Практическое задание 2.3.1. Написать простую модель, использующую лингвистические признаки текста (любого объема на ваше усмотрение, но не менее 10 предложений), для выявления преобладающего рода (мужского или женского) во фрагменте различных произведений и статей, представленных ниже:

a) Гарри Поттер и философский камень
b) Приключения Шерлока Холмса
c) Путешествие к центру Земли
d) Елизавета II – царствующая королева Великобритании (перед обработкой нужно удалить надстрочный и подстрочный текст!)
e) Любое другое произведение или статья на ваш выбор.

Аналогично разобранному примеру написать модель, использующую лингвистические признаки текста. Для этого:

1. Определите по тексту наборы слов, которые будут использованы для распознавания предложений. Для этого из выбранного вами текста создайте 2 множества с именами MALE_WORDS и FEMALE_WORDS, содержащих ключевые слова, относящиеся к мужским и женским родам соответственно. Например: он, она, парень, девушка и т.д.

2. Создайте функцию `genderize`, которая подсчитывает количество слов в предложении, попадающих в списки MALE_WORDS и FEMALE_WORDS. Если предложение содержит только слова из MALE_WORDS, оно классифицируется как мужское. Предложение, содержащее только слова из FEMALE_WORDS, классифицируется как женское. Если предложение содержит мужские и женские слова, отнесите его к категории двуполых; а если в нем нет ни мужских, ни женских слов, определите его как имеющее неизвестный род. **Функция возвращает русские наименования категорий!**

3. Напишите функцию, которая будет подсчитывать частоту слов, признаков рода и предложений во всем тексте статьи.

4. Используя библиотеку NLTK, разбейте абзацы на предложения. Выделив отдельные предложения, разбейте их на лексемы, чтобы выявить отдельные слова и знаки пунктуации, и передайте размеченный текст функциям классификации для вывода процентов предложений и слов, относящихся к категории мужских, женских, двуполых и неизвестной принадлежности.