Problem Statement

Stack Exchange is a network of question-and-answer websites on topics in diverse fields, each site covering a specific topic, where questions, answers, and users are subject to a reputation award process. The reputation system allows the sites to be self-moderating.

The data files needed are attached.

You can find the metadata information here. You can also google if you have any confusions.

https://ia800107.us.archive.org/27/items/stackexchange/readme.txt

Create a **MapReduce** program for the questions below. You can run your code on standalone (local) Hadoop installation. The process is explained in the slides in file Hadoop Introduction. A lecture with standalone demo video is also uploaded.

1. Given a keyword, identify the comments about the keyword to be either positive or negative collectively. This can be accomplished by first looking up each word in the comment (about the keyword) in the bag of positive and negative words i.e., positive.txt and negative.txt files respectively. The process is repeated for all the comments and collectively a positive or negative score can be used to classify the keyword to either used positively or negatively. Use Comments.xml file.

Example:

Comment 1: iBooks is free, Good Reader is inexpensive

Comment Sentiment: Positive as most words are found in the bag of positive words.

Comment 2: iBooks is a average alternative.

Comment Sentiment: Negative as most words are found in the bag of negative words.

Comment 3: I've tried downloading iBooks but it hangs a lot.

Comment Sentiment: Negative as most words are found in the bag of negative words.

Keyword: iBooks used negatively overall.