Project Approach Document

Project name: **Utility to remove sensitive information**

We broke our problem into substituent parts and we have summarized our approach, as per our understanding of data redaction and sanitization of the sensitive information

1. All the zip files will be unzipped by using 7-zip (an open source software) , integrating it with our .net application which we are going to build.
2. Start updating the ANSI format files (like .txt) by searching with regular expressions and removing the sensitive strings ( such as Email Id , Phone no., Address, CIN no. , Service tax no ., Pan no., etc N.B. the names can’t be well identified using regular expressions and the details for the algorithm to remove them are discussed later ) with a symbol ( like ### ), and create a log file listing the files which are updated.
3. For .docx , .pptx, .xlsx , .doc , .xls , .ppt , we will use Microsoft open xml sdk ( open source ) which will be integrated with our .net application. Now we will search for particular regular expressions (Wild Card Strings using which email addresses and phone no. etc can be identified) through the open xml sdk 2.5 libraries and update them with symbols, the data of which will be listed in the log file again.
4. Meta data information such as author, subject, keyword or title which are a part of the office file properties will also be minimized to our best.
5. For company names, and other names of individual we can compare with English dictionary since most Indian names are in different languages so they won’t match with dictionary words and do our part of replacing them with symbols but it will not have good accuracy. The company names can be removed if phrases like “Pvt. Ltd.” is found in the document.
6. For removal of the Name type of data, for better accuracy of the system from the aforesaid method, we would like to maintain a database of names ( which may be upto to 1 lac , n.b. No websites for Indian names have more then 60 k names ) which could be indexed and then the document will be searched for those occurrences but this process would be very slow if some manual data is not provided. If manual data is entered once such as the employee names who worked for a particular project and companies involved then the database handling would be easier and those names and details can be removed from the files with better precision.
7. Generally the header and footer information are used to indicate copyright details and company name, logo etc. As per our next understanding of the problem statement, we would remove headers and footers from the open xml file types.
8. We would remove any kind of watermark which are generally used to provide the company information.
9. All the changes made will be put into database or a log file indicating the names of the file which

are updated.