IV SHO‘BA
MASHINA TARJIMASINING LINGVISTIK VA DASTURIY TA’MINOTI

COUNTVECTORIZER YORDAMIDA SO‘ZLAR STATISTIKASINI ANIQLASH

Alayev Ruhillo
O‘zbekiston Milliy universiteti dotsenti, t.f.f.d.

Maxmudjonova Gulshaxnoz
ToshDO‘TAU Kompyuter lingvistikasi magistranti.

Annotatsiya: Ushbu maqola tabiiy tilni qayta ishlash va matnga asoslangan mashinali o‘rganishda muhim vosita bo‘lgan CountVectorizer haqida umumiy ma’lumot beradi. CountVectorizer metodologiyasini tushuntiradi hamda so‘z chastotalarini hisoblash va hujjat atamasi matritsasini yaratishni o‘z ichiga oladi. CountVectorizer so‘z statistikasini aniqlash va matnni tahlil qilish vazifalariga yordam beradi.
Kalit so‘zlar: so‘zlar statistikasi, matnni qayta ishlash, chastota, matn, tokenizatsiya,

Annotation: This article provides an overview of CountVectorizer, an important tool in natural language processing and effective machine learning for text. Explains the CountVectorizer methodology and retrieves word frequencies and a document term matrix. CountVectorizer helps with word statistics and text analysis.
Keywords: word statistics, text processing, frequency, text, tokenization

Аннотация: В этой статье представлен обзор CountVectorizer, важного инструмента для обработки естественного языка и эффективного машинного обучения текста. Объясняет методологию CountVectorizer и извлекает частоты слов и матрицу терминов документа. CountVectorizer помогает со статистикой слов и анализом текста.
Ключевые слова: статистика слов, обработка текста, частота, текст, токенизация.

CountVectorizer – bu matn ma’lumotlarini vektorlashtirish uchun ishlatiladigan vosita bo‘lib, u matnni mashinani o‘rganish algoritmlarida ishlatilishi mumkin bo‘lgan raqamli CountVectorizer usuli hujjatda so‘zning paydo bo‘lish chastotasini hisoblab chiqishga asoslangan. Ushbu usul orqali korpusdagi bir necha gaplar asosida so‘zlar matritsasi aniqlanadi va u gapdagi har bi so‘zning chastotasi bilan to‘ldiriladi ma’lumotlarga aylantiradi [Elov B, Hamroyeva Sh, Xusainova Z, Xudayberganov N,. 2023: 81]. CountVectorizer dasturining asosiy g‘oyasi hujjat atamasi matritsasini yaratishdan iborat bo‘lib, unda har bir satr matn korpusidagi hujjatni, har bir ustun esa korpusda topilgan unikal so‘z yoki “termin”ni ifodalaydi. Matritsadagi qiymatlar har bir hujjatdagi har bir atamaning paydo bo‘lish chastotasini ko‘rsatadi.

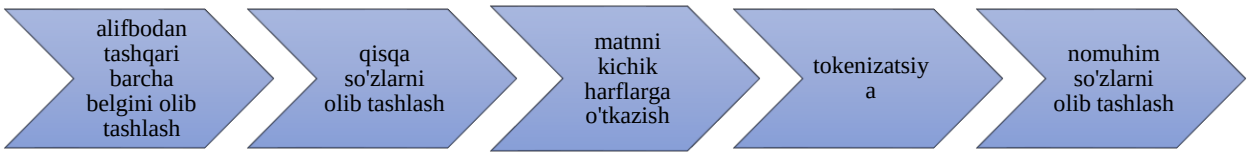
Matnni oldindan qayta ishlash deganda shovqinli, tartiblanmagan matnni aniq maqsadga tayanib tartiblash tushuniladi. Bunda jumlaning kontekstual ma’nosiga hissa qo‘shmaydigan so‘zlar olib tashlanadi. Tabiiy tilni qayta ishlashning muhim qismi bo‘lgan matnni oldindan qayta ishlash turli usullar bilan amalga oshirilishi mumkin, chunki bu usullar muammoning talabiga ko‘ra tanlanadi, ammo mashinali o‘rganish modellarining ishlashini yaxshilash uchun har bir

muammoda qo‘llanilishi kerak bo‘lgan bir nechta usullar mavjud. Mumkin bo‘lgan eng yaxshi natijalarni ta’minlash uchun usullar ma’lum bir tartibda qo‘llanilishi kerak [Aachal J., Harshada J., 2022: 878]

Matnlarini qayta ishlash jarayoni garchi iterativ bo‘lsa ham, mashinali o‘rganish modeli/algoritmi uchun muhim rol o‘ynaydi. Matnli ko‘rinishlarini ikki qismga ajratish mumkin:

1. Matnni diskret ko‘rinishlari (Discrete text representations);
2. Taqsimlangan/uzluksiz matn ko‘rinishlari (Distributed/Continuous text representations)

Korpus matnlarini diskret ko‘rinishda tasvirlashda korpusdagi so‘zlar bir-biridan mustaqil tarzda ifodalanadi. Ushbu yondashuvda so‘zlar korpus(lar) lug’atidagi o‘rniga mos keladigan indekslar bilan ifodalanadi. Ushbu toifaga CountVectorizer metodi ham kiradi [Elov, B., Hamroyeva, Sh., Alaev, R., Xusainova, Z., Yodgorov, U., 2023:117-129]. CountVectorizer imkoniyatlarini ochib berishda ixtiyoriy tanlangan to‘rtta badiiy adabiyotlarning so‘zlar statistikasi va asarda eng ko‘p ishlatilgan 10 ta so‘zlar aniqlanadi. Avvalo, badiiy asardagi matnni qayta ishlash uchun matnga qo‘yilgan maqsaddan kelib chiqib quyidagilar tanlandi:



1-rasm. Matnga qayta ishlov berish

Alifbodan tashqari barcha belgilar olib tashlandi. Alifbodan tashqari belgilarni olib tashlash matnni soddalashtiradi, tahlil qilish va qayta ishlashni osonlashtiradi. Bu, ayniqsa, matnning lingvistik mazmunini tushunishga qaratilgan NLP) vazifalarida foydali bo‘lishi mumkin.

Qisqa so‘zlarni olib tashlash. Qisqa so‘zlar odatda uzunroq so‘zlarga qaraganda kamroq semantik ma’noga ega. Ularni olib tashlash orqali matn ma’lumotlaridagi shovqinni kamaytirish mumkin, bu esa yanada mazmunli tarkibga e’tibor qaratishni osonlashtiradi. Biroq, qisqa so‘zlarni olib tashlash matnni qayta ishlash vazifasining aniq talablari va maqsadlarini hisobga olgan holda amalga oshirilishi kerak. Ba’zan, ma’lum kontekstlarda qisqa so‘zlar muhim ma’lumotlarni o‘z ichiga olishi mumkin. Shu sababli, shovqinni kamaytirish va tegishli kontentni saqlashda muvozanat juda muhimdir.

Matnni kichik harflarga o‘tkazish. Bu qadam matnni oldindan qayta ishlashning eng oddiy va samarali bosqichidir. Matn odatda qisqartma so‘zlardan yoki barcha so‘zlari bosh harflardan iborat bo‘lishi mumkin. Misol uchun matn boshida kelgan “Ona” so‘zi bosh harfda yozilib, matn ichida esa kichik harf bilan yozilgan bo‘lsa bu so‘zlar kompyuter tomonidan ikki xil so‘z sifatida qabul qilinadi va so‘zlarni joylashtirishning keyingi bosqichlarida ikki xil so‘z vektorlari hosil bo‘ladi. Shunday qilib, barcha so‘zlarni kichik harflar bilan yozish matnni qayta ishlashda eng yaxshi amaliyot bo‘ladi.[Alayev, R., Maxmudjonova G., 2023: 79]

Tokenizatsiya. Tokenizatsiya oddiy jarayo bo‘lib, qayta ishlanmagan ma’lumotlarni oladi va uni foydali ma’lumotlarga aylantiradi. Tokenizatsiya tabiiy tilni qayta ishlashda paragraflar va gaplar ma’nosini osonroq belgilash mumkin bo‘lgan kichikroq birliklarga bo‘lish uchun ishlatilinadi [Xusainova, Z., Elov, B., Yodgorov, U., 2023: 27]

Nomuhim so‘zlarni olib tashlash. Nomuhim so‘zlar hujjatning katta qismini tashkil etuvchi matn tarkibidagi so‘zlar bo‘lib, ularning umumiy xususiyati shundaki, hujjatda muhim ma’lumotga ega emas; faqat grammatika tufayli ishlatiladi. Nomuhim so‘zlarni olib tashlash ma’lum bir ma’noga ega bo‘lgan so‘zlarni necha marotaba kelganini aniqlashga yordam beradi [Maxmudjonova G., 2023: 204-211]. Phyton dasturlash tili yordamida bu jarayon amalga oshirildi. Quyida qayta ishlanmagan va qayta ishlangan matn keltirilgan (1-rasm).

	document	clean_doc
0	Quyoshni bahorning boshi va kuzning adog‘ida h...	quyoshni bahorning boshi kuzning adog‘ida qach...
1	XX asrning buyuk yozuvchisi Chingiz Aytmatov b...	asrning buyuk yozuvchisi chingiz aytmatov tani...
2	BUYUK QALB EGASI\nChingiz Aytmatov 1928-yilnin...	buyuk qalb egasi chingiz aytmatov yilning deka...
3	Chol qayiqda yolg‘iz o‘zi Golfstrimda baliq ov...	chol qayiqda yolg‘iz golfstrimda baliq ovlardi...

2-rasm. Matn qayta ishlangandan oldin va keyin

Ixtiyoriy tanlangan badiiy asarlardan olingan natijalar:

Badiiy asarlar	So‘zlar statistikasi
Ahmad Lutfiy Qozonchi “O‘gay ona”	10638
Chingiz Aytmatov, Muxtor Shoxonov “Cho‘qqida qolgan ovchining ohi-zori”	24462
Ernest Xeminguey “Chol va dengiz”	6752
Chingiz Aytmatov “O‘tar qush nolasi”	9509

1-jadval. Badiiy asarlardagi so‘zlar statistikasi

Ahmad Lutfiy Qozonchining “O‘gay ona” asarida eng ko‘p foydalanilgan 10 ta so‘zlar:

Ahmad Lutfiy Qozonchi “O‘gay ona”		Asarda eng ko‘p uchragan so‘zlar	Pythonda Count Vectorizerdan olingan natija
1	bir	539	Top 10 words for 0: 0 bir 539 husayn 176 ona 169 fotimaxonim 148 ikki 134 bor 132 odilbek 101 kun 98 bo‘lgan 85 bo‘ldi 84
2	husayn	176	
3	ona	169	
4	fotimaxonim	148	
5	ikki	134	
6	bor	132	
7	odilbek	101	
8	kun	98	
9	bo‘lgan	85	
10	bo‘ldi	84	

2-jadval. Eng ko‘p ishlatilgan so‘zlar statistikasi

Chingiz Aytmatov, Muxtor Shoxonovning “Cho‘qqida qolgan ovchining ohi-zori” asarida eng ko‘p foydalanilgan 10 ta so‘zlar:

	Chingiz Aytmatov, Muxtor Shoxonov “Cho‘qqida qolgan ovchining ohi-zori”	Asarda eng ko‘p uchragan so‘zlar	Pythonda Count Vectorizerdan olingan natija
1	bir	196	Top 10 words for 2: 0 bir 196 suv 106 bo‘lib 105 qoratoy 105 bo‘lsa 89 sen 88 bo‘ldi 70 isabekov 66 ketdi 64 qiz 62
2	suv	106	
3	bo‘lib	105	
4	qoratoy	105	
5	bo‘lsa	89	
6	sen	88	
7	bo‘ldi	70	
8	isabekov	66	
9	ketdi	64	
10	qiz	62	

3-jadval. Eng ko‘p ishlatilgan so‘zlar statistikasi

Ernest Xemingueyning “Chol va dengiz” asarida eng ko‘p foydalanilgan 10 ta so‘zlar:

	Ernest Xeminguey “Chol va dengiz”	Asarda eng ko‘p uchragan so‘zlar	Pythonda Count Vectorizerdan olingan natija
1	chol	352	Top 10 words for 3: 0 chol 352 baliq 197 bir 178 dedi 148 bo‘lsa 114 bo‘lib 96 o‘yladi 73 bola 73 qayiq 67 olib 62
2	baliq	197	
3	bir	178	
4	dedi	148	
5	bo‘lsa	114	
6	bo‘lib	96	
7	o‘yladi	73	
8	bola	73	
9	qayiq	67	
10	olib	62	

4-jadval. Eng ko‘p ishlatilgan so‘zlar statistikasi

Chingiz Aytmatovning “O‘tar qush nolasi” asarida eng ko‘p foydalanilgan 10 ta so‘zlar:

	Chingiz Aytmatov “O‘tar qush nolasi”	Asarda eng ko‘p uchragan so‘zlar	Pythonda Count Vectorizerdan olingan natija
1	bir	808	
2	bo‘lgan	379	
3	bo‘lib	335	
4	bo‘ladi	264	
5	bor	248	

6	xalq	215	Top 10 words for 1: 0 bir 808 bo‘lgan 379 bo‘lib 335 bo‘ladi 264 bor 248 xalq 215 qilib 198 bitta 183 ikki 157 bo‘lsa 157
7	qilib	198	
8	bitta	183	
9	ikki	157	
10	bo‘lsa	157	

5-jadval. Eng ko‘p ishlatilgan so‘zlar statistikasi

CountVectorizerni har tomonlama o‘rganish uning tabiiy tilni qayta ishlash va matnga asoslangan mashinali o‘rganishda muhim roli bor. So‘z chastotalarini hisoblash va hujjat muddatli matritsalarini yaratishni o‘z ichiga olib, matnni tahlil qilish va tushunishni yaxshilashga yordam beradi. CountVectorizer so‘z statistik tahlillarini osonlashtirish orqali matnli ma’lumotlar haqida chuqurroq ma’lumotga ega bo‘lish imkonini beradi va shu bilan lingvistik ma’lumotlarni qayta ishlashda mashinali o‘rganish algoritmlarining imkoniyatlarini oshiradi.

Foydalanilgan adabiyotlar:

1. Achal J, Harshada J, Bavik J, Charmi Ch “Text Pre-Processing Techniques in Natural Language Processing: A Review” International Research Journal of Engineering and Technology (IRJET), 2022. –B. 878
2. Alayev R, Maxmudjonova G, “O‘zbek tilidagi matnli hujjatlarda izlashni amalga oshirishni takomillashtirish”, Toshkent: O‘zbekistan: til va madaniyat 2023. –B. 79,
3. Elov B, Hamroyeva Sh, Xusainova Z, Xudayberganov N (2023). “O‘zbek tili korpusi matnlarini qayta ishlashda CountVectorizer, TF-IDF hamda Co-occurrence matrix usullarining ahamiyati” Elektron lug’atlar yaratishning nazariy va amaliy asoslari mavzusidagi xalqaro ilmiy-amaliy anjuman materiallari ., Andijon-2023. – B. 81
4. Elov B., Hamroyeva Sh., Alaev R., Xusainova Z., Yodgorov U., “O‘zbek tili korpusi matnlarini qayta ishlash usullari” Raqamli Transformatsiya va Sun’iy Intellekt ilmiy jurnali, 2023. –B. 117-129.
5. Maxmudjonova G., “Nomuhim so‘zlar tushunchasi va uning ahamiyati”. Kompyuter lingvistikasi: muammolar, yechim, istiqbollar Xalqaro ilmiy-amaliy konferensiya materiallari, 2023. 204-211.
6. Xusainova, Z., Elov, B., Yodgorov, U., O‘zbek tili matnlari uchun tokenizayorni ishlab chiqish.., MUHAMMAD AL-XORAZMIY AVLODLARI ilmiy-amaliy va axborat-tahliliy jurnal, 2023. –B. 27