Лабораторна робота №2
«Статистичні критерії на відкритий текст»

Пясецький Б.

Мета роботи: засвоєння статистичних методів розрізнення змістовного тексту від випадкової послідовності, порівняння їх, визначення похибок першого та другого роду.

1 Хід роботи

1. На великому тексті українською мовою, необхідно розрахувати частоти літер та біграм, а також ентропію та індекс відповідності (у тексті має бути замінена літера ґ на г; видалений символ апострофу та інші спецсимволи в тексті, включно з пробілами; великі літери замінені на малі);

2. Отримати N текстів X українською мовою для довжин ($L = 10, 100, 1000, 10000$), для кожного з яких згенерувати спотворені тексти Y. Число N, та методи спотворення текстів подані в методичці;

3. Реалізувати критерії 2.0-2.3 (критерій частих l-грам та його варіації), 4.0 (критерій розрахунку індексу відповідності) та 5.0 (критерій порожніх ящиків) та структурний. Перевірити роботу на згенерованих N текстах для кожної довжини L. Розрахувати ймовірність похибок першого та другого роду.

1.1 Множини заборонених та частих символів

Заборонені біграми	Часті біграми
аь	на
бж	ли
бф	ро
бщ	ві
бь	ов
вь	по
гє	ав
гж	ал
гь	го
гю	не
дє	ти
еи	до
еь	од
єє	ст

1.2 Результати способів спотворення

Результати застосування критеріїв до спотворених текстів та частин відкритого тексту знаходяться за посиланням.

1.3 Порогові значення для кожного з критеріїв

Табл. 1: Порогові значення для кожного з критеріїв

Номер критерію	Порогове значення для символів	Порогове значення для біграм
Критерій 2.0	3 (при L=10)	1 (при L=10)
	3 (при L=100, 1000, 10000)	10 (при L=100, 1000, 10000)
Критерій 2.1	3, 1 (при L=10)	3, 1 (при L=10)
	5, 3 (при L=100, 1000, 10000)	10, 4 (при L=100, 1000, 10000)
Критерій 2.2	1 (при L=10)	1 (при L=10)
	3 (при L=100, 1000, 10000)	10 (при L=100, 1000, 10000)
Критерій 2.3	2 (при L=10)	3 (при L=10)
	5 (при L=100, 1000, 10000)	10 (при L=100, 1000, 10000)
Критерій 4.0	0,01 (при L=10)	0,01 (при L=10)
	0,002 (при L=100, 1000, 10000)	0,002 (при L=100, 1000, 10000)
Критерій 5.0	1 (при L=10)	1 (при L=10)
	4 (при L=100, 1000, 10000)	4 (при L=100, 1000, 10000)
Структурний критерій	0,0085 (при L=10)	0,09 (при L=100, 1000, 10000)

1.4 Алгоритм стиснення для структурного критерію

У даній лабораторній роботі використовувався універсальний алгоритм стиснення даних LZW. Даний алгоритм при стисненні даних створює таблицю перетворення рядків: певним послідовностям символів ставляться у відповідність групи бітів фіксованої довжини. Таблиця ініціюється всіма символами. По мірі кодування алгоритм переглядає текст символ за символом і зберігає кожен двох-символьний рядок в таблицю у вигляді пари код/символ, де код посилається на відповідний перший символ.

1.5 Структурний критерій, що базується на основі результатів стиснення

Дивимось на модуль різниці степеня стиснення вхідного тексту та мови. Даний модуль має бути більше порогового значення.

1.6 Опис труднощів

Основними труднощями в даній лабораторній роботі були структурний критиерій (а саме, пізна реалізація його відсутності у моїй роботі) та обробка результатів, оскільки даний лабораторний практикум є досить громіздким.

2 Висновки

1. Ми засвоїли статистичні методи розрізнення відкритого тексту від випадкової послідовності. А саме, були розглянуті критерії: 2.0-2.3 (критерій частих l-грам та його варіації), 4.0 (критерій розрахунку індексу відповідності) та 5.0 (критерій порожніх ящиків) та структурний;

2. Було визначено похибку другого роду для спотворених текстів за допомогою шифру Віженера, Афінної перестановки, рівномірно розподіленої послідовності та за допомогою співвідношення, вказаного в методичці. Також визначили похибку першого роду для відкритого тексту;

3. Вибрали та описали алгоритм стиснення під назвою LZW. Також сформулювали критерій до алгоритму стиснення;

4. Проаналізувавши значення бачимо, що краще дає результат структурний критерій. Також при великих L критерії дають кращій результат.