MATNLARNING O‘XSHASHLIGINI TAHLIL QILISH

Bozorov Obid
Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universiteti
bozorov.obid@gmail.com

Annotatsiya. Ushbu ishda harflar chastotalarini matnlar tahlilida qo‘llanilishi tadqiq qilingan. Chastotali tahlil usuli yordamida o‘zbek tilidagi matnlarni moslik indeksi aniqlangan. Shuningdek, ochiq matnlarning moslik indeksini aniqlashning algoritmi taklif etilgan.
Kalit so‘zlar: *chastotali tahlil, harflar chastotasi, matematik model, moslik indeksi, o‘zaro moslik, tabiiy til.*

Abstract. In this work, the use of letter frequencies in text analysis is studied. Using the method of frequency analysis, the compatibility index of Uzbek texts was determined. Also, an algorithm for determining the compatibility index of open texts is proposed compatibility index of open texts has also been developed and implemented programmatically.
Key words: *frequency analysis, letter frequency, mathematical model, compatibility index, interchangeability, natural language.*

Аннотация. В данной работе изучается использование частот букв в анализе текста. Методом частотного анализа определен индекс сочетаемости узбекских текстов. Также предложен алгоритм определения индекса совместимости открытых текстов. Индекс совместимости открытых текстов также разработан и реализован программно.
Ключевые слова: *частотный анализ, частотность букв, математическая модель, индекс совместимости, взаимозаменяемость, естественный язык.*

Hozirda axborot himoyasi tizimlarini ishlab chiqishga va ularni takomillashtirishga alohida e’tibor qaratilmoqda. Tashkilotlarning axborot tizimlarida davlat sirlari yoki qonun bilan qo‘riqlanadigan boshqa sirlarni tashkil etuvchi ma’lumotlar, shuningdek shaxsga doir ma’lumotlar muhofaza qilinishi, ulardan ruxsatsiz foydalanishning oldini olinishini ta’minlash hamda konfidensial ma’lumotlarning noqonuniy chiqib ketmasligi yuzasidan O‘zbekiston Respublikasi qonun va hujjatlarida zarur dasturiy va tashkiliy-texnik choralar ko‘rilishi shart qilib belgilangan. Shularni hisobga olib, davlat tashkilotlarida konfidensial axborotlarni ruxsatsiz chiqib ketishini bashoratlash va bartaraf qilishni hozirgi kunning dolzarb muammolaridan biri sifatida e’tirof etish mumkin [1-2].

Harflar chastotalari va ularning matnlar tahlilida qo‘llanilishi. Matndagi harflar va so‘zlarning chastotasi matn muallifi va matnning tegishli sohasi bilan bir-

biridan farqlanadi. Shuningdek, mualliflik ishlarini aniqlashda muallifning yozish usullari tahlili yordamida harf, bigram, trigram, so‘z chastotalari, so‘z uzunligi va gap uzunliklarini hisoblay olishlari mumkin va bu usul bilan matnlarning mualliflik huquqini himoyalash uchun ham ishlatishlari mumkin bo‘ladi. O‘rtacha harflar chastotasini ko‘plab belgili matnlar tahlili orqali hisoblash mumkin. Zamonaviy hisoblash tizimlaridan foydalanish orqali katta hajmli matnlar to‘plamlarini tahlil qilgan holda kerakli aniqlikdagi natijani olish qiyin hisoblanmaydi.

Faraz qilaylik, $A=(a_1,a_2,...,a_m)$ alifbo belgilari Z_m halqadan olingan $0,1,2,...,m-1$ raqamlar bilan raqamlangan bo‘lsin. Shuningdek, N ta A alifbo belgilardandan iborat $\bar{X}=(y_1,y_2,...,y_n)$ matn berilgan bo‘lsin.

Ta’rif. $\bar{X}$ – alifbo belgilaridan tashkil topgan matn bo‘lsin. $\bar{X}$ – matnning moslik indeksi deb:

$$I(\vec{X})=\sum_{i=1}^{m}\frac{f_i(f_i-1)}{n(n-1)} \qquad (1)$$

tenglik bilan aniqlangan qiymatga aytiladi [9].

Bunda, $\vec{X}$ –tahlil qilingan ochiq matnlar to‘plami;

f_i – alifboning $i-$ belgisining matndagi uchrashlari soni;

n – matndagi barcha belgilar soni;

m – alifbodagi belgilar soni.

C# dasturlash tilida yaratilgan dastur yordamida 121 millionga yaqin belgilardan iborat o‘zbek tilidagi ochiq matnlarni (1) hisoblash formulasi yordamida tahlil qilib chiqildi va quyidagi natijalar olindi. O‘zbek tilining moslik indeksi $I_{o'zbek}=0,071698945$ qiymatiga teng bo‘ldi. Ushbu natijani ochiq manbaalardagi boshqa tillarning moslik indeksi bilan taqqoslanganda quyidagicha ko‘rinish oldi (1-jadval).

1-jadval

Til	Rus tili	Ingliz tili	**O‘zbek tili**	Italiyan tili	nemis tili	Fransuz tili
Moslik indeksi	0.0553	0.0644	**0.071698945**	0.0738	0.0762	0.0778

Hisoblash natijalaridan shuni aniqlash mumkinki o‘rganilayotgan matni qaysi tilga oidligini avtomatik ravishda aniqlash uchun uchun o‘rganilayotgan matnning belgilari chastotalari aniqlash talab etiladi.

Berilgan matnning qaysi tilga yaqinligini aniqlash uchun quyidagi qiymat hisoblanadi:

$$X=\sum_{i=0}^{m}n_ik_i \qquad (2)$$

bunda: n_i – tabiiy tilning i – harfining chastotasi, k_i – o‘rganilayotgan matnning i – harfining chastotasi, m – alifbodagi belgilar soni.

Ushbu usul orqali matnning qaysi tildaligini aniqlash uchun matn uzunligi yetarlicha katta bo‘lishi kerak. Agar, o‘rganilayotgan matn uzunligi qisqa bo‘lsa matnning chastota xarakteristikalari tabiiy tilning xususiyatlaridan farq qilishi mumkin.

Amaliyotda konfidensial axborotlarni ruxsatsiz chiqib ketishini oldini olish tizimini aylanib o‘tish uchun matndagi so‘zlar va belgilarning o‘rnini o‘zgartirgan holda olib chiqishga urinishlar bo‘ladi [3].

Faraz qilaylik, berilgan $A=(a_1,a_2,...,a_m)$ alifbo belgilaridan tashkil topgan $\bar{X}=(y_1,y_2,...,y_n)$ va $\bar{Y}=(y'_1,y'_2,...,y'_{n'})$ matnlarning qaysi tilga yoki sohaga yaqinligini aniqlash uchun matnlarning o‘zor moslik indeksi hisoblanadi:

$$MI(\vec{X},\vec{Y})=\frac{\sum_{i=1}^{m} f_i f_i'}{nn'} \qquad (3)$$

bunda, $\vec{X}$, $\vec{Y}$ – solishtirilayotgan matnlarning belgilari to‘plamlari;

f_i, f_i' – solishtirilayotgan matnlarning i – belgisining matnlardagi uchrashlari soni;

n,n' – solishtiralayotgan matnlardagi barcha belgilar soni;

m – alifbodagi belgilar soni.

Chastotali tahlil usuli yordamida o‘zbek tilidagi matnlarni moslik indeksi hisoblab chiqilishi axborot tizimlarida mavjud matnlarning o‘zbek tiliga oidligini avtomatik aniqlashning imkonini beradi. Shuningdek, konfidensial axborotlarni kiberxujumchi tomonidan matndagi so‘zlar yoki belgilarning o‘rnini almashtirish orqali ruxsatsiz olib chiqib ketilishini oldini olishga xizmat qiladi. Shuningdek, ochiq matnlarning o‘xshashligini aniqlashning algoritmi yordamida ishlab chiqilgan dasturiy ta’minot matnlarni solishtirish mumkin bo‘ladi. Matnlarning o‘zaro o‘xshashligini aniqlash tizimini yaratilishi va uning amaliyotga joriy qilinishi milliy axborot resurslarini ishonchli himoyasini ta’minlaydi.

Foydalanilgan adabiyotlar:

1. Jo‘rayev G.U., Bozorov O.N. «Elektron hukumat» sharoitida O‘zbekiston Respublikasida konfedensial axborotlarni himoyalash muammolari va yechimlari// Toshkent shahri, 2019 yil 12-aprel. 108-111 b.

2. «Shaxsga doir ma’lumotlar to‘g‘risida"gi O‘zbekiston Respublikasining Qonuni. -T., 2019 yil 2-iyul, O‘RQ–547-son.

3. Radwan R. Tahboub, Yousef Saleh. Data Leakage/Loss Prevention Systems (DLP) NNGT Journal: International Journal of Information Systems. Volume 1, 2014. –P. 13-19.