Homework 1

1. Download and install Spark 2.4.7. To make Spark work on Windows 10 you will need additionally install winutils. You can find steps here: https://medium.com/@ligz/installing-standalone-spark-on-windows-made-easy-with-powershell-7f7309799bc7.

2. Implement word count job using RDD. Print results in console and save to txt file. Create separate Maven project. Use attached text file.

3. Make it run from IDEA/Eclipse/your IDE. Find a way to run it without specifying master in code. This way you will be able to run jobs locally without changing codebase.

4. Start Spark in standalone mode on your local machine, steps: https://medium.com/@aamargajbhiye/apache-spark-setup-a-multi-node-standalone-cluster-on-windows-63d413296971.

5. Build jar file with Maven and use spark-submit command to submit word count job to local standalone cluster. Checkout Spark UI and investigate "Completed applications".  
Add to merge request spark-submit command that you used to submit a job.

Do not hesitate to ask any questions. Spark documentation is great and it’s your best friend during this task.