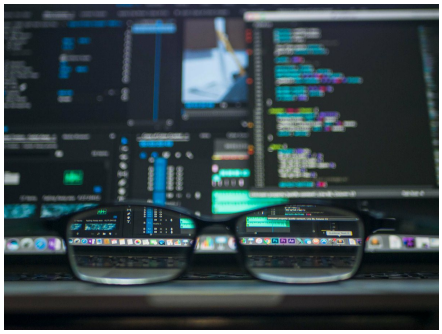
ALLAN THRAEN | 5 years ago | PDF |

.NET Development C#

READING VERY LARGE GZIPPED JSON FILES IN C#



This is a little code snippet that I often find quite handy. It's a piece of c# code that opens a gzipped json file and iterates through the items in it. Since it takes it piece by piece (as opposed to loading everything in memory) it's can pretty much handle files of any size.

I love working with large datasets - and over the years I have collected a bunch of different useful test datasets. However, they often come in files too large to deal with / handle. Often it's a CSV, XML or Json file that has been gzipped - and where even the unzipping of it could pose a problem. And even when succeeded, it can be quite challenging to load a 100 gb file into a text editor.

This code works for a large gzipped json file - but could easily be adapted to work with other compressions and formats. For example, the JsonReader could easily be replaced by an XMLReader.

It uses Newtonsoft.Json and SharpZipLib (both available as nuget packages). Replace 'Element' with the type of the object you want to deserialize to.

If you want to try it out with a large dataset, you can find gzipped json dumps of Wikipedias search index here: https://dumps.wikimedia.org/other/cirrussearch/

```
using (System.IO.Stream fs = new FileStream(filename, FileMode.Open, FileAccess.Read))
using (GZipInputStream gzipStream = new GZipInputStream(fs))
using (StreamReader streamReader = new StreamReader(gzipStream))
using (JsonTextReader reader = new JsonTextReader(streamReader))
{
    reader.SupportMultipleContent = true;
    var serializer = new JsonSerializer();
    while (reader.Read())
    {
        if (reader.TokenType == JsonToken.StartObject)
        {
            var t = serializer.Deserialize<Element>(reader);
            //Add custom logic here - perhaps a yield return?
        }
    }
}
```

.NET Development C#

CodeArt ApS
Teknikerbyen 5, 2830 Virum, Denmark
Email: info@codeart.dk
Phone: +45 26 13 66 96
CVR: 39680688