Downloading texts

```
~/Documents/work/distributed_calculations_hw1 13:49:13
$ curl -o the_brothers_karamazov.txt https://www.gutenberg.org/cache/epub/28054/pg28054.txt
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 1994k  100 1994k    0     0  1234k      0  0:00:01  0:00:01 --:--:-- 1243k
```

```
~/Documents/work/distributed_calculations_hw1/inputs 13:50:43
$ cat * > combined.txt
```

```
~/Documents/work/distributed_calculations_hw1/inputs 13:50:20
$ find . -maxdepth 1 -type f -print0 | xargs -0 du -ch | grep total$

 22M    total
```

Writing program

```
def main():
    base_path = get_base_path()
    input_text_path = base_path + 'inputs/combined.txt'
    output_path = base_path + 'results'

    if os.path.exists(output_path):
        for filename in os.listdir(output_path):
            file_path = os.path.join(output_path, filename)
            if os.path.isfile(file_path):
                os.remove(file_path)
        os.rmdir(output_path)

    sc = SparkContext('local', 'Simple WordCount on one machine')

    words = sc.textFile(input_text_path).flatMap(
        lambda text:
        text.lower().translate(str.maketrans('', '', string.punctuation)).split()
    )

    result = words.map(lambda word: (word, 1))\
        .reduceByKey(lambda a, b: a + b)\
        .sortBy(lambda a: a[1], ascending=False)
    result.saveAsTextFile(output_path)


def show_results():
    path = get_base_path() + 'results/part-00000'
    with open(path, 'r', encoding='utf-8') as f:
        for i, line in enumerate(f, start=1):
            if i > 10:
                break
            print(line.rstrip())


if __name__ == "__main__":
    main()
    show_results()
```

Testing

```
/Users/maxim-movshin/.local/share/virtualenvs/pythonProject2-8u_Nj7EF/bin/python /Users/maxim-movshin/Documents/work/distributed_calculations_hw1/main.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/06/19 13:52:58 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
/Users/maxim-movshin/.local/share/virtualenvs/pythonProject2-8u_Nj7EF/lib/python3.11/site-packages/pyspark/python/lib/pyspark.zip/pyspark/shuffle.py:65: UserWarning: Please install psutil to have better
('the', 93871)
('and', 63995)
('to', 52720)
('of', 45565)
('a', 37139)
('he', 35528)
('in', 29444)
('that', 26552)
('was', 24329)
('his', 23933)

Process finished with exit code 0
```