SparkText

A Framework for Big Data Text Mining

User Manual

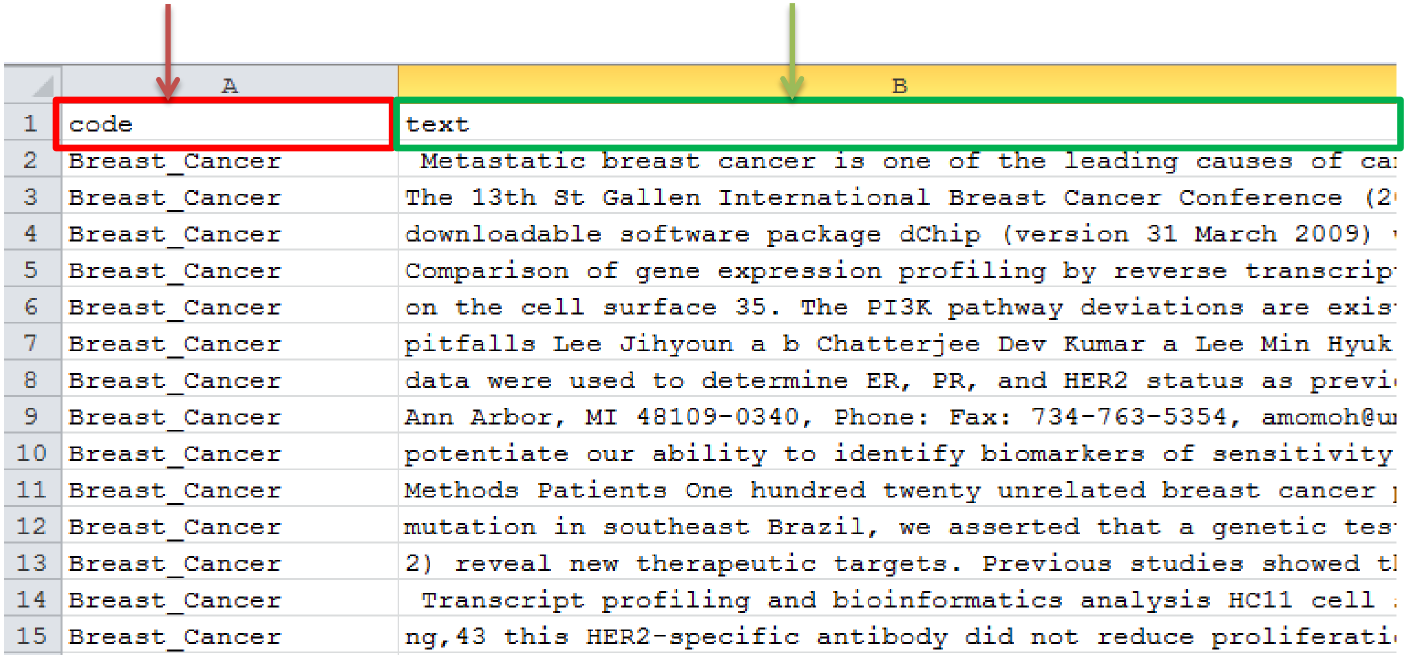
Step 1: Preparing Text Data

The first step is to prepare text data, considering pre-defined categories (classes) for each row of the data. As we have faced with a variety of text data, such as PDF, XML, Text, and any other formats, we assume that you can store your desirable text data into a CSV file.

In doing so, please save the CSV file as the following format:

Pre-defined Classes (Categories)

Text Data (e.g., Full text of scientific Articles)



	A	B
1	code	text
2	Breast_Cancer	Metastatic breast cancer is one of the leading causes of ca
3	Breast_Cancer	The 13th St Gallen International Breast Cancer Conference (2
4	Breast_Cancer	downloadable software package dChip (version 31 March 2009)
5	Breast_Cancer	Comparison of gene expression profiling by reverse transcrip
6	Breast_Cancer	on the cell surface 35. The PI3K pathway deviations are exis
7	Breast_Cancer	pitfalls Lee Jihyoun a b Chatterjee Dev Kumar a Lee Min Hyuk
8	Breast_Cancer	data were used to determine ER, PR, and HER2 status as previ
9	Breast_Cancer	Ann Arbor, MI 48109-0340, Phone: Fax: 734-763-5354, amomoh@u
10	Breast_Cancer	potentiate our ability to identify biomarkers of sensitivity
11	Breast_Cancer	Methods Patients One hundred twenty unrelated breast cancer
12	Breast_Cancer	mutation in southeast Brazil, we asserted that a genetic tes
13	Breast_Cancer	2) reveal new therapeutic targets. Previous studies showed t
14	Breast_Cancer	Transcript profiling and bioinformatics analysis HC11 cell
15	Breast_Cancer	ng,43 this HER2-specific antibody did not reduce proliferati

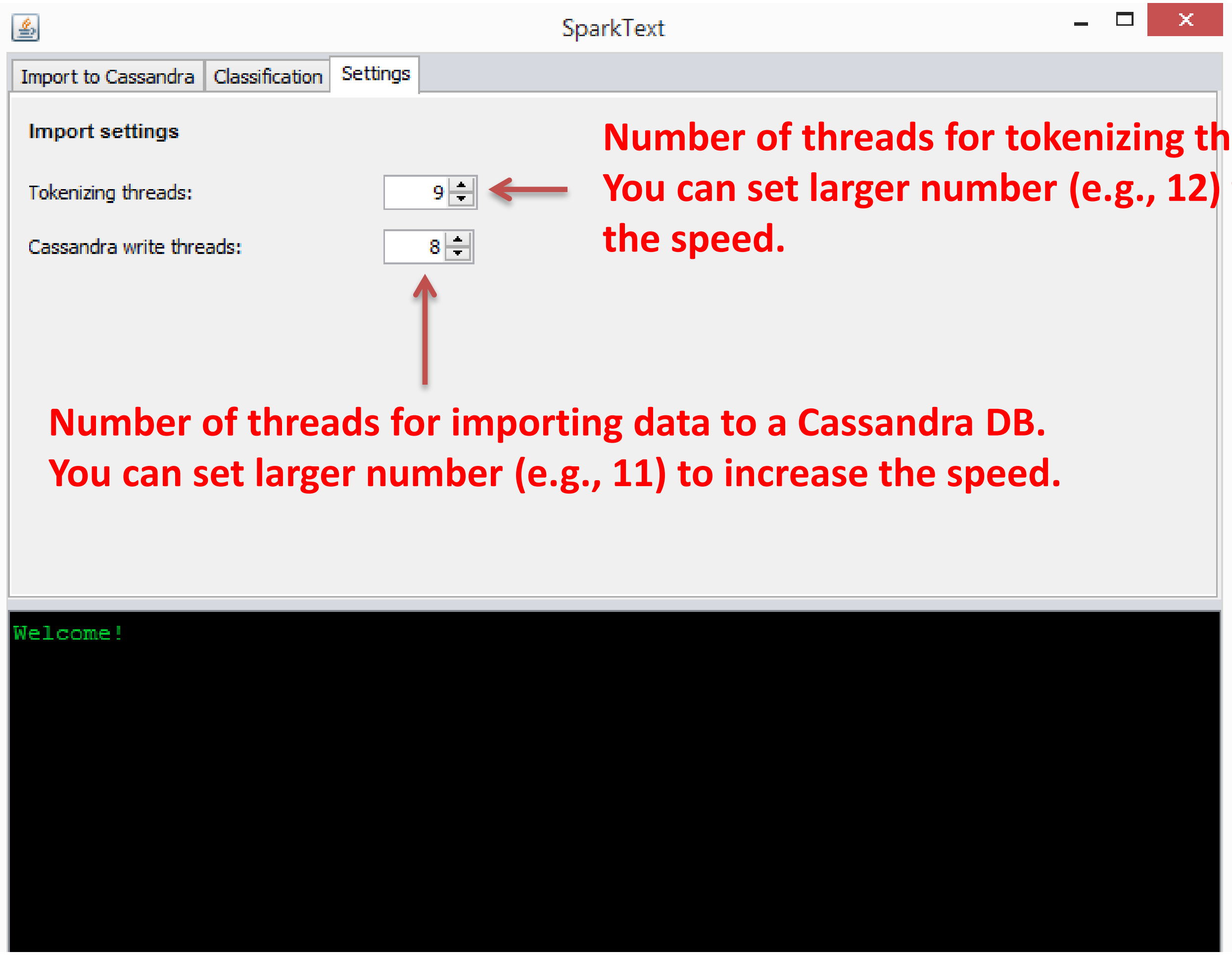
Step 2: Runing SparkText

To run the executable JAR file of the SparkText, you need to have JRE 8 and Cassandra DB on your machine. To run the executable JAR file, you need to type the following command in the command windows:

```
d:\sparktext>java –cp   sparktext.jar   mcrf.chg.textmining.gui.MainWindow
```

The local directory (path) of the SparkText package

Step 3: Setting



Import settings

Tokenizing threads: 9

Cassandra write threads: 8

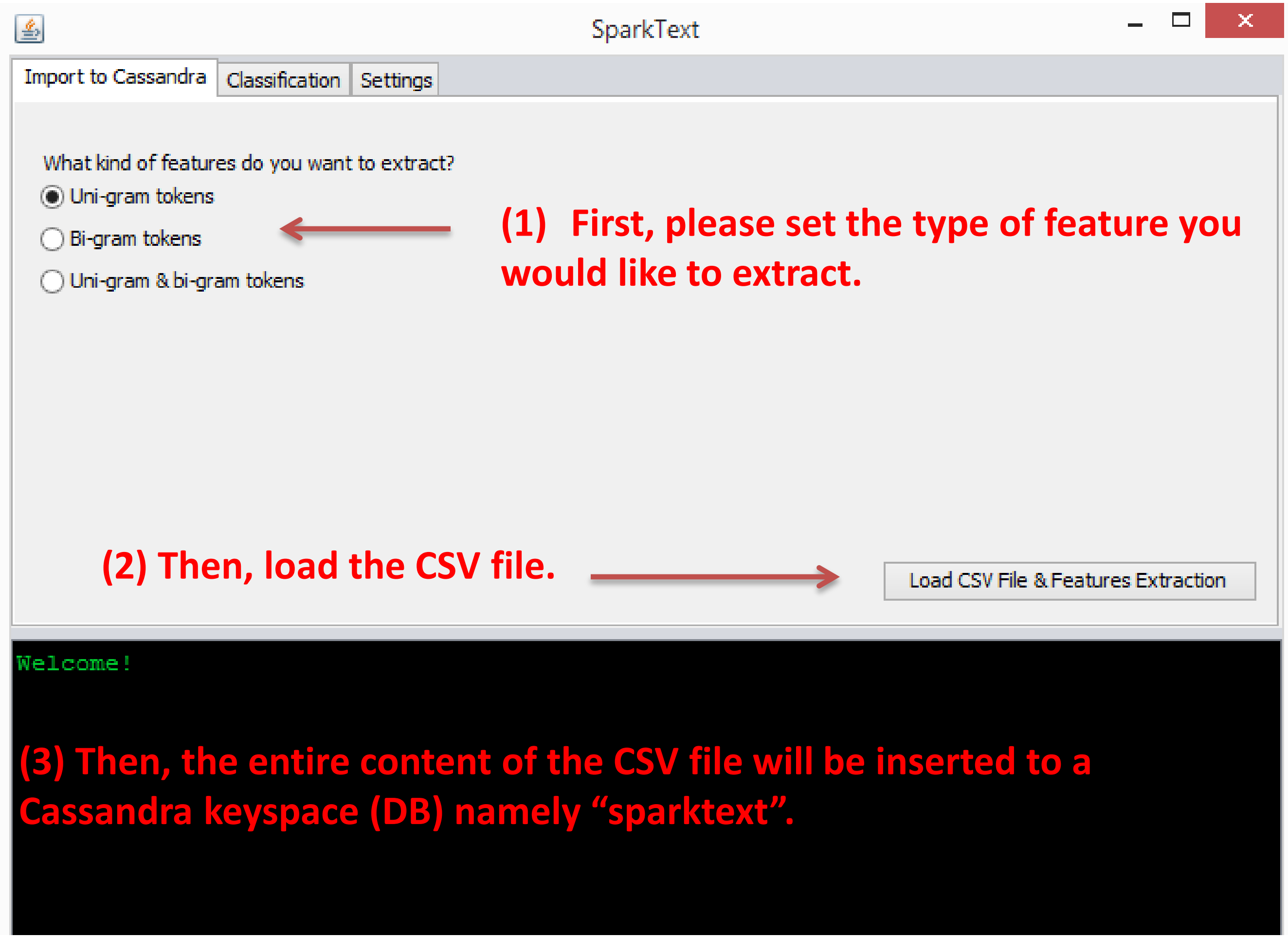
```
Welcome!
```

Number of threads for tokenizing the text data. You can set larger number (e.g., 12) to increase the speed.

Number of threads for importing data to a Cassandra DB. You can set larger number (e.g., 11) to increase the speed.

Step 4: Import to Cassandra

To make well-organized and structured data, an attempt is needed to import the CSV file and the features to a Cassandra DB.



What kind of features do you want to extract?

● Uni-gram tokens

○ Bi-gram tokens

○ Uni-gram & bi-gram tokens

(1) First, please set the type of feature you would like to extract.

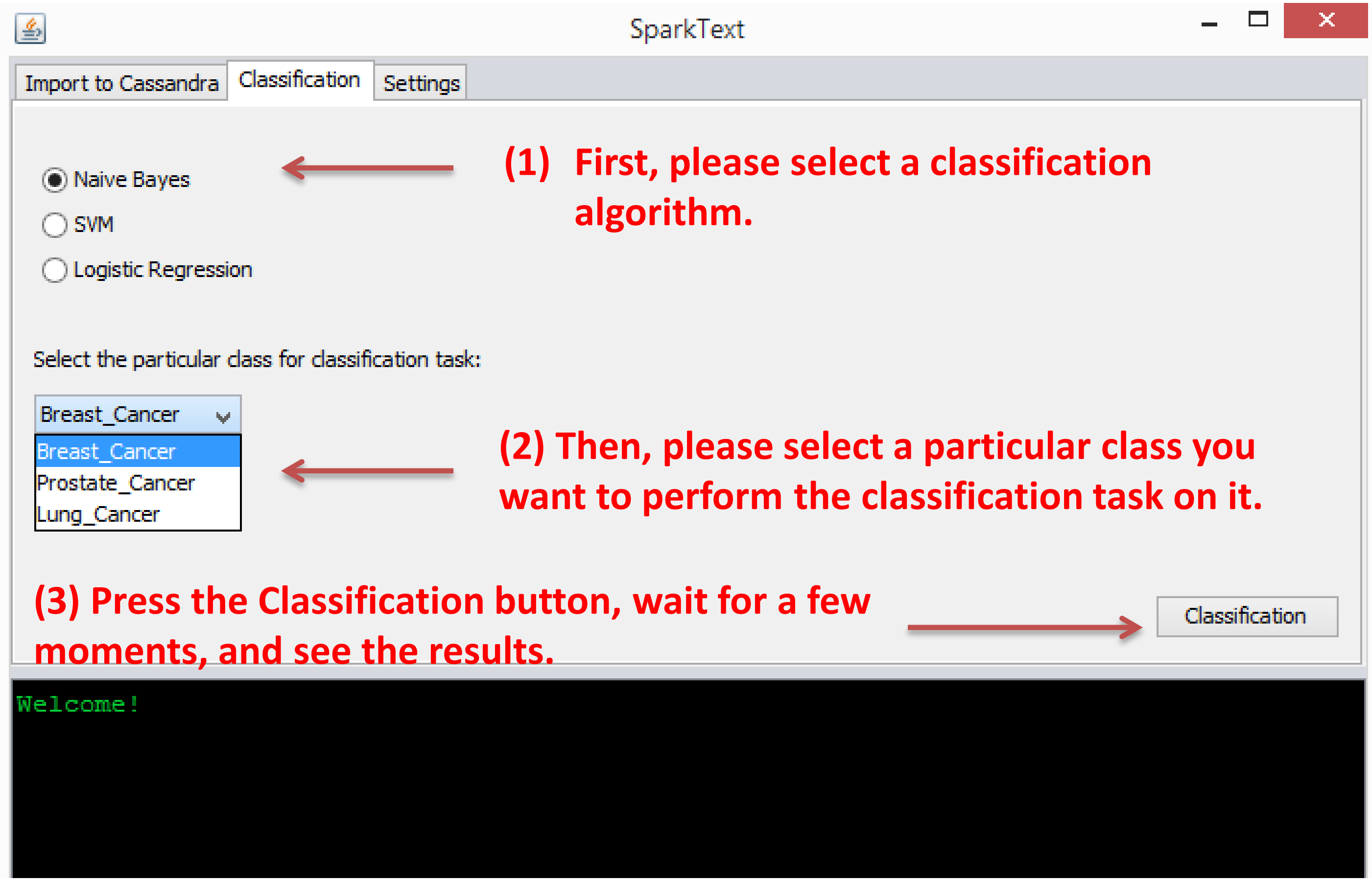
(2) Then, load the CSV file.

```
Welcome!
```

(3) Then, the entire content of the CSV file will be inserted to a Cassandra keyspace (DB) namely “sparktext”.

Step 5: Classification

Once the process of importing data into a Cassandra DB is done, you will get a prompt. After that, you can go to the classification tab, and select a classification algorithm.



● Naive Bayes

○ SVM

○ Logistic Regression

Select the particular class for classification task:

Breast_Cancer

- Breast_Cancer
- Prostate_Cancer
- Lung_Cancer

```
Welcome!
```

(1) First, please select a classification algorithm.

(2) Then, please select a particular class you want to perform the classification task on it.

(3) Press the Classification button, wait for a few moments, and see the results.

Thank you very much!