Лабораторная работа 2
Расчёт энтропии с учетом корреляционных связей

Цель: изучение основных понятий информационной энтропии и практика в вычислении энтропии с учетом корреляционных связей.

Теоретическая часть

Вычисление энтропии текста с учётом корреляционных связей, то есть энтропии пар символов или даже более длинных последовательностей, позволяет получить более точную и глубокую оценку неопределённости и информационной структуры текста. Вот несколько ключевых аспектов, почему это важно:

1. **Учёт зависимостей между символами**:
   - В языках существуют статистические зависимости между символами. Например, в английском языке после буквы "q" почти всегда следует "u". Энтропия, вычисленная без учёта этих корреляций (только по отдельным символам), может дать завышенную оценку неопределённости. При учёте пар символов энтропия становится более точной, отражая реальное количество информации.
2. **Более точная оценка эффективности кодирования**:
   - Понимание структуры зависимостей позволяет разрабатывать более эффективные алгоритмы кодирования. Кодирование, основанное на моделях, учитывающих пары символов или даже более длинные последовательности, может существенно уменьшить количество необходимых бит для представления текста по сравнению с независимым кодированием каждого символа.
3. **Улучшение алгоритмов сжатия**:
   - Алгоритмы сжатия, которые используют статистическую информацию о последовательностях символов (например, алгоритмы, основанные на контекстной модели), могут достичь лучшей степени сжатия. Знание о том, какие символы или пары символов чаще встречаются вместе, позволяет сжимать данные более эффективно.
4. **Лучшее понимание языка**:
   - Анализ корреляций между символами может выявить языковые особенности и правила, что полезно не только в теории информации, но и в лингвистике, машинном обучении, анализе текстов и других областях.
5. **Применение в машинном обучении и искусственном интеллекте**:
   - В моделях машинного обучения, таких как нейронные сети, которые обрабатывают текст для задач перевода, суммаризации или генерации текста, учёт корреляций между символами помогает создать более точные и естественные модели языка.

Таким образом, вычисление энтропии, учитывая корреляционные связи, не только повышает точность оценки информационного содержания текста, но и способствует разработке более совершенных алгоритмов обработки и передачи информации.

Для вычисления энтропии пар символов используется формула, аналогичная основной формуле Шеннона, но применённая к парам символов. Это помогает учесть взаимные зависимости между символами в тексте. Вот шаги и формула для вычисления энтропии пар символов:

1. **Определение вероятностей пар символов**:
   - Для каждой возможной пары символов (x_i,x_j)) в тексте, определяется вероятность $p(x_i,x_j)$, что символ x_i следует за символом x_j.
2. **Формула энтропии для пар символов:**
   - Энтропия пар символов $H(X,Y)$ вычисляется по формуле:

$$H(X,Y) = -\sum_{x_i \in X}\sum_{x_j \in Y} p(x_i, x_j)\log_2 p(x_i, x_j)$$

где X и Y обозначают множества символов, и суммирование производится по всем парам символов (x_i,x_j), которые встречаются в тексте.

Задание

Выполняется на основе Лабораторной работы 1.

1. Написать программу на предпочитаемом языке программирования для вычесления энтропии с учетов корреляционных связей:

Таблица 1 – N(x_i,x_j)(количество пар символов)

	А	**Б**	**В**	**…**	**Я**	**Пробел**	
А	АА	АБ	АВ		АЯ	А	$\Sigma Аa_i$
Б	БА	ББ	БВ		БЯ	Б	$\Sigma Бa_i$
В	ВА	ВБ	ВВ		ВЯ	В	$\Sigma Вa_i$
…							
Я	ЯА	ЯБ	ЯВ		ЯЯ	Я	$\Sigma Яa_i$
Пробел	А	Б	В		Я		$\Sigma\ a_i$
	Σa_iА	Σa_iБ	Σa_iВ		Σa_iЯ	Σa_i	

2. * Реализуйте визуализацию полученных результатов в виде графиков, которые показывают распределение пар частот символов. Это поможет визуально оценить однородность или разнообразие текста.