N-GRAMMGA ASOSLANGAN STEMMING ALGORITMINI ISHLAB CHIQISH, TATBIQ QILISH VA BAHOLASH

Qosimova Sarvinoz Furqat qizi
sarvinozq54@gmail.com
ToshDO‘TAU magistranti

Annotatsiya. Bugungi kunda tabiiy tilga bog‘liq bo‘lmagan stemmerlarni ishlab chiqish muhim ahamiyat kasb etadi. Tokenizatsiya jarayonining N-gramm usuli ba’zi hollarda stemlarni noto‘g‘ri aniqlashi mumkin. Shu sababli, ushbu N-gramm usuliga biroz o‘zgartirishlar kiritish orqali, yangi stemmer ishlab chiqildi. Ushbu maqola n-grammlardan foydalangan holda innovatsion stemming algoritmini ishlab chiqish, amalga oshirish va baholashni taqdim etadi. Puxta kodlash va sinchkovlik bilan baholash orqali algoritmning samaradorligi baholanadi va uning tabiiy tilni qayta ishlash vazifalariga qo‘shadigan potensial hissasi haqida qimmatli tushunchalarni taqdim etadi.

Abstract. Today, the development of stemmers that do not depend on natural language is of great importance. The N-gram method of the tokenization process can sometimes incorrectly identify stems. Therefore, by making some modifications to this N-gram method, a new stemmer was developed. This paper presents the development, implementation, and evaluation of an innovative stemming algorithm using n-grams. Through careful coding and careful evaluation, the performance of the algorithm is evaluated and provides valuable insights into its potential contribution to natural language processing tasks.

Аннотация. Сегодня большое значение имеет разработка стеммеров, не зависящих от естественного языка. N-граммный метод процесса токенизации иногда может неправильно идентифицировать основы. Поэтому, внеся некоторые изменения в этот метод N-грамм, был разработан новый стеммер. В этой статье представлены разработка, реализация и оценка инновационного алгоритма стемминга с использованием n-грамм. Благодаря тщательному кодированию и тщательной оценке оценивается производительность алгоритма и предоставляется ценная информация о его потенциальном вкладе в задачи обработки естественного языка.

Kalit so‘zlar: *Axborot olish (IR), Stemming, Stemmer, N-gramm, statistik metod, adekvat, tokenizatsiya, algoritm, gipoteza.*

Kirish

Bugungi kunda zamonaviy qidiruv va indeksatsiya tizimlari samaradorligini oshirish uchun **stemming jarayonidan** foydalanadi. Stemming jarayoni lemmatizatsiya jarayoniga o‘xshash bo‘lib, so‘zning morfologik variantlarini bir shaklga keltirish amallarini bajarishga asoslangan. Qidiruv tizimlari hujjatlarni indekslash jarayonida matn so‘zlaridagi qo‘shimchalar (affiks/prefikslar)ni

qisqartirishni bajradi. Natijada hosil qilingan o‘zaklar to‘plami IR tizimida olingan hujjatlar sonini oshiradi. Hozirda mavjud stemmerlar aniqlik ko‘rsatkichi to‘liq natijalarni qaytarmasligi ba’zi olimlar tomonidan qayd etilsa-da, zamonaviy IR tizimlari bugungi kunda ham stemmingdan muhim vosita sifatida foydalanadi.

Stemming jarayoni ustida bir qator ilmiy izlanishlar olib borilgan va muhim natijalar olingan. Jumladan, **qoidalariga asoslangan yondashuvli** stemmerlarni ishlab chiqish uchun til morfologiyasi bo‘yicha zarur bilimlar talab qilinadi. **Til korpusining statistik tahlili yondashuviga** asoslangan stemmerlar muayyan til uchun yaxshi natijalarni qaytaradi. Biroq ushbu lingvistik stemmerlar bitta katta kamchilikka ega, ular ko‘p tilli matnlarni tahlil qila olmaydi. *Belgilar chastotalariga asoslangan statistik algoritmlar* ushbu kamchilikni bartaraf qilishi mumkin.

Statistik metodlarda bizning e’tiborimizni tortgan eng qiziqarli va samarali usul Jeyms Mayfild va Pol MakNamining *Single N-gram Stemming* [Mayfield, 2003]. Mualliflar N belgining bir-biriga o‘xshash ketma-ketliklaridan foydalanib, stemming foydasiga tildan mustaqil ravishda erishish mumkin, deb taxmin qilishdi. Ular so‘zlardan olingan, N-grammlarning ba’zilari (aslida, kam tez-tez) so‘zning morfologik o‘zgarishlarni ko‘rsatmaydigan qismlari uchun umumiy bo‘lishini, ya’ni ular aynan bir xil o‘zak o‘rnini bosuvchi bo‘lib xizmat qilishi mumkinligini ta’kidladilar. Ular so‘zning “*pseudo stem”*(soxta o‘zak) sifatida *Inverse Document Frequency* (IDF) bilan ichki N-gramm so‘zini tanlashni taklif qilishdi. Ular affikslar bilan namoyon bo‘ladigan ma’lum bir morfologik o‘zgarish ko‘plab turli so‘zlarda takrorlanishi shuning uchun ham past IDF ko‘rsatishi haqida ajoyib asoslar berdi.. Masalan, ular o‘zlarining kontsepsiyalariga ko‘ra, “juggling” so‘zining (CLEF 2002 to‘plamidan) va taklif qilingan psevdo-4 va psevdo-5 so‘zlarining turli N-grammlarining hujjat chastotalarini jadvalga kiritdilar.

Bizning hozirgi tadqiqotimizning kaliti yuqorida aytib o‘tilgan jadvallarda joylashgan. Currency va warrens so‘zlari uchun psevdo-4 o‘zak mos ravishda *rren* va *rens* edi. Juggler, currency va warrens uchun psevdo-5 poyasi mos ravishda *ggler*, *rency* va *rrens* edi. Umumiy tushunchaga ko‘ra,o‘zak boshqa joydan emas, balki so‘zning boshlang‘ich belgisidan boshlanishi kerakligini aytadi. Xo‘sh, bu allaqachon oqlangan usulni takomillashtirish imkoniyati bormi?

Yuqoridagi savolga javob berishga harakat qilish bizning maqolamizning mohiyatidir. Biz n-grammlarni yaxshiroq tatbiq qilish uchun faoliyat olib boramiz, ularning chastotalarini hisoblash g‘oyasini saqlab qolamiz. IDFning oddiy kontsepsiyasi o‘rniga, biz o‘zaklarni yaxshiroq taxmin qilish imkonini beradigan yangi algoritmni ishlab chiqamiz.

2. Tavsiya etilgan usul:

Mualliflar [Mayfield, 2003] o‘zlarining eksperimental natijalariga ko‘ra, “Snowball stems [Porter, 2001], psevdo stems va 4-gramm so‘zlaridan eng yaxshi natija ko‘rsatish tendensiyasi 4 grammda kuzatilgan. Ular 4 grammlar, har qanday so‘zning boshlang‘ich 4 ta belgisi so‘z o‘zagining eng yaxshi ifodasidir degan xulosaga kelishdi. Bu bizning to‘g‘ri (qat’iy lingvistik nuqtayi nazardan emas) va

ajoyib taxmin qilingan o‘zak uchun u boshidan boshlanishi kerakligi haqidagi tushunchamizni mustahkamlaydi. Xo‘sh, bir-biriga o‘xshash N ta belgining chastotalarini (ya’ni, ichki N-gramm so‘zlari) topish o‘rniga, 4 gramm, 5 gramm, 6 gramm va hokazo chastotalardan kelib chiqqan taxminiy asosni topsak nima bo‘ladi? Istiqbolli yo‘nalish ko‘rinadi. Bizning fikrimiz, o‘zakni topish uchun dastlabki taxminimiz sifatida 4 grammni olishdir. Shubhasiz, bu uzunligi 3 yoki undan kam bo‘lgan so‘zlarni istisno qiladi. Bu ajoyib afzallik bo‘ladi, chunki to‘xtash so‘zlarining ko‘pchiligi 3 yoki undan kam uzunlikda. Endi ma’lum bir so‘z uchun bizda 4-gramm, 5-gramm va shunga o‘xshash chastotalarni topish va ulardan eng yaxshisini tanlash qoladi. Yondashuvimizni boshlashdan oldin, keling, COCAning turli N-gramm chastotalarini ko‘rib chiqaylik [COCA, 2014] Biz xuddi shu jungling so‘zidan boshlaymiz.

Jadval 2.1: “Joggling” so‘zi uchun COCA dan turli N-gramm chastotalari

N-gram	*Chastotasi*	*N-gram*	*Chastotasi*
jugg*	915	jugglin*	328
juggl*	729	juggling*	328
juggli*	328		

Biz o‘zakning adekvat (aynan bir xil) yondashuvi bo‘lib xizmat qiladigan so‘zning N-grammasi korpus orqali nisbatan yaxshi chastotaga ega bo‘lishi kerak degan xulosaga keldik. Bundan kam N-grammlarning barchasi tez-tez bo‘lib va undan yuqori chastotalar esa asta-sekin pastroq bo‘lib, berilgan so‘zning umumiy chastotasi bilan tugaydi (ikki yoki undan ortiq ketma-ket yozuvlar bir xil bo‘lishi mumkin).

Yuqoridagi jadval bizning taxminimizga muvaffaqiyatli javob beradi. Shunday qilib, bizda qisman so‘z uzunligi (N-gram) bir birlikka ko‘payishi bilan qiymatlari pasayishga (yoki doimiy bo‘lib qolishga) moyil bo‘lgan butun son o‘zgaruvchiga mavjud. So‘zning (o‘zakning) o‘zgarmas qismidan qo‘shimcha qismiga o‘tishda bu pasayish juda keskin bo‘lishi ehtimoldan holi emas. Ammo nisbiy maksimal tushishni qanday ushlay olamiz? Buning bir usuli - har ikki ketma-ket yozuvning qaytishini olishdir. Ko‘rsatkichni quyidagi tarzda belgilaymiz

(2.1)

$$\lambda_i = \begin{cases} abs(F_i - F_{i-1}) & if\ i > 4 \\ 0 & else \end{cases}$$

Bu yerda Fi - i^{th} chastotasi

N-gram. Keyin biz ikkinchi tartibli og‘ishlarni hisoblaymiz.

(2.2)

$$\Delta_i = \lambda_i - \lambda_{i-1}$$

N uzunlikdagi N-grammga mos keladigan o‘zgaruvchi bo‘lsa. Biz protseduramizni ikki bosqichda aniqlaymiz:

1-bosqich:

1) Initializatsiya:

bu erda M - juda yuqori butun qiymat

$$\psi_N = 4$$

2) Rekursiya:

Calculate λ_i, $4 < i \leq |word|$

If $\lambda_i > \Gamma$

$$\psi_N = \underset{4<i\leq|word|}{argmax}[F_i, F_{i-1}]$$

Else

$$\psi_N = i$$

3) Tugatish:

a.) If N=|word|, stop. Else, calculate Δ_i.

b.) If Δ_i>0, stop.

Bu yerda Γ - chegara qiymati bo‘lib, u ikki N-gramm chastotali burilishning qanchalik maqbulligini belgilaydi. U nolga yoki minimal qiymatga saqlanadi. Rekursiya bosqichi kiritmoq so‘zining ortib borayotgan N-grammlari orqali aylanishni amalga oshiradi. Agar 1-bosqich oxirida $\psi_N = |W|$ bo‘lsa, 2-bosqichga o‘ting.

2-bosqich:

Agar oxirgi uchta N-grammning chastotalari ψ_N bir xil bo‘lsa, kiritilgan so‘zning oxirgi uchta belgisini o‘chiring, agar ularni o‘chirish natijasida hosil bo‘lgan asosiy uchdan katta bo‘lsa, ya’ni

$$\psi_N = \psi_N - 3 \qquad \text{If } \psi_N - 3 > 3$$

Yuqoridagi protsedura oxirida ψ_N ortiqcha talab qilingan o‘zakka yaqin bo‘lgan N gramm qiymatini o‘z ichiga oladi. 2-bosqich umumiy qo‘shimchalarni (-ing kabi) ko‘rsatadigan so‘zlarni boshqaradi, agar ular 1-bosqichda ajratilmagan bo‘lsa.

3. *Illyustratsiyalar va tajribalar:*

Keling, o‘z uslubimizni ingliz, ispan va portugal so‘zlari klasteriga qo‘llaylik. Bizga kerak bo‘lgan narsa - bu boy korpusdan N-grammlarning ketma-ket chastotalari. Ingliz tili uchun biz COCA ga tayandik [COCA. 2014]. Ushbu korpusning ikkita afzalligi bor, birinchidan u inglizcha so‘zlarning boy to‘plamiga ega, ikkinchidan, wild card [*] yordamida N -gram chastotalarini hisoblash oson. Ispan va portugal so‘zlari uchun ketma-ket chastotalar mos ravishda CorpusDelEspanol [Corpus Del Español, 2014] va Corpus Do Português [Corpus Do Português, 2014] dan olingan. Ba’zi nomzodlar bo‘yicha test natijalari quyida keltirilgan:

3.1-jadval. So‘z turkumlari va o‘zaklari

Klaster raqami.	So‘z	O‘zak	Klaster raqami.	So‘z	O‘zak
1 (Inglizcha)	create	creat	2 (Ispancha)	trabajan	traba
	creates	creat		trabajar	traba
	creating	creat		trabajado	traba
	created	creat		trabajador	traba
	creation	creat	3 (Portugalcha)	dificil	dific
	creative	creat		dificilmente	dific

Keling, bizning stemmerimizni empirik tarzda (tajribaga asoslangan holda) o‘rnatilgan lingvistik stemmer bilan taqqoslaylik, deydi Porterning stemmeri [Porter, 1980]. Sinov tili sifatida biz ingliz tilini tanlaymiz, chunki Porter algoritmining natijalarini Snowball loyihasida osongina olish mumkin [Porter, 2001]. Biz 100 ta tasodifiy inglizcha so‘zlardan namuna olamiz va Porter’s Snowball tizimi va ular ustidagi protseduramizni qo‘llaymiz. Natijalar 3.2-jadvalda keltirilgan. Ikkala algoritm tomonidan yaratilgan bir xil ildizlar qalin qilib ko‘rsatilgan. To‘rtta POS, ot, sifat, fe’l va qo‘shimchalar hisobga olindi. Biz COCAda gazetalarni qidirish maydonini tanlaymiz.

Snowball (Porter) stemmer va bizning N-gram stemmerning natijalarini solishtirish uchun biz to‘g‘ridan-to‘g‘ri baholash usulidan foydalandik. Chris D. Paise [Paice, 1994] tomonidan amalga oshirilgan taqqoslash kabi, uzunlikni qisqartirishni asosiy chiziq sifatida ishlatib, biz Levenshteyn masofasidan [Levenshteyn, 1966] asosiy o‘lchov sifatida foydalandik. Masofa - bu manba qatorini maqsadli satrga aylantirish uchun zarur bo‘lgan o‘chirishlar, qo‘shishlar yoki almashtirishlar soni. Shunday qilib, Levenshteyn masofasi so‘zning stemmer tomonidan qancha birlikdan ajratilganligini ko‘rsatadi.

Biz faraz qilamizki, agar bizning stemmerimiz Porterniki kabi samarali bo‘lsa, u holda Levenshteyn masofalarining so‘z va uning ajratilgan o‘zak o‘rtasidagi taqsimoti bir xil bo‘ladi. Berilgan so‘z uchun bizda ikkita to‘g‘rilash usuli bor, Porter algoritmi va N-gramm protseduramiz. Shunday qilib, berilgan so‘z uchun bizda bir juft LD mavjud. Taqqoslash uchun biz H0 nol gipotezasini o‘rnatdik, bu ikkita stemmer o‘rtasida hech qanday farq yo‘q.

Ikki chora-tadbirlar to‘plamini bir xil namunaga bog‘langan ikkita o‘lchov sifatida ko‘rib chiqish mumkinligi sababli, biz juftlashtirilgan namunalar uchun statistik testdan foydalanishga qaror qildik.

Xususan, parametrik bo‘lmagan statistik test, Wilcoxon imzolangan darajali test bizda qo‘llaniladi, chunki bu masofalarning taqsimlanishi haqida hech qanday dalil yo‘q edi.

Wilcoxon testi N kuzatilgan qiymatlarning ikkita juftlashgan seriyasiga (xi, yi) asoslanadi, taqqoslanadigan kuzatilgan X, Y o‘zgaruvchilarning har biri uchun bir qator. Uilkoxon testi N kuzatilgan qiymatlarning ikkita juftlashgan seriyasiga (xi, yi) asoslanadi, taqqoslanadigan X va kuzatilgan Y o‘zgaruvchilarning har biri uchun bir qator. Mutloq farqlarning di=abs (xi-yi) va ishorasi (di) mutlaq farqlari hisoblanishi kerak. Keyin mutlaq qiymatlar nollardan voz kechib tartiblanadi.

Agar namuna hajmi kichraytirilgan bo‘lsa, testning yakuniy statistikasi katta N uchun Oddiy o‘zgaruvchiga yaqinlashtiriladi.

4. *Eksperimental natijalar:*

Bizning namunaviy ma’lumotlarimiz bo‘yicha Wilcoxon testini o‘tkazganimizda, biz p qiymatini olamiz 0,54. p>0,05 sifatida, H0 nol gipotezasini rad eta olmaymiz. Shunday qilib, bizning N-gramm Stemmerimiz Porter’s Stemmerdan kam emasligi haqida kuchli dalillarga egamiz.

Xulosa

Biz statistik stemming texnikasini o‘zgartirishga harakat qildik va natijalarni ancha tajribali lingvistik stemmer bilan solishtirish mumkin bo‘lgan usulni ishlab chiqdik.

Bizning N-gram stemmerimiz N-gramm korpus chastotalari mavjud bo‘lgan har qanday tilda ishlashga qodir.

Tilning neytral yondashuvi har doim lingvistik bilim yoki tahlilning zaruriy sharti bo‘lgan texnikadan afzalroq bo‘lganligi sababli, bizning natijalarimiz istiqbolli ko‘rinadi.

Foydalanilgan adabiyotlar:

1. Elov B., Hamroyeva Sh., Abdullayeva O., Xusainova Z., Xudayberganov N. / O‘zbek, turk va uyg‘ur tillarida POS teglash va stemming. O‘zbekiston: til va madaniyat 2023/1. – P. 40-64
2. Elov B., Alayev R., Xusainova Z. O‘zbek tilida stemmingni amalga oshirishning gibrid statistik yondashuvi. 2023
3. B.B.Elov, Sh.M.Khamroeva, Z.Y.Khusainova. Pipeline conveyer of NLP (natural language processing). Descendants of Muhammad al-Khwarazmi. Scientific-practical and information – analytical Journal, 1 (23) / 2023, 181-192 pp.
4. Xusainova Z.Y. Tabiiy tilni qayta ishlash (NLP)da stemming jarayoni tavsifi / BuxDU ilmiy axboroti 3/2023. – B. 113-119
5. Corpus Del Español < http://www.corpusdelespanol.org> manzilida, 2014-yil 13-yanvarda tashrif buyurilgan.
6. Corpus Do Português. <http://www.corpusdoportugues.org/> manzilida, 2014-yil 13-yanvarda tashrif buyurilgan.
7. Zamonaviy Amerika ingliz tili korpusi (COCA). <http://corpus.byu.edu/coca/> manzilida, 2014-yil 13-yanvarda tashrif buyurilgan.

8. Douson Jon (1974). Qo‘shimchalarni olib tashlash va so‘zlarni birlashtirish. ALLC byulleteni, 2-jild, № 3, 33-46.
9. Funchun Peng, Nawaaz Axmed, Xin Li and Yumao Lu (2007) Veb-qidiruv uchun kontekstga sezgir. ACM SIGIR 30-chi yillik xalqaro konferentsiyasi ma’lumotlarini qidirishda tadqiqot va rivojlantirish bo‘yicha ma’ruzalar, 639-646
10. Hafer M. va S. Weiss (1974). Harf o‘rinbosar navlari bo‘yicha so‘zlarni segmentatsiyalash. Ma’lumotni saqlash va qidirish, 10, 371-85.
11. Harman Donna (1987). Onlayn muhitda qo‘shimcha qo‘shish cheklovlari bo‘yicha muvaffaqiyatsizlik tahlili. Axborot qidirishda tadqiqot va ishlanmalar bo‘yicha 10-yillik xalqaro ACM SIGIR konferentsiyasi materiallari, 102-107.
12. Harman Donna (1991). Qo‘shimchalar qanchalik samarali? Journal of the American Society for Information Science, 42, 7-15.
13. Krovetz Robert (1993). Morfologiyani xulosa chiqarish jarayoni sifatida ko‘rib chiqish. Axborot qidirishda tadqiqot va ishlanmalar bo‘yicha 16-yillik ACM SIGIR konferentsiyasi materiallari, 191-202.
14. Levenshtein V. I. (1966). O‘chirish, kiritish va teskari o‘zgarishlarni tuzatishga qodir ikkilik kodlar. Sovet fizikasi- Doklady, 10(8),707710.
15. Lovins, J.B. (1968). Striping algoritmini ishlab chiqish. Mexanik tarjima va hisoblash tilshunosligi, 11, 22-31.
16. Majumder Prasenjit, va boshq. (2007). YASS: Yana bir qo‘shimcha striptizator. Axborot tizimlari bo‘yicha ACM operatsiyalari. 25(4), №18-modda.
17. Mayfield Jeyms va McNamee Paul (2003). Yagona N-gramm stemming. Axborot qidirishda tadqiqot va ishlanmalar bo‘yicha 26-yillik xalqaro ACM SIGIR konferentsiyasi materiallari, 415-416.
18. Melucci Massimo va Orio Nicola (2007). Avtomatik Stemmer yaratish metodologiyasini loyihalash, amalga oshirish va baholash. Amerika axborot fanlari va texnologiyalari jamiyati jurnali, 58(5), 673–686.