Answer 1.

Explanation for the approach to check plagiarism

a) **Collection of webpages and previous student assignment from which current student might copy their assignment**

  - Through Focused Crawling approach download the web pages, which are related to the topic of the assignment.
  - Crawled pages about a topic should have a links to other pages on the same topic.
  - Use of the text classifier to verify the crawl page is related to topic or not.
  - Desktop crawler to collect all the previous student assignment present into the local repository about the particular topic.

b) **Conversion of the file into the HTML or XML format and simultaneously do character encoding using conversion tool.**

  - Convert the text, which is stored into MIRCROSOFT WORD, ODF, PDF, EXCEL and POWERPOINT into the HTML or XML format.
  - Use UTF-32 for internal text encoding for maintaining compatibility and to save space.

c) **Now Store the many converted text document in large files instead of storing each individual document in a file**

d) **Use Compression technique to eradicate the redundant data to compress the file size without losing any of the content.**

e) **Now SimHash processes files and stores the hash keys and sum table values in a relational database.**

f) **Steps that SimHash follows to process the document**

  - Divide the document into the set of features with associated weights, features are words weighted by their frequency
  - Generate b-bit unique hash value for each word
  - For b-dimensional vector V, update the components of the vector by adding the weight for a word to every component for which the corresponding bit in the word's hash value is 1, and subtracting the weight if the value is 0.
  - Once all the words finished processing, generate a b-bit fingerprint by setting the ith bit or 1 if the ith component of V is positive, or 0 otherwise.

g) Now one file at a time, we perform SQL query on its key to find all other key values within a certain Threshold range. We set a single Threshold level, and multiply this value by the size of our target file, since we expect key values to increase proportionally to file size. For each file returned in this query, we first discard it if the file size differs too much from our first file. Next, we compute the distance between their sum tables, which is the sum of the absolute values of the difference between their entries. If this distance is within a certain tolerance, then we report the two files are similar.

Algorithmic complexity:
It should only take $O\ (logn)$ time to determine the number of key matches against a file, but if the number of matches is proportional to the total number of files, then we will need to perform $O\ (n)$ sum table comparison for each file.

Algorithm guaranteed to find the matches because the matching criteria depends on the threshold of the database if we will increase the database threshold than it will search for the more keys which having the same tag that the query is searching.

Yes the system can identify passage of various lengths. There is no specific shortest or longest passage that it can identify, the length of the passage is variable and it may vary from file to file.

Need to take care of the Spam while crawling the pages because sometimes during real time crawling, crawler administrator should check the pages whose data is similar to the data which crawler has already visited just few tags and text is different. To take care of these issues crawler should reduce the priority of the links from spam pages.