Лабораторна робота №05

«Зростання словника для корпусу текстів. Практичні рецепти бінування»

Завдання

1. Дослідити закон Гіпса для великого корпусу текстів.
   Із баз текстів, що додаються, для роботи потрібно вибрати щонайменше 150 (а ще ліпше ~ 1000 і більше) текстів, які мають помітно різні розміри. У дослідженні Ви повинні розглянути три методи бінування: (1) з однаковими довжинами бінів; (2) з різними довжинами бінів, кожен з яких містить однакову кількість емпіричних точок (довжин текстів) L_i; (3) експоненційне (або логарифмічне) бінування.
2. Побудувати біновані графіки залежності розмірів словників V від розмірів текстів L. Використовуючи лінійну апроксимацію даних, знайти числове значення коефіцієнту θ, яке притаманне тій чи іншій обраній мові.
3. Порівняти дані для параметра Гіпса θ, отримані за допомогою різних способів бінування.

Теоретичні відомості

1. конспект лекцій
2. стаття Frequency fluctuations for the linguistic elements in textual base2017.pdf.

Порядок виконання роботи та вказівки до оформлення звіту

1. В теоретичній частині описати залежності середньої частоти слова від довжини тексту, середньоквадратичного відхилення частоти слова від довжини тексту, залежності середньоквадратичного відхилення частоти слова від частоти слова, а також зміст показників степенів для цих залежностей, співвідношення між ними і поняття ергодичності властивості.

2. Для виконання роботи оберіть одну з програм +V(L)binning. Зауважте, що програми ++V(L)binning2023_1,2 мають додаткові можливості вивчення ієрогліфічних мов завдяки наявності режиму «символів» (а не тільки режиму «слів», який годиться для вивчення словників алфавітних мов). Проте в цих програмах не працюють деякі додаткові функції на зразок часткового очищення аналізованих слів від «підозріло довгих» слів (за фактом – помилок оцифрування текстів).

3. Тексти можна взяти з корпусів English corpus1not cleaned, English base2cleaned&expanded або з наявних на гугл-диску корпусів іншими мовами. Бажано узяти мінімум 150 текстів, а ще ліпше 1000–3000 текстів з помітно різними розмірами. Усі тексти повинні бути тою ж мовою. Для дослідження можна взяти й повністю один із корпусів English corpus1not cleaned або English corpus2punctuation-cleaned&reduced. Нарешті, можна обрати також повні або неповні корпуси текстів китайською або японською мовами, з якими слід працювати в режимі «символи», а не «слова».

4. Побудуйте графік залежності розмірів словника V від розмірів тексту L за прикладом даних рис. 1г зі статті, вказаній у списку літератури.

5. Використайте три основні методи бінування (1)–(3). На основі цих типів бінування побудуйте залежності $V_{сер}(L)$, $dV(L)$ та $dV(V_{сер})$ для обраного корпусу текстів (див. рис. 2 зі статті в списку літератури; проте врахуйте, що ці графіки в статті стосуються не словника, а частоти деякого конкретного слова).

6. При бінуванні кількість точок краще брати в межах 10–50 (як у статті зі списку літератури) або самостійно визначити оптимальну кількість точок для вашого корпусу текстів. Під час експоненційного бінування уважно обирайте оптимальне значення степеня A.

7. До звіту включіть залежності в звичайному (лінійному) та логарифмічному масштабах, а також дані їхніх лінійних апроксимацій, як і в звіті до лабораторної роботи №2.

8. Додаткове (необов'язкове) завдання:
   засвоїти роботу програми +pdf-to-cdf2 із лабораторної роботи №17 для побудови залежності кумулятивної імовірності (cdf або cmf) на основі залежності густини імовірності pdf (або масової ймовірності pmf).

9. Висновок повинен містити короткий виклад інформації про особливості ваших результатів; порівняння даних, отриманих різними способами бінування; визначення найліпшого, на Вашу думку, способу бінування; висновок про середній параметр Гіпса θ для корпусу текстів.