Лабораторна робота №08б

«Закони статистичної лінгвістики для лексичних n-грам для окремих текстів»

Завдання:

Використовуючи програму +proj6stats&plots, дослідити закони статистичної лінгвістики (див. лабораторні роботи №2 і №4) на лінгвістичному рівневі слів для окремих випадків словесних (тобто лексичних) n-грам із n = 1–5 для довільного тексту з бази. Побудувати спільну статистику для цих n-грам.

Теоретичні відомості:

1. конспект лекцій
2. стаття Extension of Zipf's Law to Words and Phrases.pdf

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 лексичних n-грам. Рангова залежність F(r), лексичний спектр pmf(F), залежність кумулятивної ймовірності від частоти cmf(F) і залежність розміру словника від розміру тексту для слів наближено описуються степеневими функціями, тобто для цих залежностей виконуються відповідно перший і другий закони Ціпфа, закон Парето та закон Гіпса. Ці ж закони загалом справджуються для випадків n = 2, n = 3, ... Водночас, зі зростанням n рангова залежність для n-грам поступово стає пологішою, тобто показник α першого закону Ціпфа дещо зменшується. Нарешті, теорія передбачає, що спільна статистика для всіх n-грам із n = 1–5 ліпше (точніше, якісніше) описується степеневими законами, аніж окремі статистичні дані для n = 1, n = 2, n = 3, ...

3. Оберіть один текст деякою мовою. Користуючись програмою +proj6stats&plots із лабораторної роботи №2, виконайте розрахунки статистики для окремих випадків n = 1, n = 2, ..., n = 5. Збережіть ці статистичні дані та експортуйте їх у програму, в якій Ви будете будувати та аналізувати графіки.

4. Зобразіть графічно окремі рангові залежності F(r) для лексичних n-грам (n = 1, 2, 3, 4, 5) у 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 у цьому масштабі, випишіть значення коефіцієнтів нахилу прямих (тобто, коефіцієнтів Ціпфа α з протилежним знаком), а також коефіцієнтів лінійної кореляції за Пірсоном R.
6. Побудуйте всі залежності F(r) у подвійному логарифмічному масштабі, разом із лінійними апроксимаціями, на єдиному графіку.
7. Порівняйте значення R, отримані лінійною апроксимацією в подвійному логарифмічному масштабі, для випадків різних n. Чи залежить якість апроксимації від n? Якщо залежить, то як вона змінюється зі зростанням n? На цій підставі зробіть висновки про характер залежності F(r) та його поведінку зі зростанням n.
8. Порівняйте значення параметра α, отримані лінійною апроксимацією в подвійному логарифмічному масштабі, для випадків різних n. Чи залежить параметр α від n? Якщо залежить, то як він змінюється зі зростанням n?

9. Додаткові завдання:

а) виконайте завдання за пунктами 4–8 для розподілу кумулятивної ймовірності частоти cdf(F) (закону Парето). Тут для кожного значення n слід порівняти якість лінійної апроксимації в масштабі log(cdf) = *f*(logF) (перевірка гіпотези про степеневу залежність cdf(F)). Як змінюється співвідношення між параметрами R зі зростанням n? На цій підставі зробіть висновок про незмінність або зміни характеру залежності cdf(F) зі зростанням n.

Як змінюється значення нахилу прямої лінії (тобто параметра закону Парето k) зі зростанням n? На цій підставі зробіть висновок про незмінність або зміни параметра k зі зростанням n.

б) виконайте завдання за пунктами 4–8 для залежності розмірів словника V від довжини тексту L (закону Гіпса). Тут для кожного значення n слід порівняти якість лінійної апроксимації в масштабі log(V) = *f*(logL) (перевірка гіпотези про степеневу залежність V(L)). Як змінюється співвідношення між параметрами кореляції за Пірсоном R зі зростанням n? На цій підставі зробіть висновок про незмінність або зміни характеру залежності V(L) зі зростанням n.

Як змінюється значення нахилу прямих ліній V(L) (тобто, параметр θ закону Гіпса) зі зростанням n? На цій підставі зробіть висновок про незмінність або зміни параметра Гіпса θ зі зростанням n.

10. Використовуючи ту саму програму, побудуйте **спільну** статистику для всіх лексичних n-грам із n = 1, 2, 3, 4 і 5. Збережіть результати та побудуйте спільну рангову залежність F(r) для всіх n-грам із n = 1–5. Побудуйте відповідний графік залежності F(r) у подвійному логарифмічному масштабі (беремо логарифм по осях абсцис і ординат, тобто перевіряємо гіпотезу про степеневу функцію F(r)).
11. Виконайте лінійну апроксимацію згаданої залежності F(r). Порівняйте якість цієї апроксимації з якістю лінійних апроксимацій, виконаних окремо для різних n від 1 до 5, виходячи з величин відповідних коефіцієнтів кореляції Пірсона. Випишіть ці коефіцієнти R, а також коефіцієнти нахилу прямих ліній, які з точністю до знаку дорівнюють параметрові α першого закону Ціпфа. У якому випадку проаналізовані залежності logF(log r) є ближчими до прямої лінії, тобто залежності F(r) ближчі до степеневої залежності: у випадку окремих статистик для n = 1, 2, ... – чи у випадку спільної статистики для n = 1–5?
12. Порівняйте величини параметра α першого закону Ціпфа для усіх випадків окремих статистик для n = 1, 2, ... та для випадку спільної статистики для n = 1–5.

13. Висновки повинні містити короткий аналіз отриманих Вами результатів.