DATASCIENCE – REPORT

SUNIL KUMAR
94018098

REDLINKS EXTRACTION:

- Mapper 1: While parsing the xml, for each <page>, create a tuple <title, #> for every title. For every link create a tuple as <link, title>. If a page has no links then create a tuple as <title, ZERO>
- Reducer 1: The reducer receives a list of key-value pairs as tuples. If a list does not contain a tuple <title, #>, then this list of links are a redlinks and it is discarded. Then we create tuples as <title, link> and pass it to the next mapper.

ADJACENCY GRAPH:

- Mapper 2: This creates tuples as <title, link> and sends to reducer 2.
- Reducer 2: For each title we append all the links and create the adjacency graph of title and links.

DIFFICULTIES:

- Hadoop Map Reduce documentation and setup was very confusing. The concept was difficult to understand and required time but easier once we understand the concept.
- Coming up with a strategy to remove redlinks was pretty challenging.

Changes made to match TA's output:

Assuming the wikilinks are case insensitive, we converted all the links and titles to lowercase. But later I noticed that the links are case insensitive. Hence removed the tolower function to match the TA's output.