Exercises
Data Mining: Learning from Large Data Sets
FS 2016

Series 1, Sep 29th, 2016 (MapReduce)

LAS Group, Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `http://las.ethz.ch/teaching/dm-f16`
Email questions to:
Mario Lucic, lucic@inf.ethz.ch

It is not mandatory to submit solutions and sample solutions will be published after one week. If you choose to submit your solution, please send an e-mail from your `ethz.ch` address with subject `Exercise1` containing a PDF (LaTeX or scan) to `lucic@inf.ethz.ch` until Wednesday, October 5th 2016.

Problem 1 (Approximation of the English dictionary):

In this exercise you are asked to construct an approximation of the English dictionary using a number of books written in English. The goal is to obtain a sorted list of words with their counts. We also want to make sure that all words are lower-cased and contain only letters from a-z. For example, if the only provided book contains the text "This is a (very, very) short 'book'. It is only 2 sentences long.", the output should be:

word	**count**
a	1
book	1
is	2
it	1
long	1
only	1
sentences	1
short	1
this	1
very	2

Your task is to modify the Word Count MapReduce example shown in the recitation session to incorporate the constraints discussed above.

Problem 2 (A basic English dictionary):

For some Natural Language Processing tasks you have to pre-process the data set by removing the most common words (stopwords). For the English language some examples are "the", "and", "if", "which", and "on". Your second task is to construct a dictionary such that the following constraints are met:

- There are at most 30 words for each letter.
- Each word in the dictionary has appeared at least **A** times, and at most **B** times in the data set , for some predefined **A** and **B**.
- For each letter the words are sorted alphabetically.

For example, if the subset of the output of the first exercise had been

word	count
a	788
all	123
antenna	9
auto	33
ball	15
beach	30
by	211

then for $\mathbf{A} = 10, \mathbf{B} = 35$ the final output of your MapReduce program should be

word	count
auto	33
ball	15
beach	30

You task is to write a `map` function and a `reduce` function in Python to solve this problem.