Given: Text Data size is 500 GB, words separated by space and laptop with 8 GB RAM.

Problem: Estimate the number of words that repeat 42 times in the entire document.

- **Sai Kasyap**

Method 1:

As, the data is large, we take a sample of data using **random sampling** say up to 8 GB of data (can take larger data but processing time will be longer for example: with checkpoints). As the requirement is estimate the words which appear 42 times, it implies that that the words are rare/sparse.

This is a **Multinomial Distribution** where each word's probability is exponentiated by its count, then we have calculated **Maximum Likelihood Estimate** of the words. Multiply the estimates of words to entire count of words in the 500 GB data and take the words whose values correspond to 42.

The other metric such as average log likelihood can also be calculated. To **dea**l with **unknown words**/ unigrams we have use techniques like smoothing say **Laplace smoothing** where in we add an artificial unigram a pseudo count to all the unigrams.

If the distributions of unigrams are different in the sampled data and the entire data, then we must **vary interpolation weight** to reduce the variance in the estimate.

Ref : https://medium.com/mti-technology/n-gram-language-model-b7c2fc322799

Method 2: Use of Vaex : A python open source data frame library using concepts like memory mapping, lazy evaluations etc.   to handle datasets which are too large fit in RAM. (https://github.com/vaexio/vaex)

Once we can access the entire data, we can use NLP techniques like count vectorizer or dictionaries to figure out the words which have frequency 42.

Please refer the articles for additional information:

https://towardsdatascience.com/ml-impossible-train-a-1-billion-sample-model-in-20-minutes-with-vaex-and-scikit-learn-on-your-9e2968e6f385

https://towardsdatascience.com/how-to-analyse-100s-of-gbs-of-data-on-your-laptop-with-python-f83363dda94

Method 3: Use of efficient data structures like Synopsis data structures and streaming algorithms to access all entire data set and use NLP techniques like count vectorizer to figure out the words which are 42.

Ref: http://homes.sice.indiana.edu/yye/lab/teaching/spring2014-C343/datatoobig.php