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**.**

In [1]:

```
MALE_WORDS = set(['парень',
'представитель', 'председатель',
'мужской', 'мужчина', 'он', 'ему', 'его',
'мальчик', 'бойфренд', 'бойфренды',
'мальчики', 'брат', 'братья', 'папа',
'папы', 'чувак', 'отец', 'отцы', 'жених',
'джентльмен', 'джентльмены', 'бог',
'дедушка', 'старик', 'внук', 'жених', 'он',
'сам', 'муж', 'мужья', 'король',
'мужчины', 'мистер', 'господин',
'племянник', 'племянники',
'священник', 'принц', 'сын', 'сыновья',
'дядя', 'дяди', 'официант', 'вдовец',
'вдовцы'])
#Множество мужских слов
```

In [2]:

```
FEMALE_WORDS = set(['героиня', 'представительница',
'председательница', 'женщина',
'актриса', 'женщины', 'она', 'ее', 'тетя',
'тети', 'невеста', 'дочь', 'дочери',
'женщина', 'невеста', 'девочка',
'подруга', 'подруги', 'девочки',
'богиня', 'внучка', 'бабка', 'бабушка',
'сама', 'дамы', 'леди', 'мама', 'мамы',
'мать', 'матери', 'миссис', 'мисс',
'племянница', 'племянницы', 'жрица',
'принцесса', 'королевы', 'она', 'сестра',
'сестры', 'официантка', 'вдова',
'вдовы', 'жена', 'жены', 'женщина'])
#Множество женских слов
```

1. 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!

In [3]:

```
def genderize(words): #Функция подсчета количества слов в преложении
  mwlen = len(MALE_WORDS.intersection(words)) #Запишем в переменную mwlen количество муж
ских слов, используя функцию пересечения множеств
  fwlen = len(FEMALE_WORDS.intersection(words)) #В переменную fwlen количество женских с
лов
  if mwlen > 0 and fwlen == 0: #Если количество мужских слов больше нуля и женских слов
нет совсем, то функция вернет параметр "male" - мужской
    return "мужские"
  elif mwlen == 0 and fwlen > 0: #Аналогично, если женских слов не нуль, а мужских слов
нет, то вернет параметр "female" - женский
    return 'женские'
  elif mwlen > 0 and fwlen > 0: #Если количество женских и мужских слов отлично от нуля,
то предложение будет считать двуполым
    return 'двуполые'
  else:
    return 'неизвестно' #В ином случае - неизвестно.
```

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

In [4]:

```
from collections import Counter #Из встроенной библиотеки Python подключим функцию Counte
r для подсчета частоты слов
def count_gender(sentences):
    sents = Counter() #Задаем пустую переменную sents для подсчета количества предложени
й определенного рода
    words = Counter() #Задаем пустую переменную words для подсчета количества слов в пре
дложении
    for sentence in sentences:
      gender = genderize(sentence) #Вызываем ранее созданную функцию
      sents[gender] += 1 #Считаем количество предложений определенного рода
      words[gender] += len(sentence) #Считаем количество слов в предложении
    return sents, words
```

1. 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.

In [5]:

```
import nltk #Подключаем библиотеку NLTK
nltk.download('punkt') #Скачиваем для нее нужное для работы расширение

def parse_gender(text): #На основе этой библиотеки, создадим функции для разделения предл
ожения на отдельные слова.
  sentences = [[word.lower() for word in nltk.word_tokenize(sentence)] #Для этого в двух
циклах разобьем наш текст на предложения, а
  for sentence in nltk.sent_tokenize(text)] #каждое предложения на слова.
  sents, words = count_gender(sentences) #Вызов ранее созданной функции
  total = sum(words.values()) #В переменную total запишем все слова из предложений опреде
ленного рода.
  for gender, count in words.items(): #И для каждой категории посчитаем частоту слов в пр
оцентах.
    pcent = (count / total) * 100
    nsents = sents[gender]
    print( "{:.3f}% {} ({} предложений)".format(pcent, gender, nsents))
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
```

In [6]:

```
text = '''Мистер и миссис Дурсль проживали в доме номер четыре по Тисовой улице и всегда
с гордостью заявляли, что они, слава богу, абсолютно нормальные люди. Уж от кого-кого, а
от них никак нельзя было ожидать, чтобы они попали в какую-нибудь странную или загадочную
ситуацию. Мистер и миссис Дурсль весьма неодобрительно относились к любым странностям, за
гадкам и прочей ерунде.
Мистер Дурсль возглавлял фирму под названием «Граннингс», которая специализировалась на п
роизводстве дрелей. Это был полный мужчина с очень пышными усами и очень короткой шеей. Ч
то же касается миссис Дурсль, она была тощей блондинкой с шеей почти вдвое длиннее, чем п
оложено при ее росте. Однако этот недостаток пришелся ей весьма кстати, поскольку большую
часть времени миссис Дурсль следила за соседями и подслушивала их разговоры. А с такой ше
ей, как у нее, было очень удобно заглядывать за чужие заборы. У мистера и миссис Дурсль б
ыл маленький сын по имени Дадли, и, по их мнению, он был самым чудесным ребенком на свете
.
Семья Дурслей имела все, чего только можно пожелать. Но был у них и один секрет. Причем б
ольше всего на свете они боялись, что кто-нибудь о нем узнает. Дурсли даже представить се
бе не могли, что с ними будет, если выплывет правда о Поттерах. Миссис Поттер приходилас
ь миссис Дурсль родной сестрой, но они не виделись вот уже несколько лет. Миссис Дурсль д
аже делала вид, что у нее вовсе нет никакой сестры, потому что сестра и ее никчемный муж
были полной противоположностью Дурслям.
Дурсли содрогались при одной мысли о том, что скажут соседи, если на Тисовую улицу пожалу
ют Поттеры. Дурсли знали, что у Поттеров тоже есть маленький сын, но они никогда его не в
идели. И они категорически не хотели, чтобы их Дадли общался с ребенком таких родителей.
Когда во вторник мистер и миссис Дурсль проснулись скучным и серым утром — а именно с эт
ого утра начинается наша история, — ничто, включая покрытое тучами небо, не предвещало, ч
то вскоре по всей стране начнут происходить странные и загадочные вещи. Мистер Дурсль чт
о-то напевал себе под нос, завязывая самый отвратительный из своих галстуков. А миссис Ду
рсль, с трудом усадив сопротивляющегося и орущего Дадли на высокий детский стульчик, со с
частливой улыбкой пересказывала мужу последние сплетни.
Никто из них не заметил, как за окном пролетела большая сова-неясыть.
В половине девятого мистер Дурсль взял свой портфель, клюнул миссис Дурсль в щеку и попы
тался на прощанье поцеловать Дадли, но промахнулся, потому что Дадли впал в ярость, что с
ним происходило довольно часто. Он раскачивался взад-вперед на стульчике, ловко выуживал
из тарелки кашу и заляпывал ею стены.
— Ух, ты моя крошка, — со смехом выдавил из себя мистер Дурсль, выходя из дома.
Он сел в машину и выехал со двора.
На углу улицы мистер Дурсль заметил, что происходит что-то странное, — на тротуаре стояла
кошка и внимательно изучала лежащую перед ней карту. В первую секунду мистер Дурсль даже
не понял, что именно он увидел, но затем, уже миновав кошку, затормозил и резко оглянулся
. На углу Тисовой улицы действительно стояла полосатая кошка, но никакой карты видно не б
ыло.
— И привидится же такое! — буркнул мистер Дурсль.
Наверное, во всем были виноваты мрачное утро и тусклый свет фонаря. На всякий случай мист
ер Дурсль закрыл глаза, потом открыл их и уставился на кошку. А кошка уставилась на него
.
'''
```

In [7]:

```
parse_gender(text)
```

```
28.871% двуполые (6 предложений)
28.548% неизвестно (13 предложений)
28.871% мужские (11 предложений)
13.710% женские (4 предложений)
```