

**ЎЗБЕКИСТОН РЕСПУБЛИКАСИ ФАНЛАР
АКАДЕМИЯСИ МИНТАҚАВИЙ БЎЛИМИ
ХОРАЗМ МАЪМУН АКАДЕМИЯСИ**

**ХОРАЗМ МАЪМУН
АКАДЕМИЯСИ
АХБОРОТНОМАСИ**

Ахборотнома ОАК Раёсатининг 2016-йил 29-декабрдаги 223/4-сон қарори билан биология, қишлоқ хўжалиги, тарих, иқтисодиёт, филология ва архитектура фанлари бўйича докторлик диссертациялари асосий илмий натижаларини чоп этиш тавсия этилган илмий нашрлар рўйхатига киритилган

2024-2/4

**Вестник Хорезмской академии Маъмуна
Издается с 2006 года**

Хива-2024

Бош муҳаррир:

Абдуллаев Икрам Искандарович, б.ф.д., проф.

Бош муҳаррир ўринбосари:

Ҳасанов Шодлик Бекпўлатович, к.ф.н., к.и.х.

Таҳрир хайати:

Абдуллаев Икрам Искандарович, б.ф.д., проф.
Абдуллаева Муборак Махмутовна, б.ф.д., проф.
Абдуҳалимов Баҳром Абдурахимович, т.ф.д., проф.
Агзамова Гулчехра Азизовна, т.ф.д., проф.
Аимбетов Нағмет Каллиевич, и.ф.д., акад.
Аметов Якуб Идрисович, д.б.н., проф.
Бабаджанов Хушнот, ф.ф.н., проф.
Бобожанова Сайёра Хушнудовна, б.ф.н., доц.
Бекчанов Даврон Жуманазарович, к.ф.д.
Буриев Хасан Чутбаевич, б.ф.д., проф.
Ганджаева Лола Атаназаровна, б.ф.д., к.и.х.
Давлетов Санжар Ражабович, тар.ф.д.
Дурдиева Гавҳар Салаевна, арх.ф.д.
Ибрагимов Бахтиёр Тўлаганович, к.ф.д., акад.
Исмаилов Исҳақжон Отабаевич, ф.ф.н., доц.
Жуманиёзов Зоҳид Отабоевич, ф.ф.н., доц.
Жуманов Мурат Арепбаевич, д.б.н., проф.
Кадирова Шахноза Абдухалиловна, к.ф.д., проф.
Қаландаров Назимхон Назирович, б.ф.ф.д., к.и.х.
Каримов Улғубек Темирбаевич, DSc
Курбанбаев Илҳом Жуманазарович, б.ф.д., проф.
Курбанова Саида Бекчановна, ф.ф.н., доц.
Қутлиев Учқун Отобоевич, ф-м.ф.д.
Ламерс Жон, қ/х.ф.д., проф.
Майкл С. Энжел, б.ф.д., проф.
Махмудов Рауфжон Баходирович, ф.ф.д., к.и.х.
Мирзаев Сирожиддин Зайниевич, ф-м.ф.д., проф.
Мирзаева Гулнара Саидарифовна, б.ф.д.
Пазилов Абдуваеит, б.ф.д., проф.

Раззақова Сурайё Раззоқовна, к.ф.ф.д., доц.
Раматов Бакмат Зарипович, қ/х.ф.н., доц.
Рахимов Раҳим Атажанович, т.ф.д., проф.
Рахимов Матназар Шомуротович, б.ф.д., проф.
Рахимова Гўзал Юлдашовна, ф.ф.ф.д., доц.
Рўзметов Бахтияр, и.ф.д., проф.
Рўзметов Дилишод Рўзимбоевич, г.ф.н., к.и.х.
Садуллаев Азимбой, ф-м.ф.д., акад.
Салаев Санъатбек Комилович, и.ф.д., проф.
Сапарбаева Гуландам Машиариповна, ф.ф.ф.д.
Сапаров Қаландар Абдуллаевич, б.ф.д., проф.
Сафаров Алишер Каримджанович, б.ф.д., доц.
Сирожов Ойбек Очилович, с.ф.д., проф.
Собитов Ўлмасбой Тоғажмедович, б.ф.ф.д., к.и.х.
Сотилов Гойипназар, қ/х.ф.д., проф.
Тоғжибаев Комилжон Шаробитдинович, б.ф.д., акад.
Холлиев Аскар Эргашевич, б.ф.д., проф.
Холматов Бахтиёр Рустамович, б.ф.д.
Чўпонов Отаназар Отожонович, ф.ф.д., доц.
Шакарбоев Эркин Бердикулович, б.ф.д., проф.
Эрматова Жамила Исмаиловна, ф.ф.н., доц.
Эшчанов Рузумбой Абдуллаевич, б.ф.д., проф.
Ўразбоев Ғайрат Ўразалиевич, ф-м.ф.д.
Ўрозбоев Абдулла Дурдиевич, ф.ф.д.
Ҳажиева Мақсуда Султоновна, фал.ф.д.
Ҳасанов Шодлик Бекпўлатович, к.ф.н., к.и.х.
Худайберганаева Дурдона Сидиқовна, ф.ф.д.
Худойберганаев Ойбек Икромович, PhD, к.и.х.

Хоразм Маъмун академияси ахборотномаси: илмий журнал.-№2/4 (111), Хоразм Маъмун академияси, 2024 й. – 304 б. – Босма нашрнинг электрон варианты - <http://mamun.uz/uz/page/56>

ISSN 2091-573 X

Муассис: Ўзбекистон Республикаси Фанлар академияси минтақавий бўлими – Хоразм Маъмун академияси

МУНДАРИЖА
ФИЛОЛОГИЯ ФАНЛАРИ

Abdullayeva N.B. Role of tone phonetic means in chinese phonetics and their intercompatibility	6
Abdullayeva X.N. Ingliz va o'zbek sehrli ertaklarida g'aroyib tug'ilish motivi talqini	9
Adizova O.I. Bolalar nutqini o'stirishda tez aytishlarning o'rni	13
Alihonova M. Pragmalinguistic analysis in research	16
Allaberganova A.A. "Devoni mutrib xonaxarob" lingvopoetik tadqiqot obyekti sifatida	18
Ashirmatova M.J. Qishloq xo'jaligi terminlarining leksikografik jihatdan moslashtirish	21
Axmedova M. Ogahiyning "riyozu-davla" asari onomastik birliklari tarixiy-etimologik tahlilining ba'zi masalalari	23
Azatova N.A., Ibodullayeva D. Jahon tilshunosligida etnografizmlarning o'rganilish masalasi	26
Bahromov J. O'zbek va ingliz tillari frazeologik birliklarining etimologik va madaniy xususiyatlari	28
Bekmurodova M.J. Comparative study of gerund, infinitive and participles in english and its equivalents in uzbek	32
Boltaqulova G.F. Phraseological units representing time in english and uzbek languages	35
Bo'riyeva N. J. R. R. Tolkinning "Uzuklar hukmdori" asarida sehr-jodu kontsepti	38
Buronova X.T. O'zbek tili tibbiy terminlarining miqdoriy leksikografik qiyosi	42
Dadabayeva F. "O'tkan kunlar" nemischa tarjimasini mutarjim Barno Oripova talqinida	45
Elov B.B., Alayev R.H., Xusainova Z.Yu., Yodgorov U.S. CBOV neyron tarmoqlari vositasida o'zbek tili so'zlarini bashoratlash	48
Eshmurodov M. Bayoniy tarixiy asarlarining lingvistik xususiyatlari	57
Eshniyazova M.B. Alisher Navoi's "Mahbub ul-qulub" – last work on the basis of the experiences, observations and conclusions	61
Farmonova U. "Qisasi Rabg'uziy" asarida shaxsning ruhiy holati va diniy tushunchalar doirasida shakllangan frazeologizmlar	64
Inoyatova D.I. Xunuklikning tilda leksik va grammatik o'ziga xoslikda verballashuvi	67
Isarov O.R. The linguistic essence of taxis phenomenon	69
Haydarov A.A., Sattorova Sh. Fonografik uslubiy vositalar haqida mulohazalar	72
Haydarov A., Tosheva F. Learning and linguistic foundations of modality category	76
Kaxxarova Sh.Sh. Joy nomlari bilan bog'liq leksik birliklarning lingvomadaniy tahlili	79
Kenzhebeyeva R.S. Theme of Katep in R. Ayapbergenov's poetry	82
Khamidova S.B. Linguistic and aphoristic description of paradoxical text concepts	85
Khamrakulova R.A. Analysing english diplomatic discourse and notable speeches of diplomats	88
Kurbonmurodov A.A. Ekstremistik matnning til xususiyatlari	91
Kurbonova G.I., Jalolova L.Sh. Akutagava Ryunosukening "O'rgimchak uyasining tolasi" va Xans-Xaynts Eversning "O'rgimchak" asarlarida "o'rgimchak" konseptining tipologik tahlili	93
Lolayeva G.G. Metafora va uning gazeta matnidagi lingvostilistik vazifalari	98
Maxmudov R., Masharipova R. Fransuz va o'zbek xalq iboralaridagi zoonimlarning funksiyalari	101
Mahmudova N.R. The use of linguistic gradation at the phonological level in english and uzbek	105
Maxmudov R., Davletova L. Fransuz va o'zbek tillaridagi ayrim geografik terminlarning leksik-semantik xususiyatlari	109
Maxmudov R., Ibadullayev B. O'rxun-Enasoy va Uyg'ur yozuvi manbalaridagi tarixiy antroponimlar	113
Marupova G.U. Linguistic features of sport tourism	117

CBOW NEYRON TARMOQLARI VOSITASIDA O'ZBEK TILI SO'ZLARINI BASHORATLASH

B.B.Elov, PhD, dots., Toshkent davlat o'zbek tili va adabiyoti universiteti, Toshkent

R.H.Alayev, PhD, O'zbekiston Milliy universiteti, Toshkent

Z.Yu.Xusainova, doktorant, Toshkent davlat o'zbek tili va adabiyoti universiteti, Toshkent

U.S.Yodgorov, o'qituvchi, Toshkent davlat o'zbek tili va adabiyoti universiteti, Toshkent

Annotatsiya. Ushbu maqolada bir nechta so'zlarni o'z ichiga olgan matnga mos CBOW modelini shakllantirish usullari va unga oid bir necha sodda misollar keltiriladi. O'zbek tilidagi gapni bitta yashirin qatlamga ega asosiy neyron tarmog'iga uzatish, uni o'rgatish jarayoni va matematik modeli hamda korpus matnlaridagi so'zlarni raqamlashtirishda $n=3$ parametr orqali one-hot encoding vektori hosil qilish tavsiflangan.

Kalit so'zlar: O'zbek tili korpusi, Word2Vec, CBOW, so'zlarni joylashtirish, kontekstli so'zlar, matematik model, vazn qiymati, context words, maqsadli so'z, word embedding.

Аннотация. В этой статье приведены методы построения модели CBOW, включающие несколько слов и примеров, основанных на этой модели. Рассмотрены передача узбекского предложения в основную нейронную сеть, имеющую скрытый слой, процесс его получения и математическую модель, а также в процессе оцифровки построение вектора горячего кодирования через параметр $n=3$.

Ключевые слова: корпус узбекского языка, Word2Vec, CBOW, встраивание слов, контекстные слова, математическая модель, весовое значение, контекстные слова, целевое слово, word embedding.

Abstract. In this article there given methods of constructing CBOW model that include several words and examples based on that model. There examined transferring an Uzbek sentence to the main neural network that has a hidden layer, the process of acquiring it, and mathematical model and in the process of digitization construction of one-hot encoding vector through the $n=3$ parameter.

Key words: Uzbek language corpus, Word2Vec, CBOW, word embedding, context words, mathematical model, weight value, context words, target word, word embedding.

Kirish. Word2vec – matndagi kontekstual va semantik o'xshashlikni aks ettiruvchi so'zlarning taqsimlangan va uzluksiz zich vektorli ko'rinishlarini yaratish uchun neyron tarmoqqa asoslangan model. Word2vec nazoratsiz model bo'lib, katta hajmdagi matn korpusi asosida so'zlarning lug'atini shakllantiradi; ushbu lug'atni ifodalovchi vektor maydonidagi har bir so'z uchun zich so'z

birikmalarini hosil qiladi. Odatda, Word2vec modelida soʻzni joylashtirish vektorlarining hajmini oʻrnatish imkoni boʻlib, vektorlarning umumiy soni asosan lugʻat hajmiga teng boʻladi [1,2]. Word2vec metodining ikkita asosiy yondashuvi mavjud [3,4]:

1. Continuous bag-of-words (CBOW);
2. Skip-gram.

CBOW usuli

Continuous Bag of Words (CBOW) – bu NLPning soʻzlarni joylashtirishda ishlatiladigan mashhur usuli boʻlib, tabiiy tildagi soʻzlar oʻrtasidagi semantik va sintaktik munosabatlarni qamrab oladi [5,6,7]. CBOW – neyron tarmoqqa asoslangan algoritm, u maqsadli soʻzni atrofidagi kontekst soʻzlarini hisobga olgan holda bashorat qiladi. Bu «nazoratsiz» oʻrganishning bir turi boʻlib, u teglanmagan maʼlumotlarni oʻrgatadi va *hissiyotlarni tahlil qilish, matnni tasniflash* hamda *mashina tarjimasini* kabi turli NLP vazifalari uchun ishlatiladi, mumkin boʻlgan soʻzlarni joylashtirishni oldindan tayyorlashga qoʻllanadi. CBOW – bu katta hajmdagi maʼlumotlar toʻplamida oʻqitilishi mumkin boʻlgan oddiy, samarali model boʻlib, *matnni tasniflash* va *tabiiy tilni tushunish* vazifalari uchun yaxshi tanlovdir [7,8,9].

Maqolada bir nechta soʻzlarni oʻz ichiga olgan matnga mos CBOW modeli shakllantiriladi, bir nechta sodd misollar keltiriladi. Ammo CBOW modeli toʻliq imkoniyatidan foydalanish uchun odatda milliardlab soʻzlar bilan oʻqitiladi. Til korpusini Word2Vec modeli orqali oʻqitishda *bir soʻz* yoki *soʻz birikmalaridan* foydalanish mumkin. Word2Vec metodining bir soʻzli arxitekturasini uchun CBOWni amalga oshirish quyidagi bosqichlardan iborat:

Maʼlumotlarni tayyorlash: korpus matnlarini tokenizatsiyalash.

Oʻquv maʼlumotlarini yaratish: korpusga mos lugʻatni shakllantirish, soʻzlarni **one-hot encoding** usuli orqali kodlash, soʻzlarni indekslash [10].

Modelni oʻqitish: bitta soʻzni one-hot encoding sonli formatda neyron tarmogʻiga uzatish, “yoʻqotilish”larni hisoblash orqali *xatolik darajasini* aniqlash va orqaga qaytish yordamida *ogʻirliklarni sozlash*.

Natija: oʻqitilgan model yordamida soʻz vektorini hisoblash va oʻxshash soʻzlarni topish.

1. *Maʼlumotlarni tayyorlash.* Aytaylik, bizda quyidagi matn mavjud: «*men oʻzbek tilini yaxshi koʻraman*».

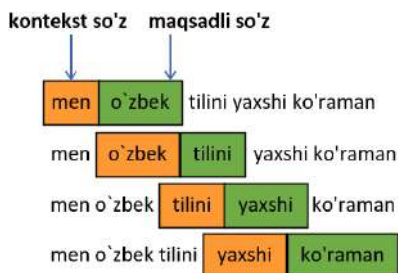
Yuqoridagi gapda bosh harflar va tinish belgilar mavjud emas. Shuningdek, berilgan matnda oʻzbek tilidagi nomuhim soʻzlar keltirilmagan. Katta hajmdagi til korpusi matnlari orqali CBOW modelini shakllantirishda *nomuhim soʻzlarni oʻchirish, sonlarni satr shaklga oʻtkazish, tinish belgilarini olib tashlash* va shunga oʻxshash matnni tozalash ishlarini bajarish kerak. Berilgan matn boshlangʻich qayta ishlash bosqichidan soʻng quyidagi tokenlar roʻyxati hosil qilinadi:

["men", "oʻzbek", "tilini", "yaxshi", "koʻraman"]

2. *Oʻquv maʼlumotlarini yaratish.* Berilgan matn asosida unikal soʻzlardan iborat lugʻatni shakllantirish lozim. Bizning misol matnimizda takroriy soʻz yoʻqligi sababli unikal lugʻat quyidagicha koʻrinishga ega:

["men", "oʻzbek", "tilini", "yaxshi", "koʻraman"]

Keyingi qadamda bitta soʻzli CBOW modeli uchun oʻquv maʼlumotlarini tayyorlashda “**maqsadli soʻz**” (**target word**)ni matndagi berilgan soʻzdan keyin keladigan **kontekstli soʻz** (**context word**) soʻz sifatida aniqlaymiz, yaʼni berilgan soʻzga mos keyingi soʻzni bashorat qilamiz. Berilgan matnni oyna yordamida skanerlash orqali *kontekstli* va *maqsadli soʻzlar* juftliklari hosil qilinadi:



Masalan, “men” kontekst so‘zi uchun maqsadli so‘z “o‘zbek” bo‘ladi. Bizning misolimizda to‘liq o‘quv ma’lumotlari matni quyidagi ko‘rinishga ega:

1-jadval.

O‘quv ma’lumotlarini tayyorlash		
O‘qitish qadami	Kontekst so‘z	Maqsadli so‘z
#1	men	o‘zbek
#2	o‘zbek	tilini
#3	tilini	yaxshi
#4	yaxshi	ko‘raman

One-hot encoding. CBOW algoritmi faqat sonli qiymatlarni qayta ishlashi sababli berilgan matndagi har bir so‘zni raqamli qiymatlarga aylantirish lozim. Masalan, lug‘atda birinchi bo‘lib kelgan “men” so‘zining kodlangan vektor qiymati: $[1,0,0,0,0]$ bo‘ladi. Lug‘atda ikkinchi o‘rinda turadigan “o‘zbek” so‘zi vektor sifatida $[0,1,0,0,0]$ kabi kodlanadi.

2-jadval.

So‘zlarni raqamlashtirish					
	men	o‘zbek	tilini	yaxshi	ko‘raman
men	1	0	0	0	0
o‘zbek	0	1	0	0	0
tilini	0	0	1	0	0
yaxshi	0	0	0	1	0
ko‘raman	0	0	0	0	1

Yuqorida keltirilgan matnga mos kontekst-maqsadli so‘zlarning umumiy to‘plamini one-hot encoding shakliga o‘tkazamiz:

3-jadval.

O‘quv ma’lumotlarini kodlashtirish		
O‘qitish qadami	Kontekst so‘zni kodlashtirish	Maqsadli so‘zni kodlashtirish
#1	$[1,0,0,0,0]$	$[0,1,0,0,0]$
#2	$[0,1,0,0,0]$	$[0,0,1,0,0]$
#3	$[0,0,1,0,0]$	$[0,0,0,1,0]$
#4	$[0,0,0,1,0]$	$[0,0,0,0,1]$

Yuqoridagi 3-jadvalda kodlangan maqsadli so‘z CBOW modeli uchun Y o‘zgaruvchisi, kodlangan kontekst so‘zi uchun X o‘zgaruvchisiga mos keladi. Keyingi qadamda modelni o‘qitish mumkin.

3. Modelni o‘qitish

Keyingi qadamda ushbu o‘quv ma’lumotlarini bitta yashirin qatlam bilan asosiy neyron tarmog‘iga uzatishimiz va uni o‘rgatishimiz kerak. Har qanday so‘zning vektor o‘lchami yashirin tugunlar soniga teng bo‘ladi. Maqolada keltirilgan matn uchun maqsadli vektor o‘lchami 3 ga teng bo‘lsin. Masalan: “men” $\rightarrow [0.021, 0.096, 0.723]$.

n-o‘lcham: bu so‘zni joylashtirish (*word embedding*) o‘lchovi bo‘lib, obyekt, uning nomi, jins va hokazo parametrlarni ifodalaydi. Ushbu n parameter 10, 20, 100 va boshqa qiymatlarga teng bo‘lishi mumkin [10, 11].

Ko‘p hollarda katta hajmli til korpuslarni o‘qitish uchun $n=300$ bo‘ladi. CBOW modelidagi neyron tarmoqlarini o‘qitish ba’zi bosqichlarga bo‘linadi [12,13]:

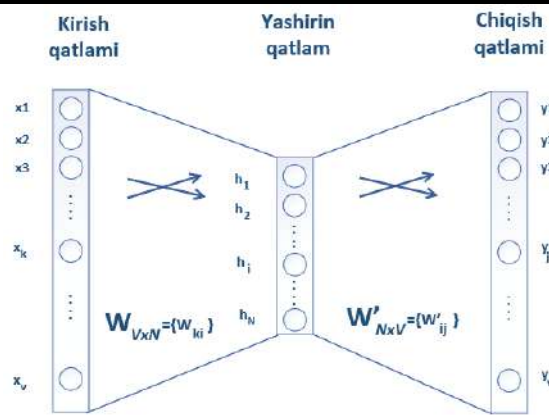
Model arxitekturasini yaratish.

Oldinga harakatlanish.

Xatoliklarni hisoblash.

Og‘irlikni sozlash orqali orqaga qaytish.

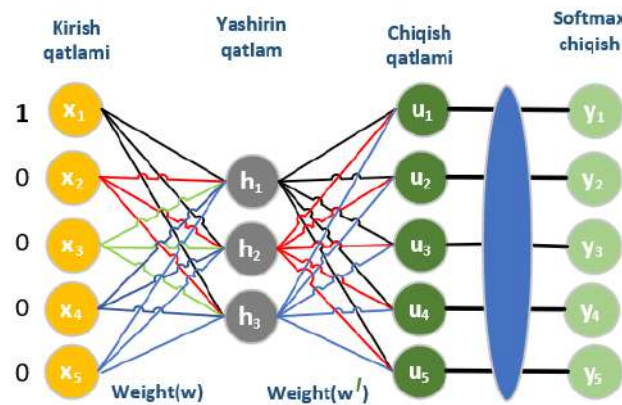
Oldinga harakatlanish bosqichidan oldin 1-rasmda keltirilgan kabi CBOW modeli arxitekturasini vektorli shaklda tushunishimiz kerak [7,14].



1-rasm. CBOW modeli arxitekturasi

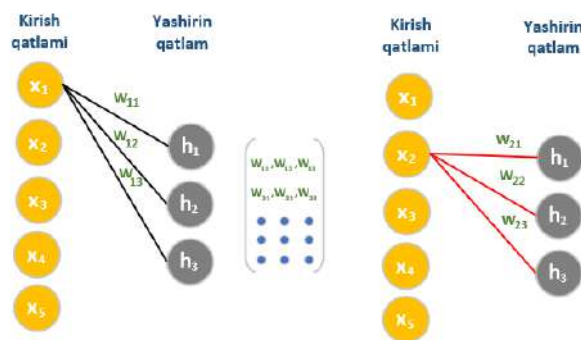
Model arxitekturasini yaratish

Quyida “**men o‘zbek tilini yaxshi ko‘raman**” matni uchun CBOW modeli ishlashini ko‘rib chiqamiz. Aytaylik, bizda kontekst so‘zi “**men**” va maqsadli so‘z “**o‘zbek**” bo‘lsin. Og‘irlikning qiymati $\mathbf{X}=(1,0,0,0)$ bo‘lgan holda, “**men**” so‘zi modelga uzatilganda, $\mathbf{y}=(0,1,0,0)$ ga teng bo‘ladi. Bizning misolda “**o‘zbek**” so‘zi uchun hisoblash amalga oshiriladi.



2-rasm. n=3 ga mos CBOW modeli arxitekturasi

Keyingi qadamda w yashirin qatlam uchun vazn matritsasini shakllantiramiz.



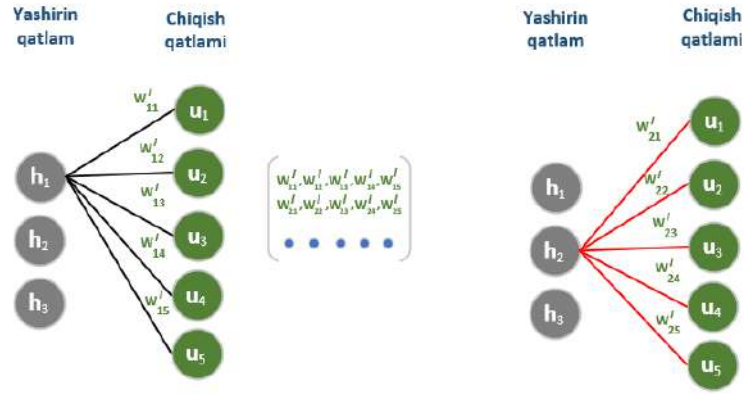
3-rasm. Yashirin qatlam uchun vazn matritsasini yaratish

Yuqoridagi 3-rasmda birinchi ikkita kirish tugunlar (x_1, x_2) uchun vaznli matritsani hosil qilish keltirilgan. Yuqorida keltirilgan tarzda x_2, x_3, x_4 va x_5 , kirish tugunlari uchun vazn matritsasi qiymatlari hisoblangach, $[3 \times 5]$ ($N \times V$) o‘lchamga ega bo‘ladi. Bu yerda,

- N : ichki/yashirin qatlamlar soni;
- V : berilgan korpusga mos unikal lug‘at hajmi.

Keyingi qadamda w' chiqish qatlam uchun vazn matritsasi shakllantiriladi.

$$w = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \\ w_{41} & w_{42} & w_{43} \\ w_{51} & w_{52} & w_{53} \end{bmatrix}$$



4-rasm. Yashirin qatlam uchun vazn matritsasini hosil qilish

Shunday qilib, yuqoridagi 4-rasmda keltirilgan usulga o'xshash tarzda barcha yashirin tugunlar uchun vazn matritsasi shakllantirilgach, uning o'lchami $[5 \times 3]$ ($V \times N$) turlicha bo'lishi mumkin.

$$w' = \begin{bmatrix} w'_{11} & w'_{12} & w'_{13} & w'_{14} & w'_{15} \\ w'_{21} & w'_{22} & w'_{23} & w'_{24} & w'_{25} \\ w'_{31} & w'_{32} & w'_{33} & w'_{34} & w'_{35} \end{bmatrix}$$

Yuqorida keltirilgan amallar bajarilganidan so'ng, CBOW modelining yakuniy shakli quyidagi ko'rinishga ega bo'ladi:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} * \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \\ w_{41} & w_{42} & w_{43} \\ w_{51} & w_{52} & w_{53} \end{bmatrix} = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix} * \begin{bmatrix} w'_{11} & w'_{12} & w'_{13} & w'_{14} & w'_{15} \\ w'_{21} & w'_{22} & w'_{23} & w'_{24} & w'_{25} \\ w'_{31} & w'_{32} & w'_{33} & w'_{34} & w'_{35} \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} \text{softmax} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix}$$

Bu yerda,

– V – unikal lug'at;

– N – yashirin qatlamlar soni.

CBOW modeli arxitekturasini shakllantirgandan so'ng oldinga harakatlanish bosqichiga o'tish mumkin.

Oldinga harakatlanish bosqichi

Ushbu bosqichda yashirin qatlam matritsasi (H)ni shakllantirish lozim:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} * \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \\ w_{41} & w_{42} & w_{43} \\ w_{51} & w_{52} & w_{53} \end{bmatrix} = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix}$$

Bu yerda,

$$h_1 = w_{11}x_1 + w_{21}x_2 + w_{31}x_3 + w_{41}x_4 + w_{51}x_5$$

$$h_2 = w_{12}x_1 + w_{22}x_2 + w_{32}x_3 + w_{42}x_4 + w_{52}x_5$$

$$h_3 = w_{13}x_1 + w_{23}x_2 + w_{33}x_3 + w_{43}x_4 + w_{53}x_5$$

Chiqish qatlami matritsasi (U) qiymatlari quyidagicha hisoblanadi:

$$\begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix} * \begin{bmatrix} w'_{11} & w'_{12} & w'_{13} & w'_{14} & w'_{15} \\ w'_{21} & w'_{22} & w'_{23} & w'_{24} & w'_{25} \\ w'_{31} & w'_{32} & w'_{33} & w'_{34} & w'_{35} \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix}$$

Bu yerda,

$$u_1 = w'_{11}h_1 + w'_{21}h_2 + w'_{31}h_3$$

$$u_2 = w'_{12}h_1 + w'_{22}h_2 + w'_{32}h_3$$

$$u_3 = w'_{13}h_1 + w'_{23}h_2 + w'_{33}h_3$$

$$u_4 = w'_{14}h_1 + w'_{24}h_2 + w'_{34}h_3$$

$$u_5 = w'_{15}h_1 + w'_{25}h_2 + w'_{35}h_3$$

Keyingi qadamda, *softmax qatlami* (y) qiymatlari hisoblanadi:

$$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} \text{softmax} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix}$$

Bu yerda,

$$y_i = \text{softmax}(u_i), \quad i = 1..5$$

Yuqoridagi formulalardan har bir sinf uchun **softmax** ehtimollikni hisoblash mumkin. Softmax funksiyasi [0..1] oralig'idagi qiymatlarni qabul qilganligi sababli eksponentlardan foydalanadi. Quyida softmax funksiyasining faqat bitta (birinchi) chiqishi keltirilgan:

$$y_1 = \frac{e^{u_1}}{(e^{u_1} + e^{u_2} + e^{u_3} + e^{u_4} + e^{u_5})}$$

Demak, yuqoridagi tenglamani umumlashtirilgan holda quyidagi formulani yozishimiz mumkin:

$$y_1 = \frac{e^j}{\sum_{j=1}^V e^j}$$

Xatoliklarni hisoblash

CBOW modelida oldinga harakatlanish amalga oshirilganidan so'ng model xatoliklarini hisoblashimiz, mos ravishda og'irliklar (\mathbf{w} , \mathbf{w}')ni yangilashimiz lozim. Model xatosini hisoblash uchun *haqiqiy qiymatni taxmin qilingan qiymat* bilan taqqoslash kerak. Bitta so'zli CBOW modelida kontekstli so'zdan keyingi so'z – maqsadli so'z hisoblanadi.



CBOW modelida xatoni hisoblash uchun quyidagi tenglamadan foydalaniladi:

$$E = -\log(w_t | w_c)$$

Bu yerda,

w_t – maqsadli so'z;

w_c – kontekstli so'z.

Shunday qilib, endi birinchi iteratsiya uchun xato/zararni hisoblaymiz. Birinchi iteratsiya uchun “o'zbek” maqsadli so'z va uning pozitsiyasi 2 ga teng.

$$\begin{aligned} E(y_2) &= -\log(w_{y_2} | w_{x_1}) = -\log \frac{e^{u_2}}{(e^{u_1} + e^{u_2} + e^{u_3} + e^{u_4} + e^{u_5})} \\ &= -\log(e^{u_2}) + \log(e^{u_1} + e^{u_2} + e^{u_3} + e^{u_4} + e^{u_5}) \\ &= -u_2 + \log(e^{u_1} + e^{u_2} + e^{u_3} + e^{u_4} + e^{u_5}) \end{aligned}$$

Yuqoridagi tenglamani umumlashtirish orqali quyidagi tenglama hosil qilinadi:

$$E = -u_{j^*} + \log \sum_{j=1}^V e^{u_j}$$

Bu yerda, j^* – chiqish qatlamidagi maqsadli so'zning indeksi.

Modelning birinchi iteratsiyasida maqsadli so'zning indeksi 2 ga teng. Shunday qilib, CBOW modelining birinchi iteratsiyasi uchun “o'zbek” so'zi gapda 2-pozitsiyada joylashgani uchun $j^*=2$ bo'ladi.

Og'irlik sozlash orqali orqaga qaytish

Berilgan matn uchun CBOW modelidagi *oldinga harakatlanish va xatoliklarni hisoblash* bosqichlari amalga oshirilganidan so'ng vazn matritsalarini sozlash orqali **orqaga qaytish (back propagation)** bosqichi bajariladi. Orqaga qaytish bosqichini bajarish uchun og'irlik matritsalarini (\mathbf{w} va \mathbf{w}') yangilash talab qilinadi. Og'irlikni yangilash uchun har bir vaznga nisbatan yo'qotish

qiymatini hisoblash, tegishli vazn bilan ko'paytirish talab etiladi. Bu usul **gradient descent** deb nomlanadi.

Ikkinchi w' vaznni yangilash uchun quyidagi misolni ko'rib chiqamiz. Ushbu bosqichda barcha neyronlarning vaznini yangilash uchun yashirin qatlamdan foydalanamiz:

1-qadam. w'_{11} ga nisbatan E gradiyentini hisoblash:

$$\begin{bmatrix} w'_{11} & w'_{12} & w'_{13} & w'_{14} & w'_{15} \\ w'_{21} & w'_{22} & w'_{23} & w'_{24} & w'_{25} \\ w'_{31} & w'_{32} & w'_{33} & w'_{34} & w'_{35} \end{bmatrix} * \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} \xrightarrow{\text{softmax}} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix}$$

$$\frac{dE(y_1)}{dw'_{11}} = \frac{dE(y_1)}{du_1} \cdot \frac{du_1}{dw'_{11}}$$

Natijada,

$$\begin{aligned} E(y_1) &= -u_1 + \log(e^{u_1} + e^{u_2} + e^{u_3} + e^{u_4} + e^{u_5}) \\ \frac{dE(y_1)}{du_1} &= -1 + \frac{d(\log(e^{u_1} + e^{u_2} + e^{u_3} + e^{u_4} + e^{u_5}))}{du_1} \\ &= -1 + \frac{d(\log(e^{u_1} + e^{u_2} + e^{u_3} + e^{u_4} + e^{u_5}))}{d(e^{u_1} + e^{u_2} + e^{u_3} + e^{u_4} + e^{u_5})} \cdot \frac{d(e^{u_1} + e^{u_2} + e^{u_3} + e^{u_4} + e^{u_5})}{du_1} \end{aligned}$$

Ushbu tenglamada bir qator hisoblashlarni amalga oshiramiz:

$$\begin{aligned} \frac{dE(y_1)}{du_1} &= -1 + \frac{1}{e^{u_1} + e^{u_2} + e^{u_3} + e^{u_4} + e^{u_5}} * u_1 \\ &= -1 + \frac{u_1}{e^{u_1} + e^{u_2} + e^{u_3} + e^{u_4} + e^{u_5}} = -1 + y_1 \end{aligned}$$

Yuqoridagi tenglamalarni umumlashtirish orqali quyidagi tenglamani hosil qilamiz:

$$\frac{dE}{du_j} = -\frac{d(u_{j*})}{du_j} + \frac{d(\log \sum_{j=1}^V e^{u_j})}{du_j} = (-t_j + y_j) = e_j$$

Izoh: $t_j = \begin{cases} 1, & \text{agar } t_j = t_{j*} \\ 0, & \text{aks holda} \end{cases}$

Yuqoridagi tenglamada t_j qiymatning haqiqiy natijasi, y_j – taxmin qilingan natija, e_j – xatolik miqdori hisoblanadi. Shunday qilib, birinchi iteratsiya uchun,

$$\begin{aligned} \frac{dE(y_1)}{du_1} &= e_1 \\ \frac{du_1}{dw'_{11}} &= \frac{d(w'_{11}h_1 + w'_{21}h_2 + w'_{31}h_3)}{dw'_{11}} = h_1 \end{aligned}$$

Yuqoridagi tenglamalardan foydalanib, asosiy ko'paytma qiymatini hisoblash mumkin:

$$\frac{dE(y_1)}{dw'_{11}} = \frac{dE(y_1)}{du_1} \cdot \frac{du_1}{dw'_{11}} = e_1 h_1$$

Demak, umumlashtirilgan yakuniy tenglama quyidagi ko'rinishga ega:

$$\frac{dE}{dw'} = e * h$$

2-qadam. w'_{11} vazn qiymatini yangilash:

$$new(w'_{11}) = w'_{11} - \frac{dE(y_1)}{dw'_{11}} = w'_{11} - e_1 h_1$$

1 va 2-qadamlarga o'xshash tarzda $w'_{12}, w'_{13} \dots w'_{35}$ qiymatlarni yangilash mumkin. Keyingi qadamda w vaznni yangilash lozim.

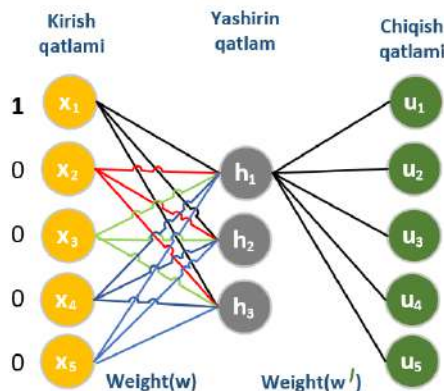
1-qadam. w_{11} ga nisbatan E gradiyentini hisoblash:

$$\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \\ w_{41} & w_{42} & w_{43} \\ w_{51} & w_{52} & w_{53} \end{bmatrix} = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix} * \begin{bmatrix} w'_{11} & w'_{12} & w'_{13} & w'_{14} & w'_{15} \\ w'_{21} & w'_{22} & w'_{23} & w'_{24} & w'_{25} \\ w'_{31} & w'_{32} & w'_{33} & w'_{34} & w'_{35} \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} \xrightarrow{\text{softmax}} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix}$$

Mazkur orqaga qaytish bosqichida kirish qatlamidan yashirin qatlamgacha bo'lgan barcha vazn qiymatlari ($w_{11}, w_{12} \dots w_{53}$) yangilanadi. Ushbu maqolada faqat bitta w_{11} vazn uchun hisoblashlar keltiriladi:

$$\frac{dE}{dw_{11}} = \frac{dE}{dh_1} \cdot \frac{dh_1}{dw_{11}}$$

E orqali $h_1, u_1, u_2, u_3, u_4, u_5$ qiymatlar yangilanadi.



5-rasm. Vazn matritsasini yangilash

Bu yerda,

$$\frac{dE}{dh_1} = \left(\frac{dE}{du_1}, \frac{du_1}{dh_1} \right) + \left(\frac{dE}{du_2}, \frac{du_2}{dh_1} \right) + \left(\frac{dE}{du_3}, \frac{du_3}{dh_1} \right) + \left(\frac{dE}{du_4}, \frac{du_4}{dh_1} \right) + \left(\frac{dE}{du_5}, \frac{du_5}{dh_1} \right) = ew'_{11} + ew'_{12} + ew'_{13} + ew'_{14} + ew'_{15}$$

u_1 va h_1 lar uchun,

$$\frac{du_1}{dh_1} = \frac{d(w'_{11}h_1 + w'_{21}h_2 + w'_{31}h_3)}{dh_1} = w'_{11}$$

h_1 xatolikka mos tarzda $\frac{du_2}{dh_1}, \frac{du_3}{dh_1}, \frac{du_4}{dh_1}, \frac{du_5}{dh_1}$ qiymatlar yuqoridagi tenglamada o'xshash tarzda hisoblanadi.

$$\frac{du_1}{dw_{11}} = \frac{d(w_{11}x_1 + w_{21}x_2 + w_{31}x_3 + w_{41}x_4 + w_{51}x_5)}{dw_{11}}$$

va nihoyat,

$$\frac{dE}{dw_{11}} = \frac{dE}{dh_1} \cdot \frac{dh_1}{dw_{11}} = (ew'_{11} + ew'_{12} + ew'_{13} + ew'_{14} + ew'_{15}) * x$$

2-qadam. w_{11} vazn qiymatini yangilash:

$$new(w_{11}) = w_{11} - \frac{dE}{dw_{11}} = w_{11} - (ew'_{11} + ew'_{12} + ew'_{13} + ew'_{14} + ew'_{15}) * x$$

1 va 2-qadamlarga o'xshash tarzda $w_{12}, w_{13} \dots w_{54}$ vazn qiymatlarini yangilash mumkin.

CBOV usuli tahlili

Korpus matnlarini yuqorida keltirilgan o'qitish jarayoni orqali modellashtirgandan so'ng *modelni sozlash va to'g'ri og'irliklarni o'rnatish* kerak. Modelni o'qitish jarayonining yakunida birinchi vazn matritsasini ko'rib chiqamiz.

Masalan: **“men o'zbek tilini yaxshi ko'raman”.**

Yuqoridagi matnga mos o'qitish modelimizning birinchi og'irligi:

$$w = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \\ w_{41} & w_{42} & w_{43} \\ w_{51} & w_{52} & w_{53} \end{bmatrix}$$

“men” so'ziga mos vazn qiymatlari $[w_{11}, w_{12}, w_{13}]$ dan iborat.

$$\begin{array}{l} \text{men} \\ \text{o'zbek} \\ \text{tilini} \\ \text{yaxshi} \\ \text{ko'raman} \end{array} \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \\ w_{41} & w_{42} & w_{43} \\ w_{51} & w_{52} & w_{53} \end{bmatrix}$$

Endi yakuniy tenglamalarni hosil qilamiz:

1. Oldinga harakatlanish boshqichi tenglamalari:

$$h = wx, \quad u = w'h$$

$$y_j = \text{softmax}(u_j) = \frac{e^j}{\sum_{j=1}^v e^j}$$

$$E = -\log(w_t|w_c) = -u_{j^*} + \log \sum_{j=1}^v e^{u_j}$$

2. Orqaga harakatlanish boshqichi tenglamalari:

2.1. w'_{11} vazn qiymatini yangilash:

$$\frac{dE}{dw'} = e * h$$

$$\text{new}(w'_{11}) = w'_{11} - \frac{dE(y_1)}{dw'_{11}} = w'_{11} - e_1 h_1$$

Umumiy tenglama:

$$\frac{dE}{dw'} = (wx) \oplus e$$

$$\text{new}(w') = w'_{old} - \frac{dE}{dw'}$$

2.2. w_{11} vazn qiymatini yangilash:

$$\frac{dE}{dw_{11}} = (ew'_{11} + ew'_{12} + ew'_{13} + ew'_{14} + ew'_{15}) * x$$

Umumiy tenglama:

$$\frac{dE}{dw} = x \oplus (w'e)$$

$$\text{new}(w) = w_{old} - \frac{dE}{dw}$$

Xulosa. Word2vec modeli Google kompaniyasi tomonidan ishlab chiqilgan; soʻzlarning kontekstual va semantik oʻxshashlikini aks ettiruvchi hamda raqamli vektorli koʻrinishlarini yaratish uchun neyron tarmoqqa asoslangan bashoratlash modelidir. Soʻzlarning raqamli vektor koʻrinishi ularning semantik maʼnosini va soʻzlar oʻrtasidagi munosabatlarni qamrab oladi hamda NLP algoritmlariga matn maʼlumotlari bilan samarali ishlash imkonini beradi. Word2vec nazoratsiz model boʻlib, katta hajmdagi matn korpusini oʻz ichiga olishi, mumkin boʻlgan soʻzlarning lugʻatini yaratishi va ushbu lugʻatni ifodalovchi vektor maydonidagi har bir soʻz uchun zich soʻz birikmalarini yaratishi mumkin. Bugungi kunda Word2vec metodining CBOW va Skip-gram kabi ikkita asosiy yondashuvi mavjud. CBOW modeli neyron tarmogʻiga asoslangan algoritm boʻlib, maqsadli soʻzni uning atrofidagi kontekst soʻzlarini hisobga olgan holda bashorat qiladi. CBOW usuli qoʻshni (kontekst) soʻzlaridan maqsadli soʻzni taxmin qilishga harakat qiladi. CBOW modeli sayoz neyron tarmogʻi boʻlib, nazorat ostidagi oʻrganish algoritmi yordamida oʻqitiladi. Ushbu maqolada til korpusini bitta soʻzli CBOW modeli orqali oʻqitishning 4-bosqichli matematik modeli keltirildi. Katta hajmdagi til korpusi matnlari orqali CBOW modelini shakllantirishda birinchi navbatda nomuhim soʻzlarni oʻchirish, sonlarni satr shaklga oʻtkazish, tinish belgilarini olib tashlash va shunga oʻxshash matnni tozalash ishlarini bajarish lozim. Maqolada $n=3$ boʻlgan hol uchun CBOW modeli arxitekturasi keltirilgan boʻlib, oddiy neyron tarmogʻi orqali oʻzbek tilidagi matnni oʻqitish jarayonining matematik modeli keltirildi. Modelni oʻqish jaronida yuzaga keladigan xatoliklarni qayta ishlash usullari, vazn matritsasini shakllantirishning matematik tenglamalari keltirildi. Neyron tarmogʻi orqali oʻqitilgan matnga mos softmax ehtimollikni hisoblash tenglamalari keltirilgan boʻlib,

w va w' og'irliklarni sozlash orqali CBOW modelidagi orqaga qaytish bosqichi misollar orqali tavsiflandi. Natijada, korpus matnlarini o'qitish jarayoni orqali modellashtirgandan so'ng modelni sozlash va og'irliklarni to'g'ri o'rnatish usullari haqida fikr-mulohaza yuritildi. CBOW modeli – bu NLP vazifalarga sezilarli hissa qo'shadigan so'zlarni joylashtirish usuli. CBOW modelining nazariy asoslarini tushunish, amaliy tatbig'ini o'rganish; afzallik va cheklovlarini tushunish orqali tabiiy tilga ishlov berish, ma'lumot olish va AIning boshqa ilovalarini ishlab chiqish imkonini beradi. NLP tadqiqotlari rivojlanishi bilan CBOW va boshqa so'zlarni joylashtirish modellari rivojlanishda davom etadi, bu esa mashinaga inson tilini tushunish, ular bilan yanada samarali ishlash imkonini beradi.

FOYDANALANILGAN ADABIYOTLAR RO'YXATI:

1. Tan, M., Zhou, W., Zheng, L., & Wang, S. (2012). A Scalable Distributed Syntactic, Semantic, and Lexical Language Model. *Computational Linguistics*, 38(3). https://doi.org/10.1162/COLI_a_00107
2. Sabharwal, N., & Agrawal, A. (2021). Introduction to Word Embeddings. In *Hands-on Question Answering Systems with BERT*. https://doi.org/10.1007/978-1-4842-6664-9_3
3. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2006). Distributed Representations of Words and Phrases and their Compositionality. *Neural Information Processing Systems, 1*.
4. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.
5. Onishi, T., & Shiina, H. (2020). Distributed Representation Computation Using CBOW Model and Skip-gram Model. *2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI)*, 845–846. <https://doi.org/10.1109/IIAI-AAI50415.2020.00179>
6. B.Elov, Z.Xusainova, N.Xudayberganov. (2022). Tabiiy tilni qayta ishlashda Bag of Words algoritmidan foydalanish. *O'zbekiston: til va madaniyat (Amaliy filologiya)*, 2022, 5(4). 31-45
7. Elov B., Aloyev N., Xusainova Z., Yuldashev A. O'zbek tili korpusi matnlarini qayta ishlash Word2Vec, GloVe, ELMO, BERT usullari // Труды XI Международной конференции «Компьютерная обработка тюркских языков» «TURKLANG 2023». Бухара, 20-22 октября 2023 г.
8. Feng, Y., Hu, C., Kamigaito, H., Takamura, H., & Okumura, M. (2022). A Simple and Effective Usage of Word Clusters for CBOW Model. *Journal of Natural Language Processing*, 29(3). <https://doi.org/10.5715/jnlp.29.785>
9. B.ELov, Sh.Khamroeva, Z.Xusainova (2023). The pipeline processing of NLP. *E3S Web of Conferences 413, 03011, INTERAGROMASH 2023*. <https://doi.org/10.1051/e3sconf/202341303011>
10. Rodríguez, P., Bautista, M. A., González, J., & Escalera, S. (2018). Beyond one-hot encoding: Lower dimensional target embedding. *Image and Vision Computing*, 75. <https://doi.org/10.1016/j.imavis.2018.04.004>
11. Elov B., Hamroyeva Sh., Matyakubova N., Yodgorov U. One-hot encoding and Bag-of-Words methods in processing the uzbek language corpus texts // Труды XI Международной конференции «Компьютерная обработка тюркских языков» «TURKLANG 2023». Бухара, 20-22 октября 2023 г.
12. Vrbanec, T., Meštrović, A. (2020). Corpus-based paraphrase detection experiments and review. In *Information (Switzerland)* (Vol. 11, Issue 5). <https://doi.org/10.3390/INFO11050241>
13. Faouzi, H., El-Badaoui, M., Boutalline, M., Tannouche, A., & Ouanan, H. (2023). Towards amazigh word embedding: Corpus creation and word2vec models evaluations. *Revue d'Intelligence Artificielle*, 37(3). <https://doi.org/10.18280/ria.370324>
14. Liu, B. (2020). Text sentiment analysis based on CBOW model and deep learning in big data environment. *Journal of Ambient Intelligence and Humanized Computing*, 11(2). <https://doi.org/10.1007/s12652-018-1095-6>