YUQORI N-GRAM MODELLARINI O’ZBEK TILI MATNLARIGA QO‘LLASH

Botir Elov[1], Ruhillo Alayev[2], Abdulla Abdullayev[3], Narzillo Aloyev[4]

[1]texnika fanlari falsafa doktori, dotsent. Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-pochta: elov@navoiy-uni.uz
[2] texnika fanlari falsafa doktori. Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universiteti
E-pochta: mr.ruhillo@gmail.com
[3]"Urganch innovatsion university" NTM ta'limni kredit tizimini boshqarish bo‘lim boshlig‘i
E-pochta: abdulla_abdullayev9270@mail.ru
[4]tayanch doktorant, Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-pochta: vip.alayev@gmail.com

K E Y W O R D S

Til modellari, n-gram til modeli, o'rtacha logarifmik ehtimollik, modelni baholash, Laplas silliqlash, mashinali o’qitish.

A B S T R A C T

Tabiiy tilni qayta ishlashda kontekstdagi keyingi qaysi so’zni bashorat qilish vazifasi tilni modellashtirish deb ataladi. N-gram tahlili tilni qayta ishlashning muhim usuli bo‘lib, u til tuzilishini tushunishga va gapdagi fragmentdan keyin qanday so`z kelishini bashorat qilishga yordam beradi. N-gramlarni tahlil qilish matn yaratish, imloni tuzatish va hissiyotlarni tahlil qilish kabi murakkab vazifalarda foydalidir. NLP modellari til qoliplarini yaxshiroq tushunishi va qaysi so'zlar birgalikda paydo bo'lishini o'rganish orqali samaraliroq bashorat qilishlari mumkin. N-gram modellari mashina tarjimasi, chatbotlar va qidiruv tizimlari kabi turli xil NLP ilovalarini yaxshilashga yordam beradi.

Kirish

N-gramlar til korpusi yoki deyarli har qanday turdagi ma'lumotlar ketma-ketligidan to'plangan elementlar ketma-ketligidir. N-gramdagi **n** hisobga olinadigan bir qator elementlarning o`lchovini, n=1 uchun unigram, n=2 uchun bigram, n=3 uchun trigram va hokazolarni belgilaydi.

Bugungi kunda tabiiy tillar bizga his-tuyg'ularimiz va fikrlarimizni ifoda etishga yordam beradi, bu esa turli madaniyatlar va jamiyatlarda o'ziga xos g'oyalar va urf-odatlarni ifodalash usulidir. Ushbu tillarni tushunadigan va matn yaratadigan sun'iy intellekt tizimlari til modellari sifatida tanilgan bo`lib, bu o'n yillikdagi eng so'nggi va trendli dasturiy ta'minot texnologiyasidir.

Tilni modellashtirish zamonaviy NLP ilovalarida hal qiluvchi element bo'lib, mashinalarga ma'lumotni samrali tushunishga imkon beradi. Har bir til modeli turi, u yoki bu tarzda, odamlar tomonidan yaratilgan sifatli ma'lumotni miqdoriy ma'lumotga aylantiradi, bu esa o'z navbatida odamlarga cheklangan darajada bir-biri bilan bo'lgani kabi, mashinalar bilan ham muloqot qilish imkonini beradi. Ushbu maqolada yoqori (n>1) N-gram til modellari va ularning o`zbek tili matnlariga qo`llanilish usullari haqida fikr mulohaza yuritiladi.

Tilni modellashtirish

Tilni modellashtirish (Language modeling, LM) – bu gapda berilgan so'zlar ketma-ketligining yuzaga kelish ehtimolini aniqlash uchun turli statistik va ehtimollik usullaridan foydalanish [1,2]. Til modellari ehtimolliklarni gap yoki so'zlar ketma-ketligi yoki oldingi so'zlar to'plamida kelayotgan so'zning ehtimolini belgilaydi. Til modellari haqida quyidagi mulohazalar o`rinli:

– Til modellari *keyingi so'zlarni bashorat qilish, mashina tarjimasi, imloni tuzatish, mualliflikni aniqlash* va *tabiiy tilni yaratish (natural language generation)* kabi ko'plab NLP ilovalari uchun foydalidir.

– Til modellarining asosiy g'oyasi so'z ketma-ketligi bo'yicha ehtimollik taqsimotidan

foydalanish bo`lib, bu ketma-ketlikning *qandaydir sohada gap sifatida sodir bo'lishini* tavsiflaydi.

–Til modellari matnlarning *tilga tegishli bo'lish ehtimolini baholaydi*. Ketma-ketliklar bir nechta elementlarga ajratiladi va til modeli oldingi elementlarni hisobga olgan holda elementning ehtimolini modellashtiradi. Elementlar sifatida *baytlar, belgilar, so'z ostilar (subwords)* yoki *tokenlar* bo'lishi mumkin. Ushbu ketma-ketlikning ehtimoli undagi elementlarning ehtimolliklari ko`paytmasidan iborat.

–Til modellari asosan ikki xilga ajaratiladi: **N-gram til modellari (N-Gram language models)** [3] va **grammatikaga asoslangan til modellari (Grammar-based language models)** [4].

–Shunigdek til modellarini **statistik til modellari (Statistical Language Models)** [5] va **neyron tili modellariga (Neural language model)** [6] ham tasniflash mumkin.

–Tilni modellashtirish yordamida har qanday dasturdan mustaqil ravishda, til modelini **tasodifiy gap generatori** sifatida ishlatish mumkin. Bu holda til modeli ehtimoli bo'yicha gaplar tanlanadi.

Statistik til modellari

Statistik til modelida *ehtimollik til korpusidagi barcha gaplar bo'yicha taqsimlanadi*. Statistik til modellari mavjud matnlar asosida so'zlarning paydo bo'lish ehtimolini o'rganadi. Oddiyroq statistic til modellarida so'zlarning qisqa ketma-ketligi kontekstini ko'rib chiqishi mumkin bo'lsa, kattaroq statistik til modellari gaplar yoki paragraflar darajasida ishlashi mumkin. Odatda, eng ko'p ishlatiladigan statistik til modellari so'zlar darajasida ishlaydi. Shuningdek, deyarli barcha til modellari gapning ehtimolini shartli ehtimollar ko`paytmasiga ajratadi. Statistik til modellari so'zlarning ehtimollik taqsimotini o'rganish uchun **N-gramm, Hidden Markov Models (HMM)** [7] va ma'lum lingvistik qoidalar kabi an'anaviy statistik usullardan foydalanadi. Ushbu modellarning barchasidan eng ommabop va oson amalga oshiriladigani **N-gram tili modellari**dir.

Neyron til modellari

Neyron til modellari tilni modellashtirish uchun *oldinga yo'naltirilgan neyron tarmoqlari (feedforward neural networks), takroriy neyron tarmoqlari (recurrent neural nets), diqqatga asoslangan neyron tarmoqlar (attention-based networks)* va *transferlarga asoslangan neyron tarmoqlar (transformers-based neural nets)* kabi turli xil yondashuvlardan foydalanadi va ularning samaradorligi statistik til modellaridan ko`ra yuqori hisoblanadi [8]. Neyron tili modellari statistik til modellariga nisbatan juda ko'p afzalliklarga ega, chunki ular katta hjamdagi ma`lumotlar tahlilini o'z ichiga oladi. Shuningdek, o'xshash so'zlarning kontekstlarida yaxshiroq umumlashtira oladi va *so'zlarni bashorat qilishda aniqroq natija beradi*.

Neyron tarmoqlarga asoslangan til modellari ancha *murakkab va sekinroq hamda o'qitish uchun ko'proq amallarni talab qiladi* va statistik til modellariga qaraganda kamroq *ancha murakkab*. Demak, ushbu modellar hisoblash quvvati va o'quv ma'lumotlari ko'p bo'lmagan amaliy masalalar uchun, ayniqsa kichikroq vazifalar uchun n-gram tili modeli kabi statistik til modellari to'g'ri vositadir. Biroq, neyron tarmoqlarga asoslangan **katta til modellari (Large language models, LLM)** NLP AIda katta yutuqlarga olib keldi [10,11]. Bugungi kunda ushbu modellar o'rganilgan bilimlar orqali og`zaki va yozma nutqdan foydalanish holatlari bilan bog'liq kundalik hayotda keng qo`llanilmoqda. So'nggi bir necha yil ichida LLM o'lchamlari ham har yili 10 barobar ortib bormoqda va bu modellar o'zlarining imkoniyatlari bilan bir qatorda murakkabligi va hajmi bo'yicha ham o'sib bormoqda.

Til modellarini baholash

Mashinali o'qitish modeli **test ma`lumotlar to`plamidagi gaplarini yaxshi bashorat qilsa**, samarali hisoblanadi. Biz ma'lumotlarning bir qismini parametrlarni baholash uchun saqlab qo'yamiz va qolgan qismini modelni sinab ko'rish uchun ishlatamiz. **Yaxshi model** test ma`lumotlar to`plamiga yuqori ehtimolliklarni belgilaydi, bunda odatda har bir gapning ehtimoli uzunligi bo'yicha normallashtiriladi.

Test ma`lumotlar to`plami namunalari qanchalik bashorat qilinganligini o'lchaydigan muqobil yondashuv **chalkashlik (perplexity)** o'lchovidan foydalanishdir [11,12,13]. Chalkashlikni kamaytiradigan modellar ehtimollikni ham oshiradi. Chalkashlikning matematik formulasini ko'rib chiqsak, bu

shunchaki ehtimollikning teskarisi bo'lib, logarifmik almashinuvlar qo'llaniladi.

Chalkashlik – tasodifiy tanlovlardagi o'lchovdir. Noaniqlik yuqori bo'lgan taqsimotlar yuqori chalkashlikka ega bo`ladi. Uniform (tekis) taqsimot yuqori chalkashlikka ega bo`lib, undan tasodifiy bashorat qilishda foydalanilmaydi. Ushbu maqolada o`zbek tili korpusi matnlari uchun yuqori n-gram modellari ishlab chiqiladi va undagi chalkashlikni kamaytirish usullari keltiriladi.

"O'zbek tili matnlari uchun unigram til modelini ishlab chiqish muammo va yechinlar" deb nomlanuvchi maqolamizda Alisher Navoiyning "Navodir un-nihoya" devoni asosida unigram tili modelini ishlab chiqilgan bo`lib, matndagi har bir so'zning ehtimolliklarini shu matnda so'zning necha marta paydo bo'lishiga qarab baholaydi [14,15].

$$P_{o`quv}(unigram) = \frac{n_{o`quv}(unigram)}{N_{o`quv}}$$

unigram lug`ati hajmi
(o`quv matnidagi unikal unigramlar soni)

Unigram modelini o'qitishda foydalanilgan matn sifatida Alisher Navoiyning "Navodir un-nihoya" devoni tanlangan. Model baholanadigan matnlar Alisher Navoiyning "Badoyi-ul-bidoya" devoni (*uzb_text1*) va mutlaqo boshqa muallif, janr va davrda yozilgan Pirimqul Qodirovning "Avlodlar dovoni" asari (*uzb_text2*)dan foydalanilgan.

Ushbu maqolada bigram (n = 2), trigram (n = 3), quadrigram (n = 4) va quinquegacha (n = 5) yuqoriroq n-gramli modellarni shakllantiriladi. Ushbu modellar avvalgi maqoladagi unigram modelidan farq qiladi, chunki modellarda so'zning ehtimolini baholashda oldingi so'zlarning konteksti hisobga olinadi.

O`quv, uzb_text1 va uzb_text2 baholash matnlari asosida aniqlangan unigram, bigram, trigram, 4-gram va 5-gramlar soni quyidagi jadvalda keltirilgan:

1-jadval.

O`quv, uzb_text1 va uzb_text2 baholash matnlari asosida aniqlangan unigram, bigram, trigram, 4-gram va 5-gramlar

Korpus	unigram	bigram	trigram	4-gram	5-gram
O`quv matn (train)	22129	77438	83719	78127	71453
Baholash matni (uzb_text1)	19724	69841	75951	90968	64802
Baholash matni (uzb_text2)	30206	115455	134168	127895	117889

Yuqori n-gram til modellari

Modelni o'qitish

Berilgan n-gram modeli uchun quyidagi mulohazalar o`rinli:

–Har bir so'zning ehtimoli uning oldidagi n-1 so'zlarga bog'liq. Masalan, trigram modeli (n = 3) uchun har bir so'zning ehtimolligi oldingi 2 so'zga bog'liq.

–Ushbu ehtimollik n-gram o'quv to'plamidagi barcha oldingi (n-1)-gramlar orasida paydo bo'lish ulushi sifatida baholanadi. Ya`ni, n-gram modelini o'qitish bu shartli ehtimollarni o'quv matnidan hisoblashdan iborat.

Quyidagi trigram modelida so'zning ehtimolini hisoblashni ko'rsatilgan:

Berilgan gap: "**Men yangi kitobni o'qidim** </s> "

$$P_{o`quv}(\text{"o'qidim"}|\text{"yangi kitobni"}) = \frac{n_{o`quv}(\text{"yangi kitobni o'qidim"})}{n_{o`quv}(\text{"yangi kitobni"})} \quad (1)$$

Til modelida barcha so'zlar kichik harflar bilan yoziladi va tinish belgilari e'tiborga olinmaydi, hamda gapning yakunida [</s>] tokeni qo`shiladi.

Gap boshidagi so'zlar

Yuqori n-gramli til modellarida har bir gapning boshidagi so'zlar yuqoridagi formulani qo'llash uchun yetarlicha uzun kontekstga ega bo'lmaydi. Ushbu holatlar uchun (1) formulani moslashtirish uchun n-gramlarga gapni boshlovchi [<s>] tokeni qo`shiladi. Quyida trigram modeliga mos ikkita misol keltirilgan:

$$P_{o`quv}(\text{"men"}) = P_{o`quv}(\text{"men"}|\text{"<s>" "<s>"}) = \frac{n_{o`quv}(\text{"<s>" "<s>" "men"})}{n_{o`quv}(\text{"<s>" "<s>"})} \quad (2.1)$$

$$P_{o`quv}(\text{"yangi"}|\text{"men"}) = P_{o`quv}(\text{"yangi"}|\text{"<s>" "men"}) = \frac{n_{o`quv}(\text{"<s>" "men yangi"})}{n_{o`quv}(\text{"<s>" men"})} \quad (2.2)$$

Yuqoridagi (2) formulalarda boshlang'ich belgilarni o'z ichiga olgan n-gramlar boshqa har qanday n-gram kabi ekanligini ko'rish mumkin. Ushbu tokening yagona farqi ularni faqat gap boshida turganlarida sanalishida. Va nihoyat, faqat [<s>] tokenlarini o'z ichiga olgan n-gramlar soni, tabiiyki, bizning o'quv matnimizdagi gaplar soniga teng:

$$P_{o`quv}(\text{"men"}|\text{"<s>""<s>"}) = \frac{n_{o`quv}(\text{"<s>" "<s>" "men"})}{n_{o`quv}(\text{"<s>" "<s>"})} = \frac{n_{o`quv}(\text{"men" @gap boshlandi})}{n_{o`quv}(\text{"<s>" "men"})} \quad (3.1)$$

$$P_{o`quv}(\text{"yangi"}|\text{"<s>" "men"}) = \frac{n_{o`quv}(\text{"<s>" "men yangi"})}{n_{o`quv}(\text{"<s>" men"})} == \frac{n_{o`quv}(\text{"men yangi" @gap boshlandi})}{n_{o`quv}(\text{"men" @gap boshlandi})} \quad (3.2)$$

Noma'lum n-gramlar

Unigram modeliga o'xshab, yuqori n-gramli modellar baholash matnida hech qachon o'quv matnida ko'rinmagan n-gramlarga duch keladi. Bu muammoni Laplas silliqlash ehtimollik formulasining surati va maxrajidagi n-gramlarga **k** psevdo-sonini qo'shish orqali hal qilish mumkin. Biroq, avvalgi maqolada aytib o`tilganidek, Laplas silliqlash n-gram modelini yagona model bilan interpolyatsiya qilishdan boshqa narsa emas.

$$P_{o`quv}(unigram) = \frac{n_{o`quv}(unigram) + k}{N_{o`quv} + kV} = \frac{n_{o`quv}(unigram)}{N_{o`quv} + kV} + \frac{k}{N_{o`quv} + kV} = \frac{N_{o`quv}}{N_{o`quv} + kV}\frac{n_{o`quv}(unigram)}{N_{o`quv}} + \frac{kV}{N_{o`quv} + kV}\frac{k}{kV} = \frac{N_{o`quv}}{N_{o`quv} + kV}\frac{n_{o`quv}(unigram)}{N_{o`quv}} + \frac{kV}{N_{o`quv} + kV}\frac{1}{V} \quad (4)$$

silliqlanmagan unigram ehtimollik

tekis ehtimollik

Unigram modeli uchun Laplas silliqlash: har bir unigramga **k** psevdo soni qo'shiladi. Bunda,

N – o`quv matnidagi so'zlarning umumiy soni;

V – trening matnidagi unikal unigramlar soni.

Shunday qilib, soddalik uchun, o'quv matni mavjud bo`lmagan, biroq baholash matnida ko'rinadigan n-gramlarga 0 ehtimollikni belgilaymiz. Keyinchalik, usbu ehtimolliklar silliqlashtiriladi.

Modelni baholash

Har bir n-gramning shartli ehtimollari o'quv matnidan hisoblab chiqilgandan so'ng, ularni baholash matnidagi har bir so'zga tayinlaymiz. Shuningdek, baholash matnining ehtimolligi qiymati barcha n-gram ehtimolliklarning ko`paytmasiga teng:

Berilgan matn: "**Men yangi kitobni o'qidim**.
U juda qiziqarli ekan"

$$P_{baholash}(T) = P_{umumiy}(\text{"men yangi kitobni o'qidim </s>"})$$

$$P_{umumiy}(\text{"u juda qiziqarli ekan </s>"})$$

$P_{baholash}$("men yangi kitobni o'qidim </s>")
$= P_{o`quv}$("men"|"<s>""<s>")$P_{o`quv}$("yangi"|"<s>""men")

$P_{o`quv}$("kitobni"|"men yangi")$P_{o`quv}$("o'qidim"|"yangi kitobni")

$P_{o`quv}$("</s>"|"kitobni o'qidim")

$P_{baholash}$("u juda qiziqarli ekan </s>")
$= P_{o`quv}$("u"|"<s>""<s>")$P_{o`quv}$("juda"|"<s>""u")

$\cdot P_{o`quv}$("qiziqarli"|"u juda")$P_{o`quv}$("ekan"|"juda qiziqarli")

$P_{o`quv}$("</s>"|"qiziqarli ekan")

$P_{baholash}(T)$
$= P_{o`quv}$("men"|"<s>""<s>")$P_{o`quv}$("yangi"|"<s>""men")

$P_{o`quv}$("kitobni"|"men yangi")$P_{o`quv}$("o'qidim"|"yangi kitobni")

$P_{o`quv}$("</s>"|"kitobni o'qidim")$P_{o`quv}$("u"|"<s>""<s>")

$P_{o`quv}$("juda"|"<s>""u")$P_{o`quv}$("qiziqarli"|"u juda")

$P_{o`quv}$("ekan"|"juda qiziqarli") $P_{o`quv}$("</s>"|"qiziqarli ekan")

Natijada, biz n-gram modeli uchun baholash ko'rsatkichi sifatida o'rtacha logarifmik ehtimollikdan foydalanishimiz mumkin. Bizning n-gram modelimiz qanchalik yaxshi bo'lsa, uning baholash matnidagi har bir so'zga tayinlash ehtimoli yuqori bo'ladi.

$$P_{baholash}(T) = \prod_{so`z} P_{o`quv}(\text{so'z})$$

$$\log\left(P_{baholash}(T)\right) = \prod_{so`z} \log\left(P_{o`quv}(\text{so'z})\right) \qquad (5)$$

$$O'rtacha\ logarifmik\ ehtimollik_{baholash} = \frac{\prod_{so`z} \log\left(P_{o`quv}(\text{so'z})\right)}{N_{baholash}}$$

baholash matnidagi so'zlarning umumiy soni

N-gram modelining qo‘llanilishi

Modelni o`qitish

1. **NgramCounter** klassi tokenlashtirilgan matn faylini qabul qiladi va ushbu matndagi barcha n-gramlarning sonini saqlaydi. Ushbu klass unigram modelidagi **UnigramCounter** klassiga o`xshash bo`lib, ikkita qo‘shimcha xususiyatlar qo`shilgan.

```
from nltk.util import ngrams
train_counter=NgramCounter('../data/train_tok
enized.txt')
```

2. **Ngramodel** klassi **NgramCounter** obyektini qabul qiladi. Klassning **train** metodi orqali har bir n-gram uchun shartli ehtimollik hisoblanadi. Bunda o'quv matnida n-gram paydo bo'lish soni oldingi (n-1)-gram paydo bo'lish soniga bo'linadi:

Gap: Xizmat ko‘rsatgan ta’ lim sohasi xodimlarini taqdirlash marosimi o‘tkazildi.

Trigram: $P_{o`quv}$("o‘tkazildi"|"taqdirlash marosimi")=
$= \frac{n_{o`quv}(\text{"taqdirlash marosimi o‘tkazildi"})}{n_{o`quv}(\text{"taqdirlash marosimi"})}$

3. Bu qiymatlar **counts** atributidan olinadi. Bu holda, avvalgi (n-1)-gram mavjud bo'lmagan unigram modeli bundan mustasno. Buning o'rniga, unigramning o`quv matnida paydo bo'lish sonini matndagi tokenlarning umumiy soniga (hisoblagichning **token_count** atributida saqlanadi) bo'lish yetarli.
4. Aniqlangan ehtimollik qiymati modelning **probs** atributida saqlanadi. Bu har bir n-gramni o'quv matnidagi umumiy shartli ehtimollik bilan taqqoslaydigan qiymat hisoblanadu. Quyida **"marosimi o‘tkazildi"** trigramasiga misol keltirilgan:

```
train_model = Ngramodel(train_counter)
train_model.train()
print(train_counter.counts[('taqdirlash',
'marosimi', 'o‘tkazildi')]) # 39
print(train_counter.counts[('marosimi',
'o‘tkazildi')]) # 536
print(train_model.probs[('taqdirlash',
'marosimi', 'o‘tkazildi')]) # 0.0727 = 39/536
```

5. Agar ushbu n-gram o'quv matnidagi biron bir gapning boshida paydo bo'lsa, uning boshlang'ich shartli ehtimolini hisoblash kerak:

Gap: G‘oliblarni taqdirlash marosimi o‘tkazildi.
$\mathbf{4-gram}$: $P_{o`quv}$("marosimi"|"<s> g‘oliblarni taqdirlash")=
$= \frac{n_{o`quv}(\text{"<s> g‘oliblarni taqdirlash marosimi"})}{n_{o`quv}(\text{"<s> g‘oliblarni taqdirlash"})} =$
$= \frac{n_{o`quv}(\text{"g‘oliblarni taqdirlash marosimi" @gap boshlanishi})}{n_{o`quv}(\text{"g‘oliblarni taqdirlash" @gap boshlanishi})}$

5
– ***gram***: $P_{o`quv}$("marosimi"|"<s> <s> g‘oliblarni taqdirlash")=
$= \frac{n_{o`quv}(\text{"<s> <s> g‘oliblarni taqdirlash marosimi"})}{n_{o`quv}(\text{"<s> <s> g‘oliblarni taqdirlash"})} =$
$= \frac{n_{o`quv}(\text{"g‘oliblarni taqdirlash marosimi" @gap boshlanishi})}{n_{o`quv}(\text{"g‘oliblarni taqdirlash" @gap boshlanishi})}$

6. Bu holda, n-gram va unga mos keladigan (n-1)-gramlar soni **NgramCounter** hisoblagichning **start_counts** atributida topiladi. Unigramlarga mos keladigan (n-1)-gram mavjud emasligi sababli, har bir unigramning boshlanish ehtimoli uning gap boshida paydo bo‘lish sonining o‘quv matnidagi gaplar soniga nisbatiga teng.
7. Hisoblangan ehtimollik qiymati modelning **start_probs** atributida saqlanadi. Bu har bir n-gramni o'quv matnidagi boshlang'ich shartli ehtimoli bilan taqqoslaydigan qiymat hisoblanadu. E'tibor bering, 4 va 5 gramli modellar uchun boshlang'ich ehtimollar yuqoridagi misolda bir xil. Natijada, ular faqat bir marta hisoblab chiqiladi va **start_probs** to`plamida bitta kalit sifatida saqlanadi.

```
print(train_counter.counts[('taqdirlash',
'marosimi', 'o‘tkazildi')]) # 24
print(train_counter.counts[('marosimi',
'o‘tkazildi')]) # 215
print(train_model.probs[('taqdirlash',
'marosimi', 'o‘tkazildi')]) # 0.111 = 24/215
```

Modelni baholash

O'quv matnidan barcha n-gram shartli ehtimollar hisoblab chiqilgandan so'ng, biz ulardan baholash matnidagi har bir so'zga ehtimollikni belgilash uchun foydalanishimiz mumkin. Bu holda, unigram modelidan farq qiladigan yondashuv qo'llanadi: matnning n-gram darajasida logarifmik ehtimolligini hisoblash o'rniga – baholash matnidagi har bir unikal n-gram sonini o'quv matnidagi logarifmik ehtimolligiga ko'paytiriladi.

Gapdagi har bir so'z uchun biz har bir n-gram modeli (shuningdek, uniform model) uchun ushbu so'zning ehtimoli hisoblab chiqiladi va bu ehtimollarni baholash matnining ehtimollik matritsasida qator sifatida saqlanadi. Natijada, bu ehtimollik matritsasi quyidagi o`lchamlarga ega bo'ladi: 1 ta uniform model + 5 n-gramli modellardan iborat ustun hamda baholash matnidagi so‘zlar soni (lu`gat o`lchami)ga teng satrlardan iborat. Ushbu maqolada tanlangan baholash matnlaridagi tokenlar soni quyidagicha: *L(uz_text1)=353110, L(uz_text2)=450066.*

Gapdagi token o`rni		0	1	2	3	4
		uniform	**unigram**	**Bigram**	**trigram**	**4-gram**
0	**men**	$\frac{1}{V}$	$\frac{n(\text{"men"})}{N}$	$\frac{n(\text{"<s> men"})}{N(\text{"<s>"})}$	$\frac{n(\text{"<s><s> men"})}{N(\text{"<s> <s>"})}$	$\frac{n(\text{"<s><s><s> men"})}{N(\text{"<s> <s><s>"})}$
1	**yangi**	$\frac{1}{V}$	$\frac{n(\text{"yangi"})}{N}$	$\frac{n(\text{"men yangi"})}{N(\text{"men"})}$	$\frac{n(\text{"<s>men yangi"})}{N(\text{"<s> men"})}$	$\frac{n(\text{"<s><s> men yangi"})}{N(\text{"<s> <s> men"})}$
2	**kitobni**	$\frac{1}{V}$	$\frac{n(\text{"kitobni"})}{N}$	$\frac{n(\text{"yangi kitobni"})}{N(\text{"yangi"})}$	$\frac{n(\text{"men yangi kitobni"})}{N(\text{"men yangi"})}$	$\frac{n(\text{"<s> men yangi kitobni"})}{N(\text{"<s> men yangi"})}$
3	**o’qidim**	$\frac{1}{V}$	$\frac{n(\text{"o`qidim"})}{N}$	$\frac{n(\text{"kitobni o`qidim"})}{N(\text{"kitobni"})}$	$\frac{n(\text{"yangi kitobni o`qidim"})}{N(\text{"yangi kitobni"})}$	$\frac{n(\text{"men yangi kitobni o`qidim"})}{N(\text{"men yangi kitobni"})}$
4	**</s>**	$\frac{1}{V}$	$\frac{n(\text{"</s>"})}{N}$	$\frac{n(\text{"o`qidim </s>"})}{N(\text{"o`qidim"})}$	$\frac{n(\text{"kitobni o`qidim </s>"})}{N(\text{"kitobni o`qidim"})}$	$\frac{n(\text{"yangi kitobni o`qidim </s>"})}{N(\text{"yangi kitobni o`qidim"})}$
0	**u**			…	…	

1-rasm. n-gram modeli ehtimollik matrisasi.

Bu yerda,

V	o`quv matnidagi unikal unigramlar soni;
N	o`quv matnidagi tokenlarning umumiy soni;
= belgisi	modelning **start_probs** lu`gatida faqat bir marta saqlanadigan bir xil boshlang'ich ehtimolliklar

Ehtimollar matritsasi

Uniform modelida har bir so'z uchun bir xil ehtimollikdan foydalaniladi, ya'ni o'quv matnidagi **1/(unikal unigramlar soni)**. Natijada, ehtimollik matritsasining birinchi ustunini ushbu ehtimolga mos uniform_prob atributida saqlanadi.

```
prob_matrix[:, 0] = self.uniform_prob
```

N-gram ehtimolini aniqlash uchun n-gram har doim gapdagi joriy so'z bilan tugashi sababli:

```
ngram_end = token_position + 1
```

Berilgan n-gramning boshlanishi tabiiy ravishda n-gram uzunligini olib tashlagan holda oxirgi pozitsiyasi:

```
ngram_start = token_position + 1 – ngram_length
```

Agar bu boshlang'ich pozitsiya qiymati manfiy bo'lsa, bu so'z n-gram modeli uchun yetarli kontekstga ega bo'lish uchun gap boshida paydo bo'lishini anglatadi. Yuqoridagi (1) misoldagi gapda trigram modelida **"yangi"** so'zi uchun: **token_position=1** va **ngram_length=3** ga teng. Natijada, **ngram_end=1+1=2, ngram_start=2-3=-1** o`rinli. Shu sababli, **"yangi"** bilan tugaydigan trigrama boshlang'ich **<s>** tokeni bilan to'ldirilishi kerak:

Trigram: $P_{o`quv}$("yangi"|"<s> men")=

$$= \frac{n_{o`quv}(\text{"<s> men yangi"})}{n_{o`quv}(\text{"<s> men"})} =$$

$$= \frac{n_{o`quv}(\text{"men yangi" @gap boshlanishi})}{n_{o`quv}(\text{"men" @gap boshlanishi})}$$

Bu holda, shartli ehtimollik boshlang'ich shartli ehtimolga aylanadi:
"<s> men yangi" trigramasi **"men yangi"** boshlang'ich n-gramasi bo'ladi.
Agar n-gram o'quv matnida mavjud bo`lmasa bo'lsa, avval aytib o'tilganidek, unga **0** ga teng ehtimollik beramiz.

```
if ngram_start < 0:
    ngram = tuple(sentence[0:ngram_end])
    prob_matrix[row, ngram_length] =
self.start_probs.get(ngram, 0)
```

Aks holda, agar n-gramning boshlang'ich pozitsiyasi **0** ga teng yoki undan katta bo'lsa, demak, n-gram gapda mavjud va uni oddiygina boshlang'ich va oxirgi pozitsiyasi bilan ajratib olish mumkin. Keyingi qadamda o'quv modelining **probs** atributidan uning shartli ehtimolligini aniqlanadi. Agar lug`atda n-gram mavjud bo`lmasa, uning ehtimoli ham **0** ga teng bo'ladi.

Trigram: $P_{o`quv}$("o`qidim"|"yangi kitobni")=

$$= \frac{n_{o`quv}(\text{"yangi kitobni o`qidim"})}{n_{o`quv}(\text{"yangi kitobni"})}$$

```
else:
    ngram =
tuple(sentence[ngram_start:ngram_end])
    prob_matrix[row, ngram_length] =
self.probs.get(ngram, 0)
```

Yuqoridagi barcha amallar **Ngramodel** klasining **evaluate** metodi vositasida amalga oshiriladi, bu esa tokenlashtirilgan baholash matnining faylni kiritish ma`lumoti sifatida qabul qiladi. **Evaluate** metodi tokenlashtirilgan matndagi har bir so'zni o'qiydi va ehtimollik matritsasidagi ushbu so'zning tegishli qatorini to'ldiradi. Quyidagi dastur kodida o`quv ma`lumotlari to`plami uchun n-gram modellarini o'qitish va ularni **uzb_text1** to`plamida baholash kodi keltirilgan. **Uzb_text1** to`plamidagi modellarni baholashda ehtimollik matritsasining eng birinchi 3 qatori oxirida ko'rsatilgan.

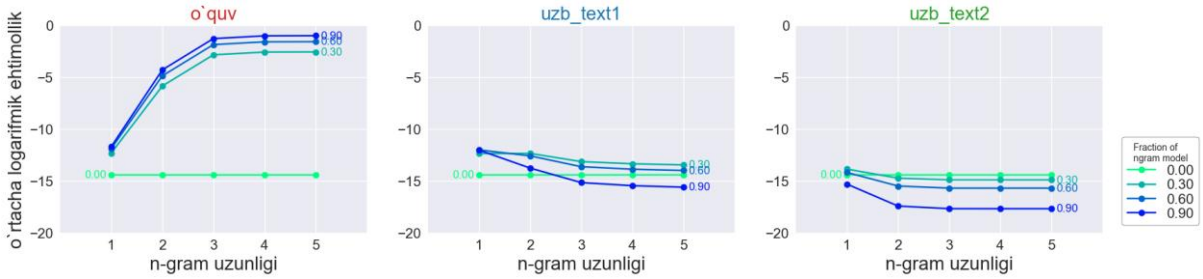
```
train_model = Ngramodel(train_counter)
train_model.train()
dev1_prob_matrix =
train_model.evaluate('../data/dev1_tokenized.txt
')
print(dev1_prob_matrix[:3])
# [[8.26787929e-05 3.11072953e-06 3.91987770e-05
3.91987770e-05 3.91987770e-05 3.91987770e-05]
#  [8.26787929e-05 7.93578210e-02 1.00000000e+00
1.00000000e+00 1.00000000e+00 1.00000000e+00]
#  [8.26787929e-05 5.59246955e-02 1.03602368e-01
1.03602368e-01 1.03602368e-01 1.03602368e-01]]
```

Natijalar tahlili

N-gram modellarini quyida keltirilgan 3 ta konfiguratsiyada baholashni amalga oshiramiz:

1. **N-gram uzunligi**: unigramdan (n=1) 5 gramgacha (n=5);
2. **N-gram interpolyatsiya og'irligi (vazni):** Har bir n-gram modeli uchun uni har xil og'irligidagi yagona model bilan interpolyatsiya qilinadi:
   0 dan (0% n-gram, 100% bir xil) **0.3, 0.6** va **0.9** (90% n-gram, gram, 10% bir xil) gacha. Bunda, n-gram modeli hech qachon 100% bo'la olmaydi, chunki 0 ehtimoli bo'lgan n-gramlar o'rtacha log ehtimolini manfiy cheksizlik tomon tortadi.
3. **Baholash matni:** Modelni o'quv matni (**train.txt**) va ikkita baholash matnida (**uzb_text1** va **uzb_text2**) baholaymiz.

Quyidagi grafikda n-gram modellari, interpolyatsiya og'irliklari va baholash matni bo'yicha o'rtacha ehtimolliklar ko'rsatilgan:



2-rasm. n-gram modellari, interpolyatsiya og'irliklari va baholash matni bo'yicha o'rtacha ehtimolliklari

Yuqoridagi grafikdagi keltirilgan ma`lumotlarni tahlil qilamiz.

1. *O`quv ma`lumotlari (chap tomondagi 1-grafik)*

N-gram uzunligi oshgani sayin, o'quv matnida n-gram modeli shunchalik yaxshi bo'ladi. Bu tabiiy, chunki n-gram qanchalik uzun bo'lsa, bir xil kontekstga ega bo'lgan n-gramlar shunchalik kam bo'ladi. Natijada, bu n-gram (shartli) ehtimollik chegarasining katta qismini egallashi mumkin. Ushbu hodisa turli n-gram modellari asosida **"men bugun yangi kitobni o'qidim"** gapida **"o`qidim"** so'zining ehtimolini baholash quyidagi misolida tasvirlangan:

1-jadval. Turli n-gram modellari asosida "o`qidim" so'zining ehtimolini baholash

Model	Ehtimollik	Baho	# n-gram	# (n-1)gram	Kontekstdagi boshqa n-gramlar
Unigram	P("o'qidim")	0.00054	174	3215468	</s>, gazetani, jurnalni, …
Bigram	P("o'qidim"|"kitobni")	0.032	1	31	kitobni oldim, kitobni berdim, …
Trigram	P("o'qidim"|"yangi kitobni")	0.125	1	8	yangi kitobni oldim, yangi kitobni uzatdim, …
5-gram	P("o'qidim"|"bugun yangi kitobni")	0.5	1	2	bugun yangi kitobni berdim
6-gram	P("men bugun yangi kitobni o'qidim")	1	1	1	–

Yuqoridagi jadvaldagi qiymatlaridan **"o`qidim"** so'zining misolidan **"kitobni" - "kitobni oldim", "kitobni berdim"** kabi bir xil kontekstli ko'plab bigramalar mavjud bo'lsa-da, bir xil kontekstli 4 gram kamroq ekanligini qayd etish mumkin: **"bugun yangi kitobni o`qidim" - "bugun yangi kitobni berdim"**. Natijada, oxirgi modelda **"o`qidim"** so`zi birinchi modellarga qaraganda ancha yuqori ehtimolga ega bo`ladi.

2. *Baholash ma`lumotlari (o`rtadagi 2-grafik)*

– Unigramdan bigram modeliga o'tganimizda, **uzb_text1** matnining o'rtacha logarifmik ehtimolligi biroz ortadi. Bu, ehtimol, ikkita kitobning umumiy bigramalari bilan bog'liq, chunki ular bir muallifning bir xil seriyasidan.

$P_{o`quv}$("botir") = *0.00000622*

$P_{o`quv}$("botir"|"elov") = *1*

– Unigram bilan solishtirganda bigram ehtimollik bahosi eng katta yaxshilanishga ega bo'lgan holatlar asosan atoqli otlar (ismlar, joy nomlari). Masalan, "botir" unigramasini hisobga olsak, unga yuzaga kelishi mumkin bo'lgan barcha mumkin bo'lgan unigramalardan yuqori ehtimollik berish juda qiyin. Ammo, agar oldingi so'z "elov" ekanligini bilsak, keyingi so'z "botir" ekanligiga ishonchimiz komil, chunki bu ikki so'z har doim o'quv matnida bigrama sifatida birga keladi.

– Biroq, bigramdan yuqori n-gramli modellarga o'tayotganimizda, o'rtacha logarifmik ehtimollik keskin pasayadi! Bu asosan, **uzb_text1** baholash matnida mavjud, lekin o`quv matnida bo'lmagan noma'lum n-gramlarning ko'pligi bilan bog'liq. Ular **0** teng ehtimolga ega bo'ladi, bu esa baholash matnining o'rtacha logarifmik ehtimolini sezilarli darajada pasaytiradi. Quyida har bir model uchun noma'lum n-gramlar soni, shuningdek, baholash matnidagi noma'lum so'zlarning foizi ko'rsatilgan:

2-jadval. Modeldagi noma'lum n-gramlar soni va baholash matnidagi noma'lum so'zlarning foizi

Model	noma'lum n-gramlar soni	noma'lum n-gramlarning foizi
unigram	11500	3.2
bigram	108509	30.7
trigram	233681	66.2
5-gram	284413	80.5
6-gram	296048	83.8

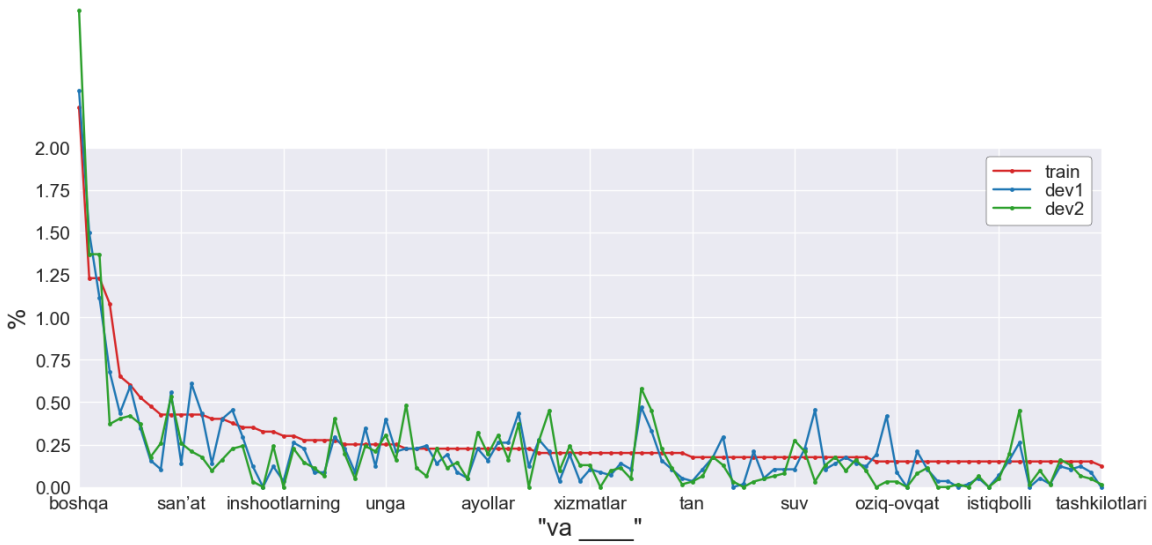
Yuqoridagi natijalari asosiy mashinali o'qitsh tamoyilini tasdiqlaydi:
Murakkabroq model har doim ham yaxshiroq emas, ayniqsa o'quv ma'lumotlari kichik bo'lsa.

Bizning holatda, kichik hajmdagi o'quv ma'lumotlari o'quv matnida ko'rinmaydigan ko'plab n-gramlar bo'lishini anglatadi. Bu muammo murakkabroq modeldan foydalanilganda yanada kuchayadi: o‘quv matnidagi 5 gramlik bigramaga qaraganda boshqa matnda takrorlanish ehtimoli kamroq bo`ladi.

3. *Baholash ma`lumotlari (oxiridagi 3-grafik)*

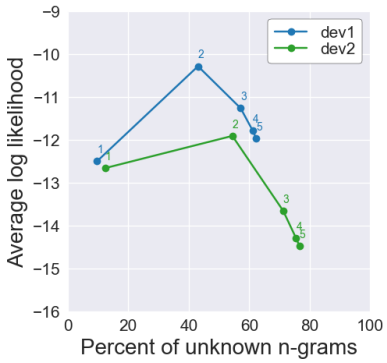
uzb_text2 baholash matnida bir xil n-gram modellari baholanganda, biz n-gram modelidan yoki uniform model bilan qanchalik interpolyatsiya qilinganidan qat'i nazar, **uzb_text2** matnidagi samaradorlik ko`rsatkichi odatda **uzb_text1** matnidan kamroq ekanligini ko'ramiz. Buni quyidagi 2 ta sabab bilan izohlash mumkin:

N-gram taqsimotidagi farq: “O’zbek tili matnlari uchun unigram til modelini ishlab chiqish muammo va yechinlar” deb nomlanuvchi maqolamizdan biz bilamizki, model yaxshi ishlashi uchun o'quv matnining n-gram taqsimoti va baholash matni bir-biriga o'xshash bo'lishi kerak. Shu nuqtai nazardan, **uzb_text2** matni **uzb_text1** matniga qaraganda yomonroq ishlashi mantiqan to'g'ri. Buni quyida **"va"** so'zi bilan boshlanadigan bigramlar uchun taqsimotda ko'rish mumkin:



Yuqoridagi grafikdan **"va"** so`zi bilan boshlanadigan bigramaning ehtimollik taqsimoti o`quv va **uzb_text1** matni o'rtasida taxminan o'xshashligini qayd etish mumkin. Chunki ikkala kitobda ham umumiy ("Alisher Navoiy" kabi) so`z birikmalari mavjud. Bundan farqli o'laroq, **uzb_text2** matnining taqsimoti o`quv matnidan juda farq qiladi: Pirimqul Qodirovning “Avlodlar dovoni” asarida "alisher navoiy" bogramasi yo'qligi aniq.

– **Noma'lum n-gramlar**: o`quv va **uzb_text1** matni ikki xil vaqt, janr va mualliflarning ikkita kitobi bo'lganligi sababli, **uzb_text1** matnida o`quv matnida mavjud bo`lmagan ko'plab n-gramlarni o'z ichiga oladi. Ushbu "noma'lum" n-gramning (**uzb_text1** va **uzb_text2**) ulushiga nisbatan baholash matnining o'rtacha logarifmik ehtimolining taqqosi quyida keltirilgan grafikda keltirilgan (har bir ma'lumot nuqtasi ustidagi raqam modelning n-gram uzunligini ifodalaydi):

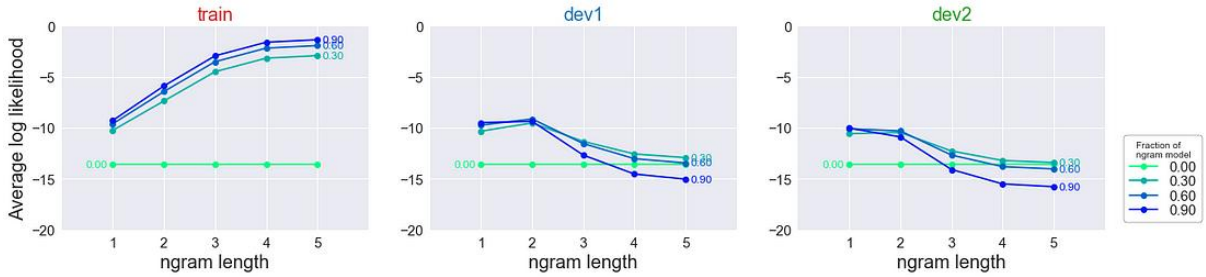


– Noma'lum n-gramlarning ulushi va o'rtacha logarifmik ehtimoli o'rtasida kuchli manfiy korrelyatsiya mavjud. Ayniqsa trigram, 4 gram va 5 gram kabi yuqori n-gram modellari uchun ushbu tasdiqni aniq ko`rish mumkin.

– Berilgan n-gram modeli uchun **uzb_text2** matni har doim **uzb_text1** matniga qaraganda noma’lum n-gramning katta qismiga ega bo`ladi. Demak, **uzb_text2** matnining odatda **uzb_text1** matniga qaraganda o'rtacha logarifmik ehtimolligi past bo'lishi ajablanarli emas.

Model interpolyatsiyasining ta'siri

Ushbu kuzatishlar bo'yicha umumiy jihat shundan iboratki, baholash matnidan (**uzb_text1** va **uzb_text2**) va n-gram modelidan qat'i nazar (unigramdan 5 gramgacha), odatda modelni biroz uniform model bilan interpolyatsiya qilish modelning o'rtacha logarifmik ehtimolligini yaxshilaydi.



Ushbu interpolyatsiyaning ta'siri “O’zbek tili matnlari uchun unigram til modelini ishlab chiqish muammo va yechinlar” deb nomlanuvchi maqolada batafsilroq tasvirlangan, xususan:

1. Uniform model bilan interpolyatsiya qilish noma'lum n-gramlarga kichik ehtimollik beradi va modelning 0 ehtimollik bilan n-gramga ega bo'lishini bartaraf qiladi. Bu esa interpolyatsiya amalining ayniqsa yuqori n-gramli modellar (trigram, 4 gram, 5 gram) uchun foydali ekanligini tushuntiradi: **yuqori n-gram modellar o'quv matnida mavjud bo`lmagan juda ko'p noma'lum n-gramlarga duch keladi.**
2. Uniform model bilan interpolyatsiya qilish o'quv matniga modelning haddan tashqari mos kelishini kamaytiradi. Albatta, o`quv matni uchun grafikda aniq ko'rinib

turganidek, modelning ishlashi yomonlashadi. Biroq, model **uzb_text1** va **uzb_text2** **b**aholash matnlari uchun grafiklarda ko'rsatilganidek, yangi matnlarga yaxshiroq umumlashtirishi mumkin.

Xulosa

Ushbu maqolada tabiiy tillarni qayta ishlashda hali ham qo'llaniladigan muhim mashinali o'qitish tamoyilini ta'kidlandi: *o'quv ma'lumotlari kichik hajmli bo'lsa, murakkabroq model yanada yomonroq bo'lishi mumkin!*
N-gram modellari uchun bu muammo **siyraklik (sparsity)** muammosi deb ham ataladi. Chunki o'quv matni qanchalik katta bo'lmasin, undagi n-gramlar o`zbek tilidagi n-gramlarning cheksiz ko'rinadigan o'zgarishlarini hech qachon qamrab olmaydi. Boshqacha qilib aytganda, ko'p n-gramlar modelga "noma'lum" bo'ladi va n-gram qanchalik katta bo'lsa, muammo yanada ortadi.
Keyingi ilmiy izlanishlarda ushbu n-gram modelini yaxshilash usullari bo`yicha tadqiqotlar olib boriladi. Misol uchun, har bir n-gramli modelni uniform model bilan interpolyatsiya qilish o'rniga, biz barcha n-gram modellarini birlashtirish mumkin. Biz **expectation-maximization** algoritmidan foydalanib, ushbu modellarning kombinatsiyalangan og'irliklarini yanada optimallashtirishimiz mumkin.

Foydalanilgan adabiyotlar

1. Booch, G., Jacobson, I., & Rumbaugh, J. (1996). The unified modeling language. *Unix Review*, *14*(13), 5.
2. Tan, M., Zhou, W., Zheng, L., & Wang, S. (2012). A scalable distributed syntactic, semantic, and lexical language model. *Computational Linguistics*, *38*(3), 631-671.
3. Brown, P. F., Della Pietra, V. J., Desouza, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational linguistics*, *18*(4), 467-480.
4. Hockey, B. A., & Rayner, M. (2005, July). Comparison of grammar-based and statistical language models trained on the same data. In *Proceedings of the AAAI Workshop on Spoken Language Understanding* (pp. 9-10).
5. Mikolov, T. (2012). Statistical language models based on neural networks.
6. Kim, Y., Jernite, Y., Sontag, D., & Rush, A. (2016, March). Character-aware neural language models. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 30, No. 1).
7. B.Elov, Sh.Sirojiddinov, Sh.Hamroyeva, E.Adalı, Z.Xusainova (2023, September). Pos Taging of Uzbek Text Using Hidden Markov Model. In *2023 8th International Conference on Computer Science and Engineering (UBMK)* (pp. 63-68). IEEE.
8. B.Elov, Sh.Khamroeva, R.Alayev, Z.Khusainova, U.Yodgorov (2023). Methods of processing the uzbek language corpus texts. *International Journal of Open Information Technologies*, *11*(12), 143-151.
9. Liu, Y., & Zhang, M. (2018). Neural network methods for natural language processing.
10. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1-45.
11. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., ... & Zaremba, W. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
12. Chen, S. F., Beeferman, D., & Rosenfeld, R. (1998). Evaluation metrics for language models.
13. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1-45.
14. B.Elov (2022). N-gram til modellari vositasida o 'zbek tilida matn generatsiya qilish. Computer linguistics: problems, solutions, prospects, 1(1).
15. B. Elov, A. Abdullayev, A., N.Xudoyberganov. (2024). O'zbek tili korpusi matnlari asosida til modellarini yaratish. «Contemporary technologies of computational linguistics», 2(22.04), 344-353