Python: ДЗ 3

Задача 1. Генератор текстов

Напишите программу которая будет генерерировать английские тексты с помощью цепей Маркова. Цепь Маркова – это последовательность случайных событий, в которой будущее не зависит от прошлого. Наш генератор будет генерировать новое слово исходя из знания двух предыдущих. План возможной реализации представлен ниже:

1. **Добыть обучающий корпус**

   Можно воспользоваться корпусом приложенным к заданию на вики или создать свой, подойдет куча статей из английской википедии по одной тематике, архив какого-нибудь новостного сайта или куча близких по тематике книг. Чем больше текстов – тем лучше результат, не скупитесь. Соберите их, руками или скриптом и сложите на диск удобным для вас способом. Минимальный размер корпуса – 15 миллионов слов.

2. **Посчитать статистику по корпусу**

   По всем вашим текстам посчитайте статистику встречаемости слов, для каждого слова и каждой пары слов нужно посчитать частоту возникновения каждого возможного слова после них. Незабудьте очистить вход от знаков препинания, лишних пробелов итд. Статистику эту надо сохранить на диск, с этим помогут модули *pickle* или *json*.

3. **Сгенерируйте текст**

   Загрузив посчитанную на предидущем шаге статистику сгенерируйте текст заданой длинны. Каждое предложение должно начинаться со случайного слова с большой буквы, второе слово должно выбираться из распределения частот слов идущих после заданого первого, а каждое последующее - из распределения заданого предыдущими двумя. Точка должна обрабатываться как отдельное слово – при генерации которого происходит начало нового предложения. Поддержка точек и больших букв обязательна, остальные символы реализуйте если хотите.

   Постарайтесь добиться от вашего текста максимальной человечности. Можно разбить его на абзацы, можно придумать как генерировать сложную пунктуацию, или как умнее выбирать первое слово для предложения, поддержать русский язык (очень трудно). Любые геройства будут награждены бонусными баллами.

4. **Отчёт и проверка**

   Выполнив домашнее задание пришлите отчёт на ys.python@gmail.com, в котором опишите своё решение, к письму приложите образец текста сгенерированного вашим скриптом (длинной не менее 10 тысяч слов). Код загрузите на github и приложите ссылку на репозиторий к письму (как пользоваться github – смотри ссылку на вики). Тема письма – **python ДЗ-3 Отделение, Имя Фамилия,** .

   Максимальная оценка 15 баллов, оценивается код (стиль и дизайн) и результат. На сдачу задания - 1 попытка.