Комбинаторные методы сжатия данных, работа 2 **Вариант 1 – для всех**

0. Выбрать большой литературный текст, художественный / научпоп / публицистику, на английском языке (или любом другом, все буквы которого есть в ASCII). Можно брать несколько текстов одного автора. Текст нужно брать в восьмибитной кодировке (ASCII), либо сконвертировать в нее. Напомню, что utf8 совпадает с ASCII только в диапазоне кодов 0–127. Объём текста 1–2 Мб (не меньше 0.8 Мб). Рекомендуется также почистить текст от экзотических символов, неразрывных пробелов, и т.п. Текст у каждого должен быть уникальный. Нужно прислать мне сам текст (`автор_название_студент.txt`) или ссылку на него.

1. Декапитализация — прием, иногда используемый для уменьшения энтропии текста на естественном языке. Все заглавные буквы заменяются на строчные, а информация о заменах кодируется максимально компактным способом. Нужно сформулировать набор правил капитализации и сохранить список исключений, лучше всего как битовый массив длины $|T|$, в котором единицы указывают на позиции исключений. (Т.е. либо в этой позиции заглавная буква не по правилу, либо по правилу должны быть заглавная, но стоит строчная.) Этот массив нужно максимально компактно закодировать, используя изученные на лекциях методы кодирования. Выбор метода необходимо обосновать, указать длину получившегося кода и дать ссылку на использовавшиеся скрипты.

Стандартные правила капитализации: первая буква текста; первая буква после точки с пробелом и/или переводом строки; то же самое для !?; буква, следующая сразу за двумя заглавными; I в окружении не-буквенных символов. Можно добавить свои правила, например, для распространенных в тексте имен (но список имен добавляется к длине закодированного списка исключений).

2. Кодирование марковского источника. Для декапитализированного текста построить все контекстные модели 3-го порядка (для каждого контекста собрать всю информацию, как в лекции 7). Посчитать необходимый объём памяти для хранения всех моделей в несжатом виде. Выбрать и обосновать способ кодирования для передачи модели декодеру, указать длину получившегося кода и дать ссылку на использовавшиеся скрипты.