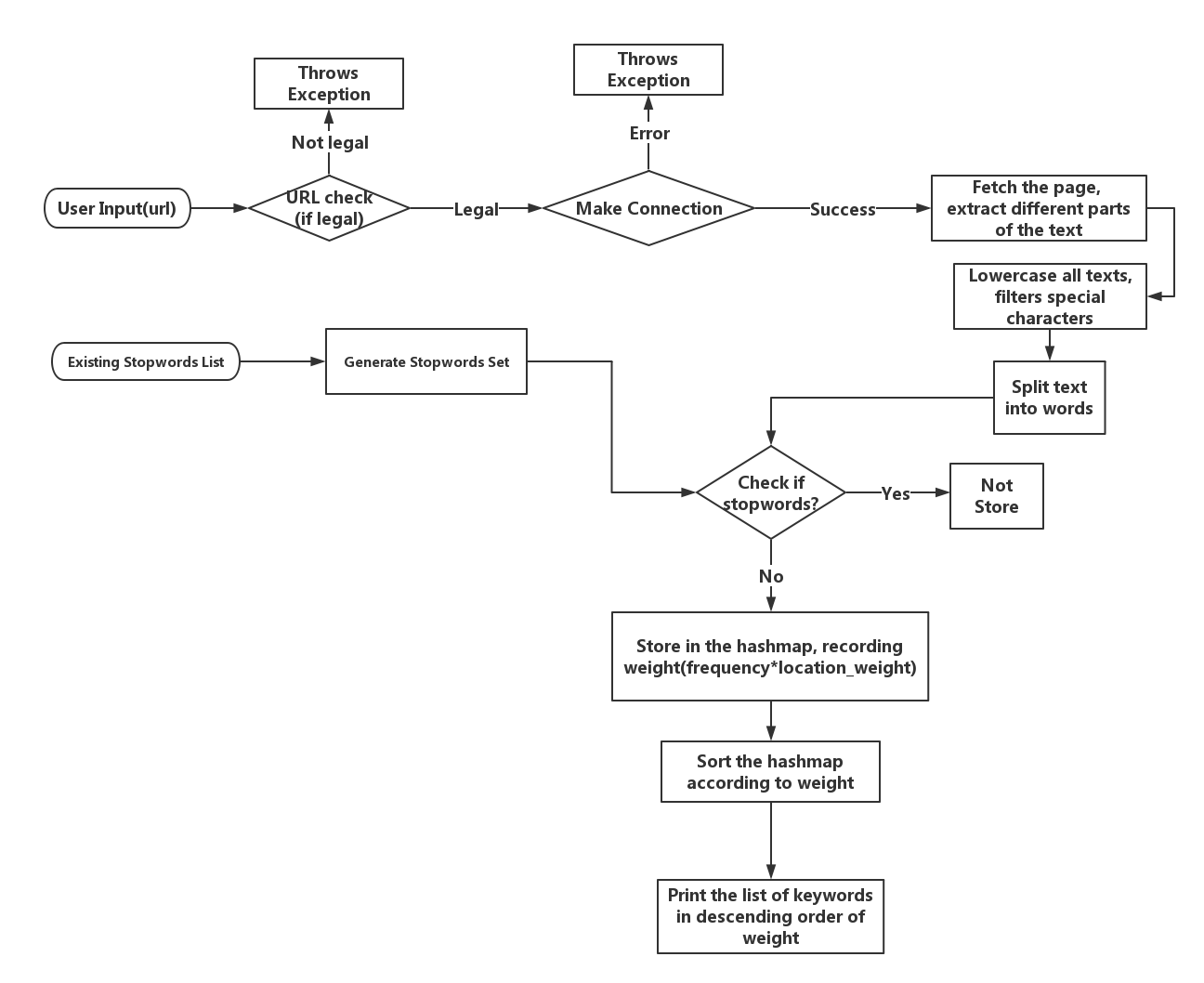
Word Density Analysis

# Purpose

The purpose of this program is:

* Receive users’ input, which is a web page’s url
* Parse the web page, extracting title, metadata(keywords) and body of the web page
* Split the text of the title, metadata and body into words, counting the frequency of each word in the texts
* Calculate the weight of each word with frequency and weight (based on location)
* Return a list of keywords and the corresponding weight in the web page.

# Design



# Implementation

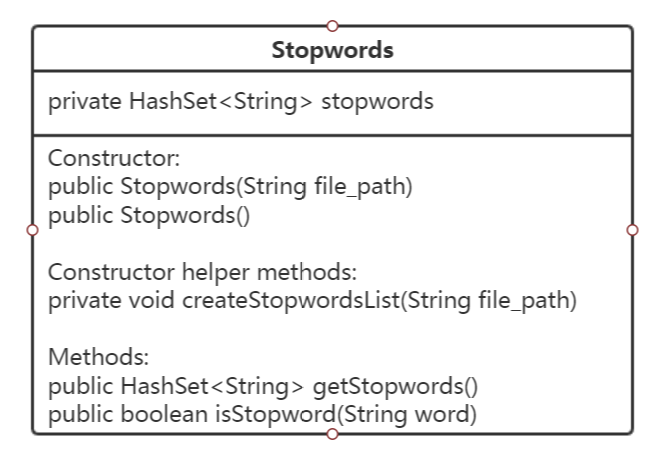
There are 6 classes in the program:

## Stopwords

Stopwords class generates stopwords set(Using HashSet<String> to store stopwords list)

input: file path of stopwords(optional)

output: a HashSet containing a list of stopwords

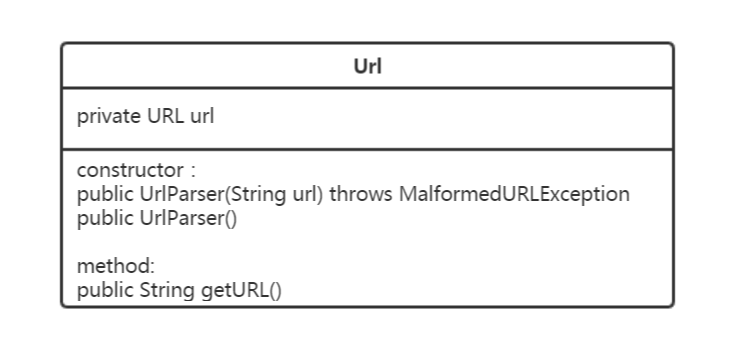


## UrlParser

UrlParser class: read in url and check if the url is correctly formed

intput:user input string

output:checked url(String)



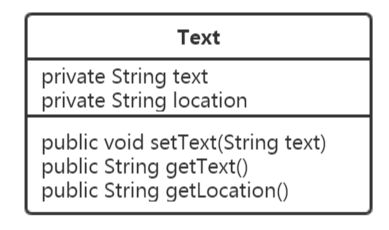
## Text

Helper class for PageParser

Store the text and location of the text(title,metadata,body)

Input:text contents(String)

Output:text contents(String), location(String)



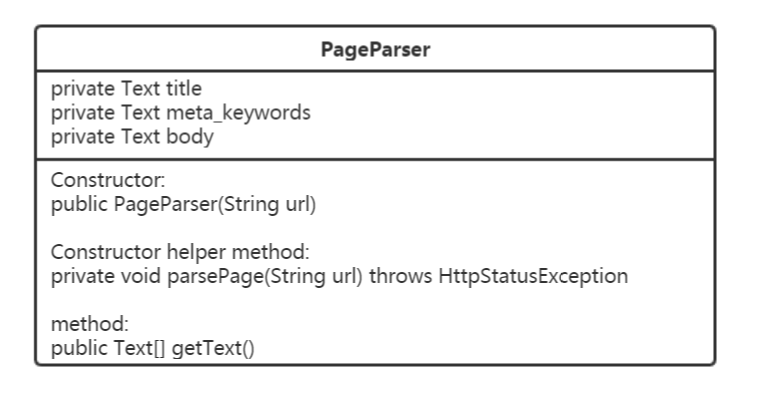
## PageParser

Import external library: Jsoup

The class fetches web page and extracts different part of page

input: url(String)

output: Text[] array(contains different part of text contents and location)



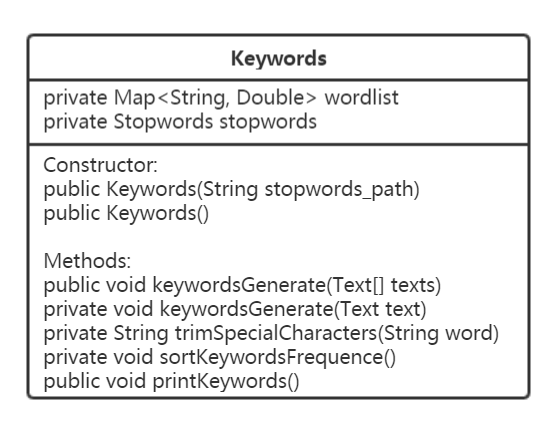
## Keywords

This class does following functions:

* process each part of text: lowercase, remove tags, split the text into words
* for each word, remove special characters
* store keywords and frequency(weight) into hashmap
* sort the hashmap
* print keyword list

input: stopwords list, each part of text

output: sorted keywords list



## Main

Only contains main method, the integration of all operations.

# Sample Output

Input: “<http://www.cnn.com/2013/06/10/politics/edward-snowden-profile/>”

Part of the output is as follows:

