Лабораторна робота №08а

«Закони статистичної лінгвістики на лінгвістичних рівнях букв (символів) і буквених (символьних) n-грам для окремих текстів»

Завдання:

Використовуючи програму +proj6stats&plots, дослідити закони статистичної лінгвістики (див. лабораторні роботи №2 і №4) на лінгвістичних рівнях букв (символів). Розглянути статистичні закони для буквених і символьних n-грам для окремих випадків n = 1–4 для деякого тексту. Побудувати спільну статистику для цих n-грам.

Теоретичні відомості:

1. конспект лекцій
2. стаття Rank-frequency dependences for the symbolic N-grams in natural texts2015.pdf
3. файл-приклад sample_Heaps_n=2(symbols+register-space)Don Quixote.opj.

Порядок виконання роботи та вказівки до оформлення звіту

1. Вимоги до оформлення звіту див. у лабораторній роботі №2.
2. У теоретичній частині зверніть увагу на особливості рангових залежностей, частотних розподілів і закону зростання словника для буквених і символьних n-грам. Рангова залежність для букв наближено описується логарифмічною функцією, закон Парето – експоненційною функцією, а закон Гіпса – експоненційною функцією складнішої форми. Водночас, зі зростанням n рангова залежність для n-грам поступово змінюється від логарифмічної до степеневої.

3. Оберіть один текст деякою мовою. Для одержання результатів, які буде простіше інтерпретувати, рекомендуємо вивчати буквені, а не символьні n-грами, а пробіл виключати зі статистичного аналізу.

4. Зобразіть графічно рангові залежності F(r) для буквених n-грам (n = 1, 2, 3, 4) у напівлогарифмічному (беремо логарифм тільки по осі абсцис і перевіряємо гіпотезу про логарифмічну залежність F(r)) і в подвійному логарифмічному масштабі (беремо логарифм по осях абсцис і ординат – перевіряємо гіпотезу про степеневу залежність F(r)).
5. Виконайте лінійну апроксимацію залежностей F(r) для різних n, побудованих в обох згаданих вище масштабах, випишіть значення коефіцієнтів нахилу прямих та коефіцієнтів лінійної кореляції за Пірсоном R.
6. Порівняйте значення R, отримані лінійною апроксимацією в напівлогарифмічному та подвійному логарифмічному масштабах, для випадків різних n. Якість якої із двох альтернативних лінійних апроксимацій вища при n = 1 і при n = 4? Якість якої з них підвищується зі зростанням n? На цій підставі зробіть висновки, як змінюється характер залежності F(r) зі зростанням n.

7. Додаткове завдання:
   виконайте завдання за пунктами 4–6 для розподілу кумулятивної ймовірності частоти cmf(F) (закону Парето). Тут для кожного значення n слід порівняти якість апроксимації в масштабах log(cmf) = *f*(F) (перевірка гіпотези про експоненційну залежність cmf(F)) і log(cmf) = *f*(logF) (перевірка гіпотези про степеневу залежність cmf(F)). Як

змінюється співвідношення між параметрами R зі зростанням n? На цій підставі зробіть висновок про зміни характеру залежності cmf(F) зі зростанням n.

8. Додаткове завдання:
   Вивчіть закон зростання «словника» букв V (тобто реально використаного в тексті алфавіту) зі зростанням довжини тексту L. Для цього за методом проб і помилок оберіть невелику початкову ділянку L тексту (орієнтовно L = 0–500 або, скажімо, L = 0–5000) і на цій ділянці побудуйте залежність V(L), обравши оптимальні значення параметрів біжучого вікна на закладці «Гіпс» програми +proj6stats&plots.
   За найпростішою теорією, зростання словника букв V описується формулою
   $$V = V_0[1 - \exp(-L/L_0)],$$
   де V_0 – це максимальний розмір алфавіту даної мови ($V_0 = 26$ для англійської мови), L_0 – деяка характеристична довжина тексту. Для перевірки цієї гіпотези розрахуйте допоміжний параметр $y = 1 - V/V_0$ і побудуйте залежність y(L). Перейдіть до логарифмічного масштабу по осі ординат y. Перевірте, чи справді одержана залежність $\log y(L)$ є лінійною, виконайте лінійну апроксимацію та визначте коефіцієнт лінійної кореляції за Пірсоном R.

9. Використовуючи ту саму програму, побудуйте **спільну** статистику для всіх буквених n-грам із n = 1, 2, 3 і 4. Для цього використайте поле «Спільна статистика». Збережіть результати та побудуйте спільну рангову залежність F(r) для всіх n-грам із n = 1, 2, 3 і 4. Побудуйте відповідний графік залежності F(r) у напівлогарифмічному масштабі (беремо логарифм тільки по осі абсцис, тобто перевіряємо гіпотезу про логарифмічну функцію F(r)) та подвійному логарифмічному масштабі (беремо логарифм по осях абсцис і ординат, тобто перевіряємо гіпотезу про степеневу функцію F(r)).
10. Виконайте лінійну апроксимацію залежностей в обох масштабах і порівняйте якість цих апроксимацій, виходячи з величин відповідних коефіцієнтів кореляції Пірсона. Випишіть ці коефіцієнти. У якому масштабі проаналізована залежність має більшу близькість до прямої лінії? Яка з гіпотез про логарифмічну або степеневу залежність F(r) ліпше підтверджується експериментальними даними?

11. Додаткове завдання:
    проаналізуйте характер статистичних залежностей, розглянутих вище, для випадку символів і символьних n-грам.

12. Висновки повинні містити короткий аналіз отриманих Вами результатів.