді, ал парақ түйіндері c_i категориялары бойынша белгіленеді. Мұндай жіктеуіш салмақты рекурсивті сынау арқылы d_j тест құжатын санатқа қосады. Ішкі түйіндерді таңбалайтын сөздер d_j бейнесінде, c_i парағының түйініне жеткенге дейін; осы жапырақ түйінінің жапсырмасы d_j-ге тағайындалады [11;178]. Мәтіндер жанрын анықтауға арналған белгілерді зерттеу және талдаудағы шешімдер ағашы классификаторының нәтижесін келесі суреттен көре аламыз (2-сурет):



2-сурет. Мәтіндер жанрын анықтауға арналған белгілерді зерттеу және талдаудағы шешімдер ағашының классификаторы

Кездейсоқ орман алгоритмі. Кездейсоқ ормандар алгоритмін «кездейсоқ орман» көмегімен sklearn кітапханасы арқылы пайдаланылды. Бұл әдіс кездейсоқ шешім ағаштарының санын тудырады және оларды модельді сынау кезінде бір-бірімен бірге қолданады. Бұл модельді пайдалану әдетте бір

шешім ағашына қарағанда жоғары дәлдікті тудырады, өйткені ол нақты сплиттерге мүмкіндік береді, дегенмен сплит санының көп болуына байланысты деректерді артық орналастыру мүмкіндігіне әкелуі мүмкін [12;6]. Кездейсоқ орман алгоритмінің алғашқы іске қосылуы 10 кездейсоқ ағашты, содан кейін 100 кездейсоқ ағашты және тіпті 1000 кездейсоқ ағашты құру арқылы жүзеге асырылды. Мәтіндер жанрын анықтауға арналған белгілерді зерттеу және талдаудағы кездейсоқ орман классификаторының нәтижесін келесі суреттен көре аламыз(3-сурет):



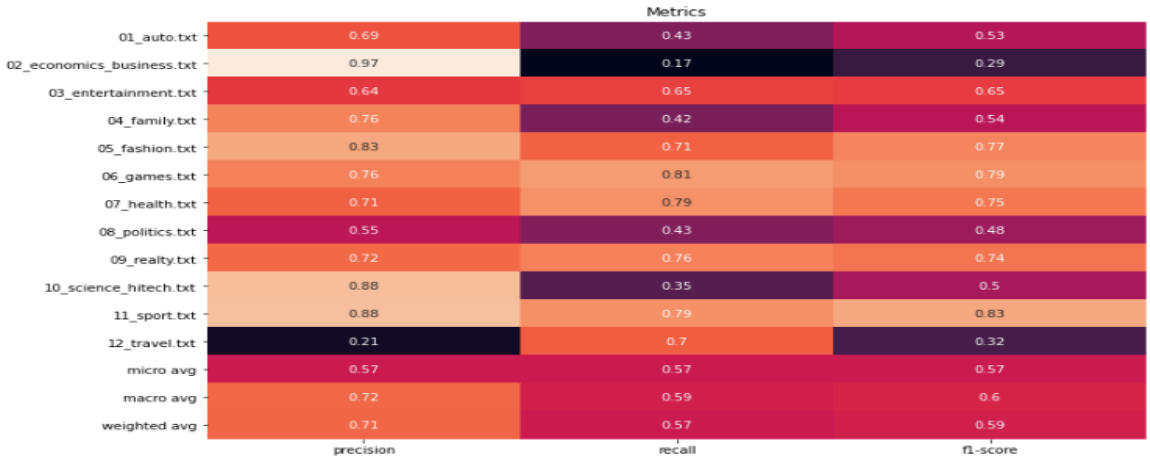
3-сурет. Мәтіндер жанрын анықтауға арналған белгілерді зерттеу және талдаудағы кездейсоқ орман классификаторы

MLPClassifier. MLP - қабатты ациклдік графикамен құрылған, алға жылжитын жүйке желісі класы. MLP кем дегенде үш қабаттан және сызықтық емес активациядан тұрады. Бірінші қабатты кіріс, екінші қабатты жасырын және үшінші қабатты шығару қабаты деп атайды [13;7]. Үш қабат толығымен байланысты, демек жасырын қабаттағы барлық түйіндер басқа қабаттардағы барлық түйіндерге қосылады. Кіріс қабаты өңделетін кіріс сигналын алады. Болжау және жіктеу сияқты қажетті тапсырманы шығыс қабаты орындайды. Кіріс және шығыс деңгейінің арасында орналасқан жасырын қабаттардың ерікті саны MLP-дің шынайы есептеуіші болып табылады. MLP-де мәліметтер кіріс бағытынан шығыс қабатқа алға бағытта ағады. MLP қателік функциясына қатысты градиенттің түсуін есептеу арқылы салмақтары жаңартылатын backpropagation көмегімен оқытылады. Қолдау векторлары немесе Naive Bayes классификаторы сияқты басқа жіктеу алгоритмдерінен айырмашылығы, MLPClassifier жіктеу тапсырмасын орындау үшін негізгі нейрондық желісіне сүйенеді [14;1033]. Мәтіндер жанрын анықтауға арналған белгілерді зерттеу және талдаудағы MLPClassifier-дің нәтижесін келесі суретте көрсетілген (4-сурет):

4-сурет. Мәтіндер жанрын анықтауға арналған белгілерді зерттеу және талдаудағы MLPClassifier-дің нәтижесі

AdaBoostClassifier. Жоғары деңгейде AdaBoost кездейсоқ орманға ұқсайды, өйткені олардың әрқайсысы орман ішіндегі әр класс ағаштарының соңғы классификация туралы шешімін болжайды. Алайда кейбір айырмашылықтар бар. Мысалы, AdaBoost-та шешім ағаштары 1 тереңдікке ие (яғни 2

жапырақ). Сонымен қатар, әр шешім ағашының болжаулары модель жасаған соңғы болжамға әртүрлі әсер етеді [15;112]. Мәтіндер жанрын анықтауға арналған белгілерді зерттеу және талдаудағы AdaBoostClassifier-дің нәтижесін келесі суреттен көре аламыз (5-сурет):



5-сурет. Мәтіндер жанрын анықтауға арналған белгілерді зерттеу және талдаудағы AdaBoostClassifier -дің нәтижесі

Жалпы қолданған датасеттердің негізінде, мәндерге сүйене келе, MLP Classifier-дың тиімді екендігі дәлелденді. Қарастырылып жатқан 12 сыныптың әрқайсысына жеке-жеке көңіл бөлсе:

1-кесте

Сынып аты	Логистикалық регрессия	Шешім ағашы алгоритмі	Кездейсоқ орман алгоритмі	MLP Classifier	AdaBoost Classifier
	F1	F1	F1	F1	F1
Автомобиль	0,82	0,52	0,62	**0,84**	0,53
Экономика және бизнес	**0,76**	0,48	0,55	0,75	0,29
Шоу-бизнес және ойын- сауық	**0,91**	0,64	0,75	0,90	0,65
Отбасы	0,87	0,64	0,72	**0,88**	0,54
Сән	0,88	0,68	0,76	**0,89**	0,77
Компьютер ойындары	0,91	0,70	0,79	**0,92**	0,79
Денсаулық және медицина	0,88	0,66	0,82	**0,89**	0,75
Саясат	**0,79**	0,49	0,57	0,78	0,48
Меншік	**0,85**	0,67	0,77	0,85	0,74
Ғылым мен технология	0,79	0,53	0,63	**0,81**	0,50
Спорт	0,96	0,78	0,88	**0,97**	0,83
Туризм	0,84	0,58	0,67	**0,84**	0,32

Әрбір сыныпқа жеке-жеке сипаттау арқылы осы алгоритмдердің ішінде, автомобиль сыныбы бойынша MLP Classifier, экономика және бизнес сыныбы бойынша логистикалық регрессия классификациясы, шоу-бизнес және ойын сауық сыныбы бойынша логистикалық регрессия, отбасы сыныбы бойынша MLP Classifier, сән сыныбы бойынша MLP Classifier және басқа сыныптарда салыстырмалы түрде MLP Classifier арқылы дәлдігі басқалармен салыстырғанда тиімді болып отыр.

Датасетке қолданылған әр әдіс бойынша алынған нәтижелердің (плоттардың) мәндерінде көрсетілген алгоритмдердің дәлдіктері бойынша салыстырсақ та, MLP Classifier жақсы нәтиже көрсеткенін 2-кестеден көре аламыз.

2-кесте

Алгоритм дәлдігі	Логистикалық регрессия	Шешім ағашы алгоритмі	Кездейсоқ орман алгоритмі	MLP Classifier	AdaBoost Classifier
micro_avg	0,85	0,61	0,71	**0,86**	0,57

Бұл мақаланың мақсаты - тілді табиғи өңдеуде жанрды алуан түрлі салада қолдануға мүмкіндік жасау. Қалған негізгі техникалық қиындықтар - жаттығу кезінде шамадан тыс қалып қоюды болдырмау үшін өзгермелі таңдаудың тиімді стратегиясын табу. Нейрондық желілердің өнімділігі орташа және кейбір кемсітушіліктер үшін әлдеқайда жоғары екендігі (өнімділіктің өзгергіштігі жоғары болғанымен) шамадан тыс үйлесімділік пен айнымалы өзара әрекеттесудің маңызды проблемалар екенін көрсетеді.

Теориялық жағынан мәтіндер негізіндегі жанрларды өмірде көп қолданылатын сыныптар бойынша жасалынды. Мәселені қарастырудың бұл тәсілі әртүрлі жанрлар арасындағы байланысты оларды бірлік ретінде қарастырудың орнына анықтауға мүмкіндік береді. Жаңа жанрларды әлі көрінбейтін қырлар жиынтығына орналастыру үшін негіз бар. Ақырында, жанрларды қыр-сырына қарай бөле отырып, белгілі бір қолдануда қандай жалпы аспект маңызды болатынына назар аудара алады.

Пайдаланылған әдебиеттер

1. Douglas B. (2015) Dimensions of Register Variation: A Cross-Linguistic Comparison. Cambridge University Press, Cambridge, England.
2. Xiao Hu, J. Stephen Downie, and Andreas F. (2009) Ehmann Lyric Text Mining in Music Mood Classification. American music.
3. Kotsiantis S B, Zaharakis I., and Pintelas P. (2007) Supervised machine learning: A review of classification techniques. Frontiers in Artificial Intelligence and Applications, 160, 3.
4. Sebastiani F. (2002) Machine learning in automated text categorization, *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1-47.
5. Seaar Al-Dabooni, Donald Wunsch, (2019), Model Order Reduction Based on Agglomerative Hierarchical Clustering. IEEE Transactions on Neural Networks and Learning Systems (TNNLS) Volume 30, Issue 7 – July, P: 1928-1942
6. https://github.com/zamgi/lingvo--classify
7. Albitar S., Fournier S. and Espinasse B. (2014) An effective TF/IDF-based text-to-text semantic similarity measure for text classification, International Conference on Web Information Systems Engineering, pp. 105-114, October.
8. Bafna, P., Pramod, D., & Vaidya, A. (2016) Document clustering: TF-IDF approach. In Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on (pp. 61-66). IEEE, March.
9. Stamatatos E., Fakotakis N., Kokkinakis G. (2010) Text Genre Detection Using Common Word Frequencies, Proc. of the 18th International Conference on COLING2000.
10. Kessler B., Nunberg G., Schütze H. (1997) Automatic Detection of Text Genre, ACL'97, pages 32-38, July.
11. Gershman A., Meisels A., Lüke K.H., Rokach L., Schclar A., Sturm A. A Decision Tree Based Recommender System. InIICS 2010 Jun 3 (pp. 170-179).
12. Ualiyeva I.M., Mussabayev R.R. (2019) A Close Look at Features for Text Categorization (in publication)
13. Breiman L. (2001) Random forests. Mach. Learn.
14. Sung-Bae Cho, Jee-Haeng Lee (2003) Learning Neural Network Ensemble for Practical Text Classification, Lecture Notes in Computer Science, Volume 2690, Aug, Pages 1032 – 1036.
15. Yatsko V.A., Starikov M.S., and Butakov A. V. (2010) Automatic genre recognition and adaptive text summarization. Autom. Doc. Math. Linguist., vol. 44, no. 3, pp. 111–120.

Исследование и анализ признаков для выявления жанров текстов

С.Ш. Шаншар*, И.М. Уалиева
Казахский национальный университет им. Аль-Фараби, Алматы, Казахстан,
e-mail*: saniyashanshar@gmail.com

В статье рассматриваются алгоритмы, которые можно использовать при изучении и анализе символов для определения жанра текстов. Существуют различия в определении жанра текстов. Алгоритм также определяется описанием текста, удалением ненужных символов, оставлением только текста и сравнением его с базой данных. В статье описан практический метод автоматического распознавания жанра текста по всем параметрам. Сравнивая алгоритмы логистической регрессии, дерева решений, случайного леса, MLPClassifier, AdaBoostClassifier, GaussianNB, рассматривался выбор наиболее важных параметров для текстов. Определение жанра текстов сейчас актуально во всех сферах информационного общества.

Ключевые слова: Жанр текста, исследование символов, алгоритмы машинного обучения, логистическая регрессия, дерево решений, случайный лес, MLPClassifier, AdaBoostClassifier.

Feature selection for automatic detection of text genre

S.Sh. Shanshar*, I.M. Ualiyeva

Kazakh national university named after Al-Farabi, Almaty, Kazakhstan,
e-mail*: saniyashanshar@gmail.com

This article discusses the algorithms that can be used in the study and analysis of symbols to determine the genre of texts. There are differences in defining the genre of texts. Algorithm is also defined by describing the text, removing unnecessary characters, leaving only the text, and comparing it with the database. The article describes a practical method of automatic recognition of the text genre based on all parameters. Comparing the logistics regression, solution tree, random forest, MLPClassifier, AdaBoostClassifier, svm, GaussianNB algorithms, the choice of the most important parameters for the texts was considered. Defining the genre of texts is now relevant in all areas of the information society.

Keywords: Text genre, symbol research, machine learning algorithms, Logistic regression, solution tree, random forest, MLPClassifier, AdaBoostClassifier.

Редакцияға 15.02.2021 түсті.

4-бөлім
ТАРИХ
ЭКОНОМИКА
ҚҰҚЫҚ

Раздел 4
ИСТОРИЯ
ЭКОНОМИКА
ПРАВО

Section 4
HISTORY
ECONOMICS
LAW