Отчет по третьем домашнему заданию.
Дмитриев Никита, БД.

Алгоритм решения заключается в следующем. Я обхожу все файлы в корпусе и соединяю их текста в единую строку. Затем я прохожусь по этой строке и избавляюсь от всех знаков пунктуации, заменяя их пробелами, при этом символы ‘!’, ‘?’, ‘;’ заменяю на точки. Параллельно я определяю какие точки могут быть завершением предложения. Для этого я проверяю, что бы последние два символа, стоящие перед точкой были маленькими буквами. Конечно это условие охватывает не все случаи, но большинство. Далее после этой операции я привожу строку к нижнему регистру и разделяю строку на слова, получая список слов. После этого я приступаю к анализу слов. Я создаю три словаря (**D1**, **D2**, **D3**). В **D1** будут лежать слова, которые стоят после точки (то есть начала предложений) с их количеством. В **D2** ключом будет являться одно слово, в **D3** - два слова. Значениями этих ключей так же будет словарь, в котором лежат все слова с их количеством, которые встречаются после ключа. На следующем этапе из этих трех словарей я удаляю слова, которые в тексте встречаются один раз.

Затем я приступаю к созданию распределений слов. Рассмотрим **D1** (различные начала предложения). Из него я делаю список, из кортежей. В каждом кортеже лежит слово и число **p** (отношение количества данного слова ко общему количеству слов в данном словаре). То есть по сути это число - вероятность того, что предложение начнется со данного слова. Получили следующий список: [(**s1**, **p1**), (**s2**, **p2**), … , (**sn**, **pn**)]. Далее я делаю из этого словаря следующий словарь : [(**s1**, **p1**), (**s2**, **p1** + **p2**), … (**sn**, **p1** + **p2** + … + **pn**)]. И теперь, когда я хочу сгенерировать рандомное слово, с которого может начинаться предложение, я генерирую рандомное число **p** от **0** до **1** и иду по этому списку, пока **p** > **p1** + **p2** + … + **pi**. Как только это условие прекращает выполняться я возвращаю слово, которому сопоставляется данное число. Такие списки я составлю для каждого слова из **D2**, **D3**. И в конце я записываю полученные распределения на диск.

При генерации текста, функции подается на вход количество предложений. Число этих предложений разделяется на группы от 7-15 предложений (абзацы). Далее в функции генерации абзаца вызывается определенное число раз функция генерации предложения. В свою очередь функция предложения генерирует первое слово из словаря D1, а дальше по нему генерирует слова из словарей D2 и D3. после этого полученные слова преобразуются в предложение. Первая буква предложения делается заглавной, так же буква ‘i’ трансформируется в ‘I’, а если в предложении встречаются, например ‘I’ и ‘m’, идущие подряд, то они склеиваются через апостроф. Таким образом генерируется заданное количество предложений. Полученный текст сохраняется в файл, а программа после этого спрашивает не хотим ли мы сгенерировать еще один текст.