Лабораторна робота №01

«Препроцесинг текстових документів»

Завдання
використовуючи видані програми, здійснити попереднє опрацювання окремих текстів і текстових баз різними мовами; порівняти функціонал та інтерфейс різних програм, їхні переваги та недоліки; дослідити залежність часу опрацювання великих текстових баз від їхнього розміру.

Теоретичні дані
конспект лекцій за тематикою попереднього опрацювання текстів.

Порядок виконання роботи та вказівки до оформлення звіту

1. Звіт має містити титульну сторінку, текст завдання, коротку теоретичну частину, опис виконання роботи, одержані результати та висновки.

2. У теоретичній частині коротко описати стандартні завдання попереднього опрацювання текстів, частину з яких Ви досліджуєте в цій лабораторній роботі, а також методи реалізації цих завдань: кодування текстів, основні лінгвістичні елементи текстів, препроцесинг текстів і текстових баз, злиття та поділ текстів на частини, формування формальних словників текстів і текстових баз, відмінності формального словника (словника словоформ) від справжнього словника лексем.

3. Ознайомитися із виданими програмами для препроцесингу, формування словників, інвертування, злиття та поділу текстів.

4. Обрати досліджуваний текст і одну або дві текстові бази. Для найнадійніших вимірювань часу роботи програми +Text cleaner&processor(main) текстова база повинна бути *великою*, скажімо з обсягом не меншим за 200 МБ. Якщо Ви оберете таку велику базу, то тоді для скорочення часу опрацювання бази текстів іншими програмами доречно обрати іншу, *малу* текстову базу, яка може містити, наприклад, 50–100 текстів.

   Зауваження 1. Як і у всіх решті лабораторних роботах, чітко зазначити в звіті, які саме тексти або бази текстів було обрано!

   Зауваження 2. Основні завдання цього лабораторного практикуму типово передбачають вивчення одного або кількох текстів різними мовами, одного малого текстового корпусу та одного великого корпусу. *Ідеальною ситуацією* є обрання Вами *базових* текстів і корпусів одразу перед початком виконання першої ж лабораторної роботи. Далі в наступних роботах Ви будете досліджувати ті самі тексти, поступово вивчаючи їхні різні властивості. Зазначимо, що для виконання окремих робіт (найперше роботи №19) бажано обрати тексти, зміст яких Ви хоч трохи знаєте, хоча загалом це не обов'язкова вимога. Зрештою, для роботи №19 Ви можете обрати окремий текст, проте його не повинні вивчати інші студенти!

   Приклад: Ви обираєте один англійський (або німецький, французький чи іспанський тощо) текст, один український (або китайський, японський, французький тощо) текст і один текст шведською (словацькою, білоруською тощо) мовою.

Один із цих текстів Ви можете також обрати для рандомізації в лабораторних роботах №№9–11, порівнюючи властивості вихідного природного і рандомізованого текстів.

Зазначимо, що лише в основному архіві +main text corpora2023-24.zip для Вас підготовлено тексти понад 70 мовами світу із сумарним обсягом майже 3 ГБ.

Тому неприпустимо обрання різними студентами текстів, обраних іншими студентами! Для уникнення такої ситуації просимо погоджувати обрання текстів зі старостами.

5. Використовуючи програми очищення текстів +Text cleaner&processor(main) і +Text cleaner(for English only), опрацювати один обраний Вами текст і малу текстову базу деякою мовою, які Ви потім будете використовувати в наступних лабораторних роботах. Серед завдань можуть бути переведення всіх літер до нижнього регістру, очистка від розділових знаків, невидимих символів (NPS – non-printed symbols), цифр тощо.

6. Злити докупи малу текстову базу за допомогою програм для злиття текстів +Text merger2022 і +Text merger&data statist2017.

7. Розбити обраний Вами текст на неоднакові за розмірами частини за певною ознакою (наприклад, після слова «розділ») або на однакові частини за допомогою програми розділення текстів на дві частини +TextSplitter(2halves) і програм розділення текстів на багато частин +TextSplitters1 або +TextSplitters2. Останні програми передбачають режими поділу за ознакою (буквою, символом, словом) або поділу на строго однакові (якщо це можливо) частини.

8. Інвертувати єдиний текст і малу текстову базу за допомогою програм +TextInverter2O.P.(py) і +TextInverter5V.B.(C#) на лінгвістичних рівнях символів, слів і речень.

9. Сформувати словник усіх текстів в обраній Вами малій текстовій базі або сукупний словник всієї текстової бази за допомогою програм +Text cleaner&processor(main) і +Dictionary generator. Якщо йде мова про сукупний словник, то попередньо слід злити всі тексти з бази докупи.

10. Обрати програму препроцесингу текстів +Text cleaner&processor(main). Поступово зменшуючи кількість текстів у великій текстовій базі, визначити часи t обробки текстової бази програмою залежно від розмірів текстової бази L у МБ. Побудувати графік $t(L)$ для роботи програми.

11. Встановити характер цієї залежності (степенева, лінійна, експоненційна, ...?), почергово будуючи графіки $t(L)$ в різних масштабах: $\log t(L)$, $t(\log L)$ і $\log t(\log L)$. Зробити відповідні висновки.

12. Порівняти технічні можливості, зручність інтерфейсу, функціональні можливості, відсутність багів, швидкодію, особливості використання центрального процесора комп'ютера та його оперативної пам'яті, а також інші можливі характеристики усіх програм препроцесингу текстів, на які Ви звернули увагу.

13. Висновки повинні містити короткий аналіз особливостей порівняння програм, Ваші побажання на адресу цих програм, а також висвітлення та пояснення Ваших даних і графіків.