Author: **Abhisek Mohanty**
Lab 5 Data Science CAP5771

Command in Terminal:

```
java -jar extract.jar enwiki-latest-pages-articles.xml output
```

DIFFICULTIES:

The most difficult thing in this project was getting used to the new Hadoop terminologies and setting up the main method. There were weird package issues and errors, which I somehow managed to fix. Once set up, the general logic was quite simple as it involves just two operations at each step, map and reduce.

The second difficulty was the step of removing dead links. The logic used in the program is as follows.

MapperXMLParser class:

- Parses the XML and finds Title and set of Links.
- Add the tuple <title, "~"> to the output first. (explained below)
- For each link, add the <link, title> to output, in the same order.
    - Adding <link, title> instead of <title, link> as otherwise, we won't be able to compare the links with the titles from the entire dataset (to see if it's a red link), since each reducer will have only a subset of the data.

Explanation for above logic:

Let's suppose, we have three sets of data:

Title	List of Links
A	B1, B, C
B	A1, A, C
C	B1, A1, A, B

It is clear that A1 and B1 are red links. So the expected output after removing red links is:

A - B, C ;
B - A, C ;
C - A, B ;

The logic in MapperXMLParser outputs the following:

A	~
B1	A
B	A
C	A
B	~
A1	B
A	B
C	B

C	~
B1	C
A1	C
A	C
B	C

Once the combiner is done processing the above output from mapper, the key value pairs are as follows:

A – { ~, B, C }
B – { ~, A, C }
C – { ~, A, B }
A1 – { B, C }
B1 – { A, C }

Next the logic in the **ReduceXMLParser** does the following:

ReduceXMLParser :

- For each key and values from the combiner, check if “ ~ ” is present in the values.
- If “ ~ “ is present
    - Add each <key, value> to output, except the ~ symbol
- If “ ~ “ is not present
    - Ignore, don’t add. (Ignores the sets in **BOLD** above)

So, the output is:

A – { B, C }
B – { A, C }
C – { A, B }

Which is same as the expected output.

This is then consumed by the GraphMapper and GraphReducer classes.