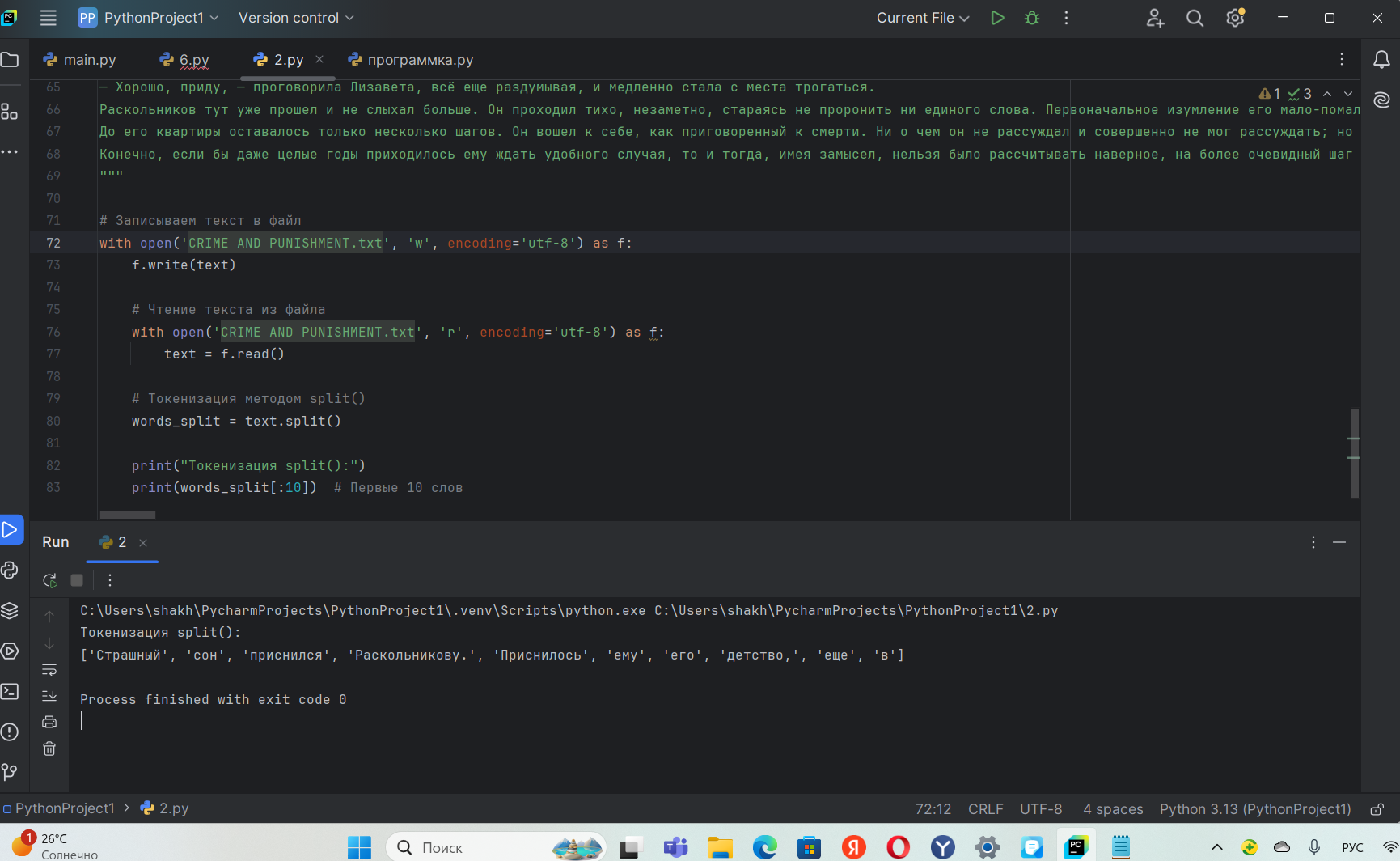
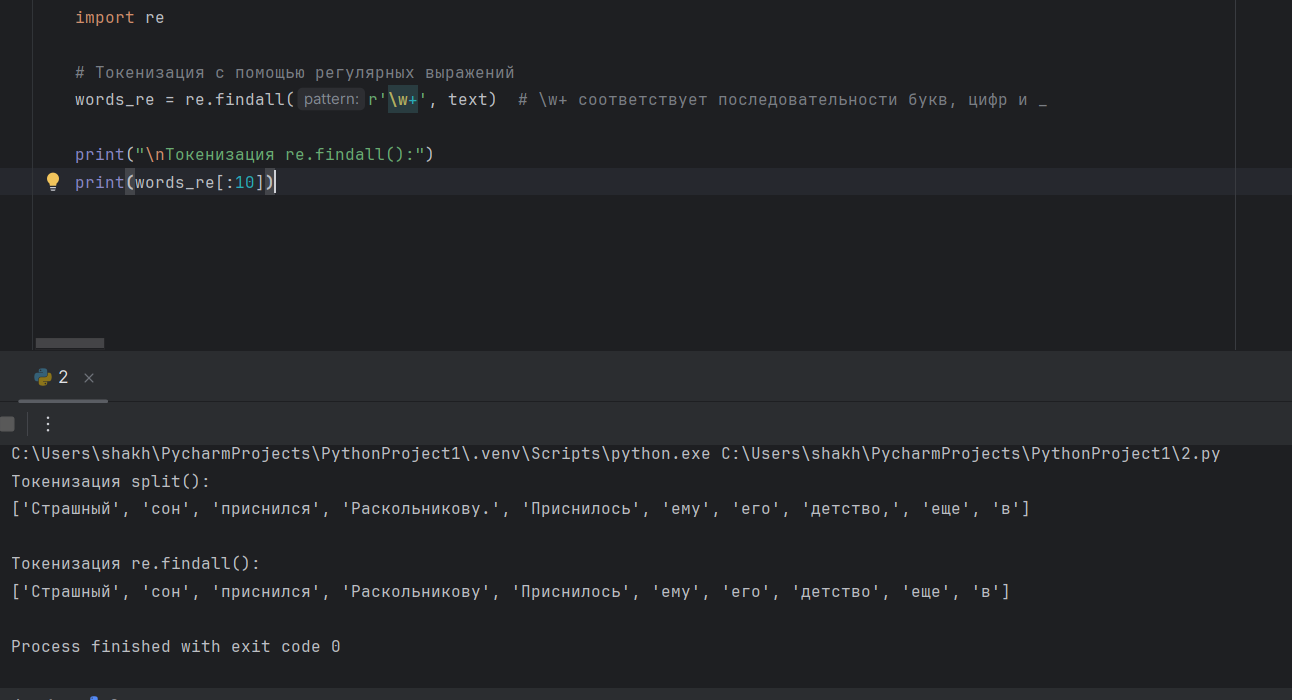
Лабораторная работа номер 14

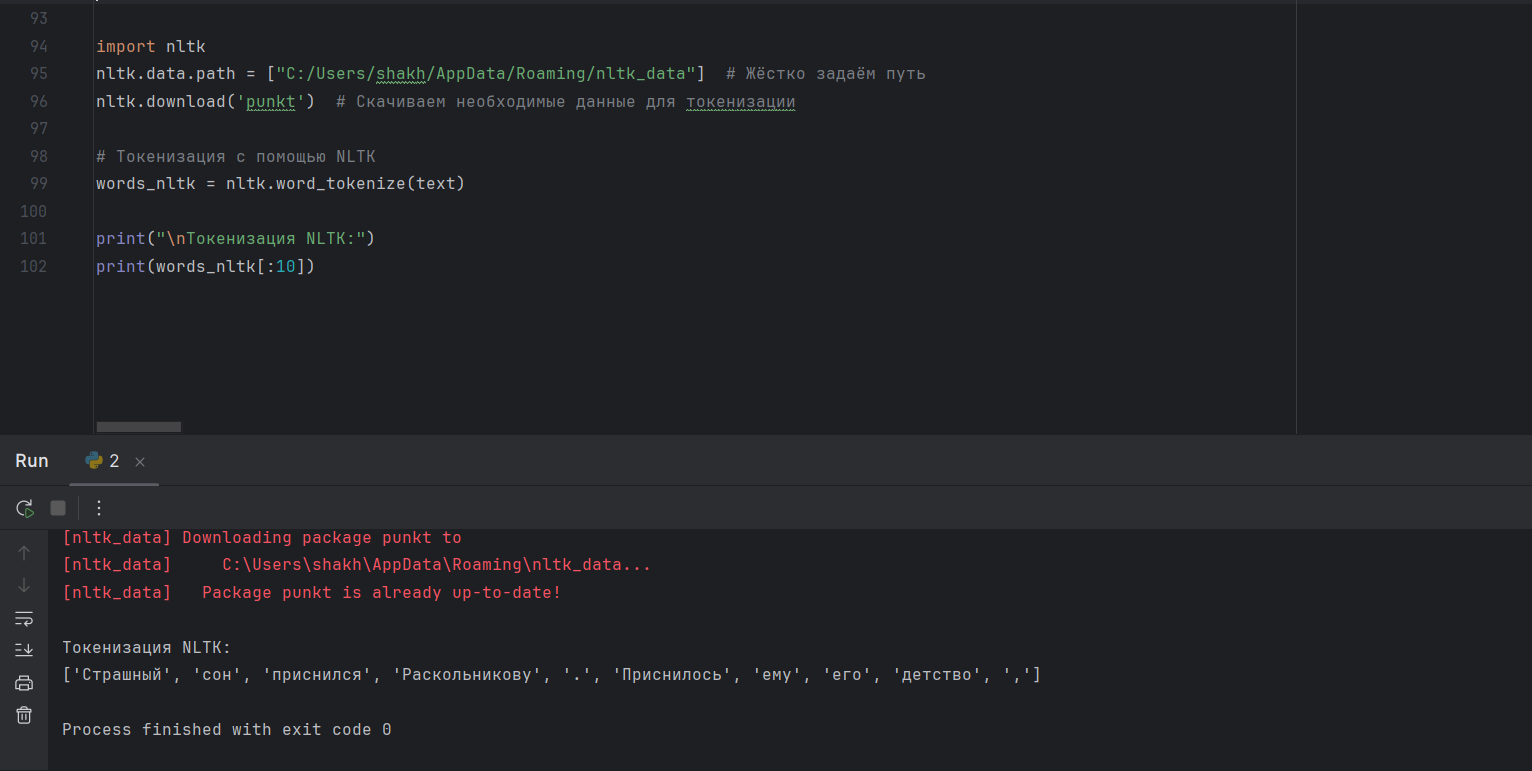
**Уровень 1:**

1. Проведите предобработку и морфоанализ текста на Питоне:  
- подгрузить текст из файла (тексты не должны повторяться)

- провести токенизацию методом .split

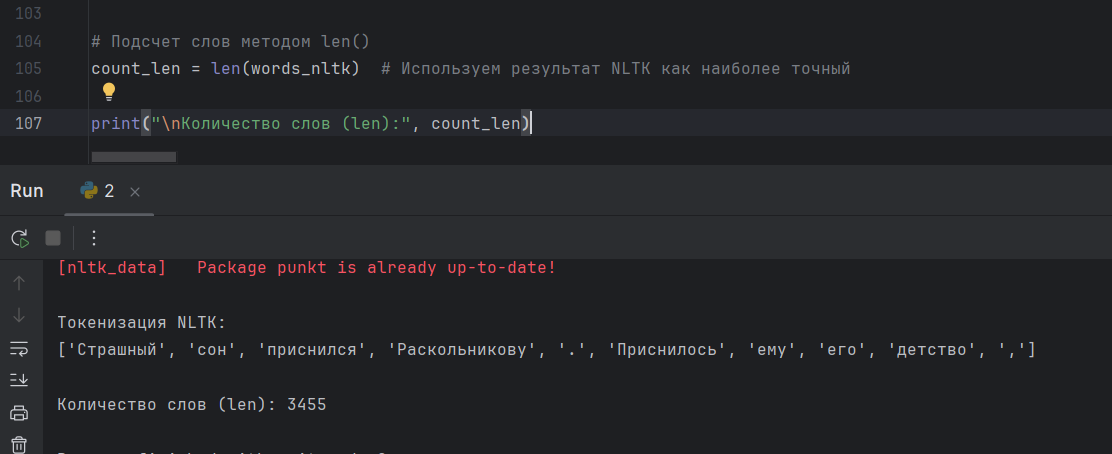
  
- провести токенизацию при помощи регулярных выражений

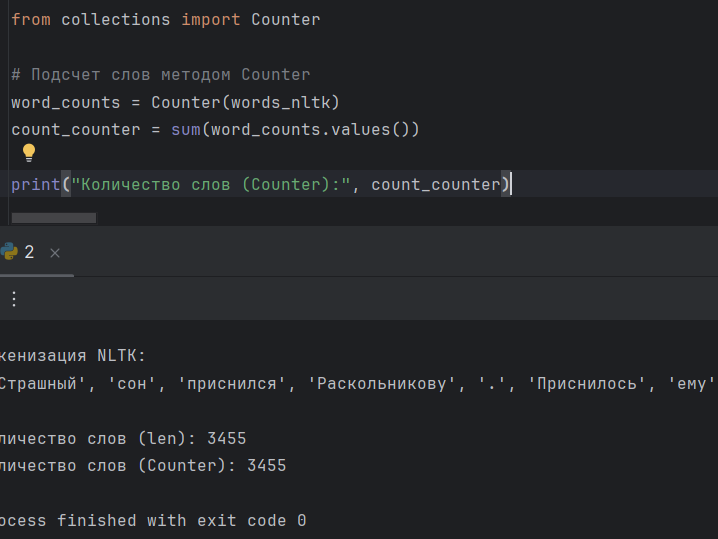
  
- провести токенизацию при помощи библиотеки NLTK

  
? в чем отличия разных способов токенизации?

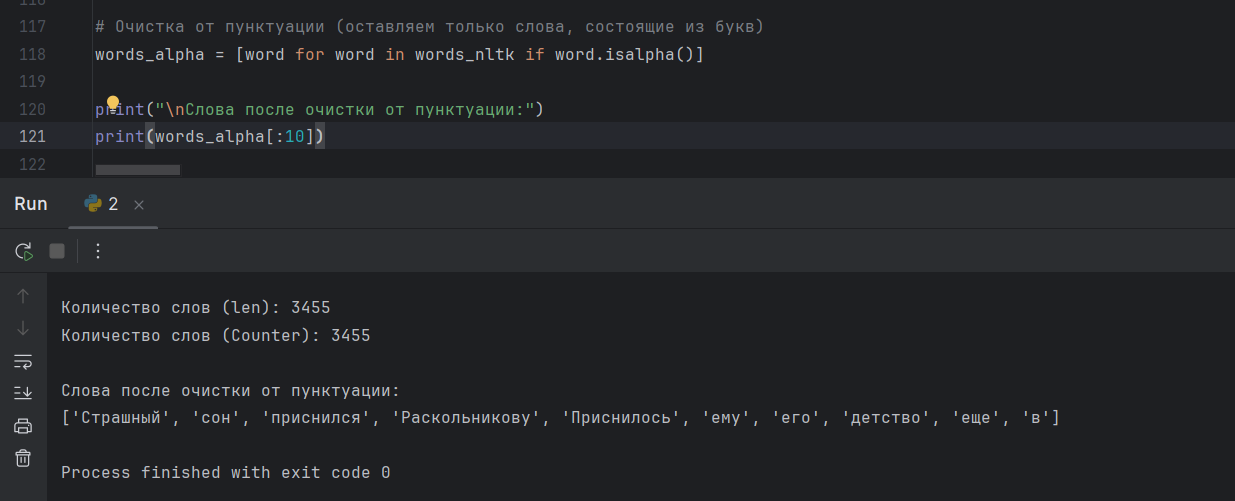
* .split(): просто разбивает текст по пробелам, не учитывает пунктуацию (она остается частью слов)
* re.findall(): более гибкий, можно настроить шаблон для выделения слов, но требует знания регулярных выражений
* NLTK: наиболее продвинутый метод, правильно обрабатывает пунктуацию, сокращения и другие особенности языка

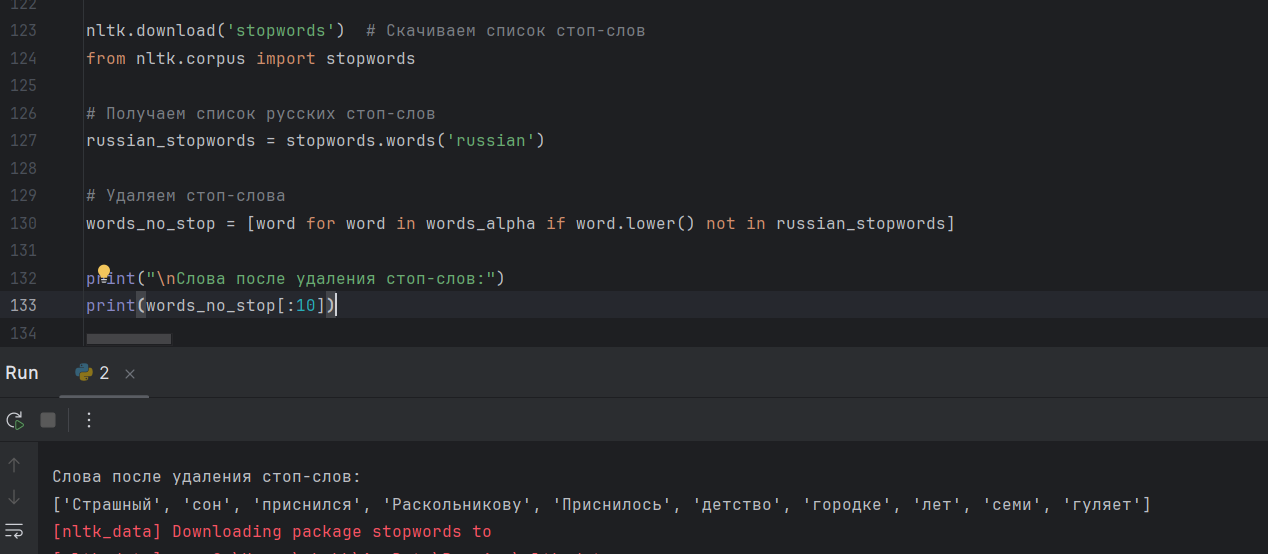
- посчитать количество слов в тексте при помощи len

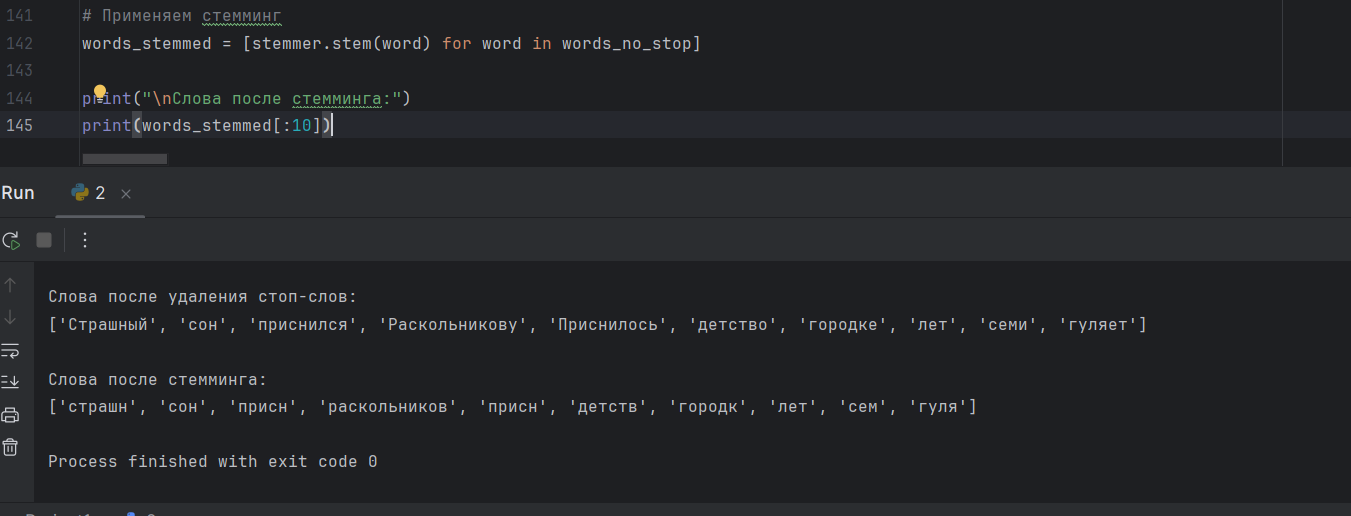
  
- посчитать количество слов в тексте при помощи Counter

  
? есть ли отличия в результате?

нет  
- очистить от пунктуации при помощи isalpha

  
- очистить текст от стоп-слов при помощи библиотеки NLTK

  
- провести стемминг при помощи библиотеки NLTK

  
- провести лемматизацию при помощи pymorphy2

? в чем отличия результатов после стемминга и лемматизации?  
- построить график наиболее частотных (лемматизированных) слов текста при помощи matplotlib  
- построить dispersion plot при помощи одноименного метода из библиотеки NLTK  
- определить грамматические характеристики слов при помощи pymorphy2  
- построить график, отражающий распределение частотности частей речи в тексте, при помощи matplotlib  
- протестировать методы similar, common\_contexts и collocations из библиотеки NLTK  
? для чего эти методы можно использовать?