**Keyword Density Analysis – Design Document (BrightEdge Challenge)**

**Jeris Alan Jawahar**

The main idea behind my design of the keyword density analyzer has been through assignment of appropriate weights to what keywords should be considered more relevant in comparison to others. In the due course of small research I did, some of the main contents of a HTML page used in SEO include the page title, page description, set of keywords associated to the page, any strong/bold content and finally the page content itself. Based on the above assumption, I assigned different weights to the varying page content:

1. Page Title – 5
2. Page Keywords – 4
3. Page Description – 3
4. Strong/Bold content – 2
5. Normal page content – 1

The page content was parsed into 1-gram, 2-gram and 3-gram and the keyword topics were assigned weights based on where they appear in the HTML page. Once all the n-grams were generated, the top 10 keyword topics having the highest weights are retrieved and displayed to the user.

There are primarily 4 classes in the whole project:

1. Parser – This class is used to establish a connection with the URL and once the page is fetched, it is parsed and the respective content information of different sections are populated here. The content is also processed here to ensure that the word density analysis can be properly run on the content at a later stage in the pipeline.
2. Stopwords – This class is used to take care of the stop word processing of the HTML page content.
3. Constants – This class stores the necessary constants used in the application.
4. Analyzer – This class performs the main analysis task by generating the n-grams and identifying the most relevant topics for the given URL. It is also used in display of results.

Requirements:

Ensure that either Java 1.7 or Java 1.8 is installed on your host system. To check the java version, enter the following command in your command prompt or terminal – “java -version”

After confirmation of Java installation, Download the kda.jar file for the github repo “https://github.com/jerisalan/KeywordDensity” and enter the following command – “java -jar kda.jar <URL>”. You should see the top 10 relevant topics being listed out for that URL specified.

References:

1. <https://stackoverflow.com>
2. <https://jsoup.org/>

External Libraries Used:

1. JSOUP – Java HTML Parser Library