CONFIDENTIAL - DO NOT SHARE

Estimating number of distinct elements in data streams

Please watch this video explaining what min-hash signatures are. The relevant parts are 22:00 - 25:30. We would like you to implement that algorithm and use to estimate the number of distinct elements in a sequence.

Here are the functions we would like you to implement. The signatures are in Python for readability, but **please use any language you are comfortable with to implement the code**:

```
# This function will return an estimate to the number of distinct elements in items
# items - a sequence of elements
# k - number of hash functions
def estimateDistinctElements(items, k):

# This function will return an estimate to the number of distinct elements in items (same as
# above) with the distinction that listsOfItems is now a list of sequences. The idea behind this
# function is to generate partial estimates on every sequence and then combine them into a
# single estimate. Note: this function should simulate a distributed environment, so
# assume the list of sequences is just an abstraction, and that every individual sequence
# is on a different physical machine and that the combined size of all those lists cannot
# fit on any single machine.
# listOfItems - a sequence of elements
# k - number of hash functions
def estimateDistinctElementsParallel(listsOfItems, k):

# This function will return the difference between the estimate and the actual number of distinct
# elements in items
# items - a sequence of elements
# estimate - a number that represents the estimate (the answer from the function above)
def calculateEmpiricalAccuracy(items, estimate):
```

If you have any question, please send them to datascience@salesforce.com with the email title being **Distinct Elements Estimate Assignment (Scala)**

Thanks and best of luck,
The Secret Data Science Team