Лабораторна робота №10

«Дослідження базових лінгвістичних властивостей рандомних символьних послідовностей»

Завдання:

1. Для досліджень використати рандомні та рандомізовані тексти, генеровані в лабораторній роботі №9.
2. Дослідити усереднений по всіх словах параметр $\overline{R}$ кластеризації слів у цих текстах і усереднений по всіх словах параметр $\overline{\gamma}$ кореляцій слів у них.
3. Дослідити ті ж параметри для такої ж кількості природних текстів, написаних різними мовами.
4. Зробити висновки стосовно схожості або відмінності параметрів $\overline{R}$ і $\overline{\gamma}$ для природних і рандомних або рандомізованих текстів.

Теоретичні відомості:

1. конспект лекцій;
2. наукові статті до лабораторних робіт №9, №19 і №21.

Порядок виконання роботи та вказівки до оформлення звіту

1. Вивчіть теоретичні відомості, пов'язані з рандомними текстами, кластеризацією та кореляціями слів у текстах. Дослідження параметра R кластеризації слів докладніше описано в лабораторній роботі №19, а дослідження параметра γ кореляцій – в лабораторній роботі №21.

2. Оберіть для дослідження рандомні та рандомізовані тексти, згенеровані Вами раніше в лабораторній роботі №9. А саме, в цій роботі вже було згенеровано *по одному тексту* з орієнтовними розмірами L ~ 100 000 слів або L ~ 500 000 букв таких типів (далі відтворено матеріал з лабораторної роботи №9):
   а. глобально рандомізований текст на лінгвістичному рівні слів (програма +shuffler(words only) з лабораторної роботи №11; кількість циклів рандомізації N ~ 50 000–500 000, вихідний природний текст – це один із обраних Вами в попередніх лабораторних роботах текстів);
   б. текст мавпи Міллера (програма +monkey_texts з лабораторної роботи №11; алфавіт M від 1 до 33; однакові ймовірності всіх букв; ймовірність пробілу 0.2; barrier = 1);
   в. текст Маркова (програма +Markov texts з лабораторної роботи №27; тип ланки – або «букви», або «слова»; довжина ланцюжка N – від 3 до 8);
   г. текст Саймона (програми +Simon_3in1 або +Simon_with memory із лабораторної роботи №11; імовірність нових слів з мішка 0.1–3.0%; решта параметрів за бажанням);
   д. текст Поля (програма +Polya texts з лабораторної роботи №11; початкові розміри урни зі словами від 10 000 до 100 000 слів; цілочисельні параметри ρ і ν обрати так: $\nu = 1, 2, ..., 5$, $\rho \leq \nu$).
   е. Додатково слід згенерувати тексти, які є рандомізованими версіями того ж обраного Вами одного природного тексту, але з використанням рандомізації інших типів. Це рандомізація на лінгвістичних рівнях букв (символів) і речень, а також локальна рандомізація на рівні слів. Для цього можна використати

програму +shuffler2023 (виконавчий файл у папках зі шляхом Build -> exe.win32-3.6) або програму +shuffler_universal із лабораторної роботи №11.

Разом у підпункті (е) Ви одержите 3 рандомізовані тексти: глобально рандомізований на рівні букв (символів), глобально рандомізований на рівні речень і локально рандомізований на рівні слів.

ж. Використовуючи програму +monkey_text_generator(letters) із лабораторної роботи №9, у цій роботі було додатково згенеровано *по одному* тексту мавпи Міллера із:

1) однаковими частотами букв
2) лінійною ранговою залежністю частоти букв
3) логарифмічною ранговою залежністю частоти букв.

У режимах 2) і 3) слід коректно обрати чисельні коефіцієнти на інтерфейсі (див. файл recipes_lin&log-monkey_corr.doc, а також кольорову підказку на інтерфейсі)!

Разом у підпункті (ж) Ви одержите 3 рандомні тексти мавпи Міллера.

Отже, сумарно Ви матимете 11 рандомних і рандомізованих текстів для подальших досліджень.

3. На додаток, оберіть для дослідження природний текст, який було взято Вами для рандомізації, і ще 10 додаткових природних текстів різними мовами. Для вибору текстів можете використати текстову базу в підпапці «diff languages_268texts» кореневої папки «smaller corpora_diff languages» в архіві «+main text corpora2023-24.zip».

   Сумарно Ви матимете 11 природних текстів.

4. Оберіть по одному природному та рандомному (або рандомізованому) тексту. У цих текстах знайдіть слова з найвищою (або однією з найвищих) абсолютною частотою F (типу слова «the» в англійських текстах), для яких статистика найбагатша, – або слова з найвищими показниками кластеризації R (такі можуть трапитися лише в природних текстах), але також із достатньо високою частотою F (скажімо, не меншою за F = 50–100). Для визначення частот F слів у тексті можете використати програму +proj6stats&plots із лабораторної роботи №2, а для визначення параметрів R слів у тексті – програму +NG.metrics_R із лабораторної роботи №19.

5. Користуючись програмою +FA010 із лабораторної роботи №21, визначте показники степеня γ, які описують кореляції для згаданих вище високочастотних слів у двох досліджуваних текстах.

   Для цього слід обрати режим роботи програми («слова» або «символи»). Далі слід виставити максимальну ширину w_{max} біжучого вікна, що приблизно дорівнює 5% від довжини тексту (тобто «максимальної позиції») L_{max}:

$$w_{max} = 0.05L_{max}.$$

   Потім оберіть мінімальне вікно w_{min}, крок переміщення (H) і крок збільшення (K) вікна так, аби приблизно дотримати рівності

$$w_{min} = H = K \text{ і } w_{min} = w_{max}/N, \text{ де } N = 100–400.$$

   Внизу на інтерфейсі запишіть обране Вами слово для досліджень в полі «Шаблон пошуку (Template)». Для решти параметрів залиште їхні дефолтні значення.

   Запустивши програму, у вікні «Alpha = ...» Ви одержите шуканий показник γ, а також інформацію про точність розрахунку цього параметра «Goodness = ...». Останній параметр не повинен бути меншим за 0.99.

   Випишіть інформацію про обрані Вами два тексти, обрані слова та їхні показники F, R, γ і Goodness.

Зробіть висновок стосовно наявності чи відсутності довгосяжних кореляцій для обраних Вами слів у природному і рандомному текстах. Нагадуємо, що $\gamma = 0.5$, якщо кореляції відсутні або короткосяжні, і $\gamma > 0.5$, якщо кореляції довгосяжні.

6. Зберіть усі досліджувані тексти (11 природних і 11 рандомних текстів) в єдину папку і використайте одну з програм +NG.metrics_R_avg1,2 із лабораторної роботи №19. У пункті меню «Файл» оберіть «Відкрити базу текстів» і вкажіть шлях до Ваших текстів. Оберіть лексичні n-грами, $n = 1$, граничні умови «РВС», фільтр частоти в межах 10–20, залежно від середнього розміру текстів (значення фільтру тим вище, що довшими є тексти). Активуйте розрахунок і збережіть результати розрахунків.

   Порівняйте усереднені по всіх словах параметри $\overline{R}$ і $\overline{\gamma}$, а також відповідні с.к.в. ΔR і $\Delta\gamma$ для природних і рандомних текстів. Які висновки випливають з цих даних? Чи існує кластеризація слів у природних і в рандомних текстах? Нагадуємо, що випадок відсутності кластеризації описується рівністю $\overline{R} \approx 1$, а випадок наявності кластеризації – нерівністю $\overline{R} > 1$.

7. Коротко ознайомтеся з інтерфейсом і роботою програми +FA&R_all words(symbols)_py із лабораторної роботи №21. Помістіть всі 22 досліджувані Вами тексти в папку «corpus» програми. Почергово опрацюйте кожен із цих текстів даною програмою.

   Послідовність дій така. Запустіть програму. У пункті меню «Choose file» оберіть текст, оберіть режим роботи програми («слова»), оберіть опції «Boundary condition = periodic», «Filter = ...» у межах Filter = 10–20, залежно від розміру тексту (параметр Filter тим більший, що довший текст). Решті параметрів залиште їхні дефолтні значення на інтерфейсі. Далі натисніть «Analyze», дочекайтеся завершення роботи програми та натисніть «Save data».

   Дані у формі таблиці Excel збережено в папці «saved_data» програми (файл типу «title.txt condition=periodic,fmin=15,n=1,w=(31,31,31,638),definition=static»). Заведіть іншу, підсумкову таблицю Excel, у якій запишіть назву тексту, його довжину в словах (див. поле «Length» програми) і 8 параметрів справа в проміжному файлі «title.txt condition=periodic,fmin=15,n=1,w=(31,31,31,638),definition=static». Це параметри, які названі в програмі як R_{avg}, ΔR, $R_{w\,avg}$, ΔR_w, b_{avg}, Δb, $b_{w\,avg}$ і Δb_w. Фактично це перейменовані параметри $\overline{R}$, ΔR, $\overline{R}_w$, ΔR_w, $\overline{\gamma}$, $\Delta\gamma$, $\overline{\gamma}_w$ і $\Delta\gamma_w$, де позначення $\overline{R}$ і $\overline{\gamma}$ уже пояснено вище, ΔR і $\Delta\gamma$ – це відповідні с.к.в., а параметри $\overline{R}_w$, ΔR_w, $\overline{\gamma}_w$ і $\Delta\gamma_w$ з нижнім індексом «w» позначають ті ж самі величини, що й $\overline{R}$, ΔR, $\overline{\gamma}$ і $\Delta\gamma$, але здобуті шляхом усереднення зі зважуванням (тому й індекс «w»).

   Далі повторіть згадані вище вимірювання для всіх 22 природних і рандомних текстів і занесіть їх у підсумкову таблицю Excel.

   Порівняйте найперше величини параметрів $\overline{R}$ і $\overline{\gamma}$ для природних і рандомних (або рандомізованих) текстів і зробіть належні висновки. Зокрема, чи помітні відмінності величин $\overline{R}$ і $\overline{\gamma}$ для природних і рандомних текстів? Чи можна використати ці відмінності для розрізнення цих альтернативних типів текстів?

8. У висновку підсумуйте отримані результати. Вкажіть, чи можна використовувати параметри $\overline{R}$, $\overline{\gamma}$ і споріднені параметри для того, аби відрізнити рандомні тексти, позбавлені змісту, від природних текстів, які наповнені змістом, – навіть не знаючи мови та синтаксису цих текстів?