Лабораторна робота №15

«Визначення середньої довжини слів і речень»

Завдання

Використовуючи програми +LoSW_sliding window(single text) і +LoSW_(corpus of texts), дослідити закономірності для довжин слів і речень для двох випадків: єдиного тексту англійською, українською та російською мовами, а також для корпусу текстів однією з цих мов.

Теоретичні відомості

1. конспект лекцій
2. стаття Statistical distribution and fluctuations of sentence lengths in Ukrainian, Russian and English.pdf
3. довідковий матеріал Farsi&Japanese punctuation.doc

Порядок виконання роботи та вказівки до оформлення звіту

1. Звіт має містити титульну сторінку, текст завдання, коротку теоретичну частину, опис виконання роботи, одержаних результатів і висновки.

2. У теоретичній частині коротко описати теоретичні закономірності, які Ви досліджуєте в лабораторній роботі: орієнтовні середні довжини слів для різних мов, формули для розрахунку частоти (ймовірності) довжин слів і речень, вигляд теоретичних формул для розподілів імовірності довжин слів і речень, особливості флуктуацій довжин слів і речень.

3. Ознайомитися із доданими до лабораторної роботи програмами +LoSW_running window_single text і +LoSW_corpus of texts. Обрати по одному тексту англійською, українською та іншою мовою для вивчення, а також текстову базу для однієї з цих мов.

Як і у всіх решті лабораторних роботах, чітко зазначити в звіті, які саме тексти та бази обрано!

4. За допомогою програми +LoSW_running window_single text визначити
   а) середні довжини $\bar{l}$ і с.к.в. Δl довжин слів у буквах (або в символах);
   б) середні довжини $\bar{l}$ і с.к.в. Δl довжин речень у словах, в буквах та символах для обраних текстів трьома мовами залежно від ширини біжучого вікна.

   Початкову ширину вікна, крок приросту ширини вікна та крок ковзання біжучого вікна обирати, виходячи з довжини тексту L (кількостей букв, символів і слів), знайдених програмою з лабораторної роботи №2.

   Вважати, що повна кількість різних ширин вікна повинна бути від 10 до 100. На основі даних для середніх довжин і с.к.в. довжин розрахувати усереднену по всіх вікнах середню довжину $\bar{l}$ і усереднене с.к.в. Δl для випадків довжин слів і речень. Представити кінцеві дані для довжин слів і речень у вигляді $l = \bar{l} \pm \Delta l$.

5. За допомогою програми +LoSW_corpus of texts знайти дані для середніх довжин та с.к.в. довжин слів у буквах (або в символах) і речень у словах, у буквах (або символах) для обраного корпусу тексту. Ці дані містяться у вихідному xls-файлі програми.

6. Побудувати за цими даними графіки залежності імовірності довжини речення (слова) від цієї довжини $p(l)$ і залежності с.к.в. імовірності від самої імовірності $\Delta p\,(p)$, де l – це довжина слова (речення), p – середня частота (тобто ймовірність) довжини речення (слова) в корпусі, Δp – с.к.в. частоти (імовірності) даної довжини речення (слова) в корпусі.

7. Знайти степінь γ, який описує степеневу залежність
$$\Delta p\,(p) \sim p^{\gamma},$$
будуючи цю залежність в подвійному логарифмічному масштабі, апроксимуючи її прямою лінією та знайшовши нахил лінії в цьому масштабі.
У звіті вказати значення цих степенів, а також загальну статистичну інформацію про апроксимацію: коефіцієнт лінійної кореляції R за Пірсоном, стандартну похибку тощо.

8. Перевірити, чи справді розподіл ймовірності Ґуда описує залежність $p(l)$. Оскільки в загальному вигляді це зробити важко через складність формули, достатньо перевірити, чи залежність $p(l)$ справді має експоненційний «хвіст» (тобто ділянку при великих l, яка задовільно описується функцією експоненти) – адже саме така ситуація притаманна розподілові Ґуда.
Для цього слід побудувати залежність $p(l)$ у напівлогарифмічному масштабі (логарифм по осі ординат) і перевірити, з якою точністю (з яким коефіцієнтом кореляції) залежність $\log p(l)$ є лінійною в даному масштабі.

9. Висновки повинні містити короткий аналіз особливостей Ваших даних, графіків і апроксимацій, отриманих показників степеня γ та ін., порівняння отриманих даних для різних мов і різних лінгвістичних об’єктів (слів, речень; вимірювань у кількостях букв, символів, слів тощо).