*Search Engine Optimization*

1. This process uses a web scraper to scrape and parse raw data from a list of URLs, filling a database with words, tags, etc; data that can be used to help us in the search engine optimization process. Therefore, the first thing that needs to be done is, using the list of URLs, run the web scraper, scraping and parsing the raw data from these websites.’
   1. First, make sure that you have a list of the URLs that you would like to scrape. These URLs should usually be links to the homepages of websites that are related to the company that we are performing Search Engine Optimization for. For example, if this is Miffy, the milk powder company, than our URLs should all be links to other milk powder companies. This list of URLs should be supplied in a .csv file and make sure to include the **entire** URL, including the “http://...”.
   2. Download the zip from the FrontCoding GitHub repository. Here is the link for the repository: <https://github.com/FrontCoding/scraper>. Check out the README for more information on how to run this web scraper.
   3. The runtime for the web-scraper (both scraping and parsing) can take many hours, and even over a day depending, therefore you should make sure that your computer will not Hibernate after just a few hours, is plugged in, and has a reliable connection to the internet.
   4. For faster scraping and parsing, it is advised that we distribute the scraping and parsing jobs to multiple machines. The tool can be downloaded and run on any computer that has PHP installed. If you do not have PHP installed, I recommend you download XAMPP.
   5. There are two methods for distributing the tool.
      1. The first and most effective method is to distribute the tool and have each machine hook up to a central database. The database connection can be configured in information.php. If all machines are connected to a central database, the “Master” machine can “drop” jobs onto the database while the “Slave” machines pick up these jobs and being parsing/scraping and populating the database. This method is more efficient because the system is set up so that these Slave machines will not overlap their jobs.
      2. The other, cruder method involves using each machines local database instead of a central database. This method involves manually slicing up the list of URLs and distributing these to machines individually. Then on each machine, you can run two threads, one thread which represents the Master, dropping jobs into a local database, and one thread which represents the Slave which does the scraping/parsing. After the whole process is done, simply export each machines database and import all of these into one central database so that you can have all the data in one central location.
   6. After the scraper/parser is run (again, see README for details on making this run), you should have a database full of thousands of webpages scraped, millions of words and their frequencies per page, and the contents of important html tags. All of this in one database, we can start to do some analysis.
2. Next, with all of the data in our database, we can start doing some analysis on these words. We want to get 300 most frequently occurring words in the database. If there is both English and Chinese, we want to get the 300 most frequently occurring words in English, and then again get the 300 most frequently occurring words in Chinese.
3. This can be done by using MySQL queries. We want to select maybe 900-1000 words at first so that once we delete all of the “unimportant” and very common words, we will have roughly 300 words. Here are the queries you can use.
   1. This first query returns the 1000 most frequently used words, along with the percentage of web pages they appeared on, their total frequency, and the average occurrence per web page. It also outputs this into a text file
      1. SELECT word, count(\*)/(SELECT MAX(id) FROM webpages) as webCount, sum(frequency) as Freq, sum(frequency)/(SELECT MAX(id) FROM webpages) FROM words\_parsed GROUP BY word ORDER BY Freq DESC LIMIT 1000 INTO OUTFILE 'company\_words.txt' FIELDS TERMINATED BY ',' OPTIONALLY ENCLOSED BY '"' LINES TERMINATED BY '\n';
   2. This next query gets the host URLs of the domains that were scraped along with the number of webpages within each host. Essentially, this gives us the size of each host that we scraped. This is also important for SEO, and should be included in the final report as well.
      1. SELECT w.host\_name, count(r.host\_id) as theCount FROM relationships as r, web\_host as w WHERE w.id = r.host\_id GROUP BY r.host\_id ORDER BY the Count DESC INTO OUTFILE 'company\_urls.txt' FIELDS TERMINATED BY ',' OPTIONALLY ENCLOSED BY '"' LINES TERMINATED BY '\n';
4. This data should be attractively put into tables and included in the report that we will send to the client. The report should be in the following format:
   1. Introduction: Here we should introduce what Search Engine Optimization is and how we plan to do it. Here is an example of a good introduction paragraph. This should be modified to fit each client differently.
      1. Google and Baidu searches can be some of the best, and cheapest, advertising. These search engines will look at the content of a webpage and determine how relevant the content is to the user’s search. Therefore, a website can be strategically designed to make these search engines consider it “more relevant” than its competitors. In order to have your company’s website listed higher up on a page of search results, we can employ a technique called Search Engine Optimization. By including keywords that Google and Baidu consider relevant to the search, your website’s listing can be pushed up the page, thus increasing your positive traffic. We analyzed the content of your competitors’ websites, scraping millions of words from thousands of web pages, and determined that the following words were the most frequently used. By including these keywords when designing your website’s content, we can optimize your search engine efficiency. While we have included the keywords that we find to be the most important, we of course want to know which keywords you find important. In the attached Excel document please list 20-30 additional words that we did not include that you find are very relevant and important to your website. Then, please send this Excel file back to us for further analysis. From the information we provided as well as the information that you will provide, we can further narrow down our analysis and find even more powerful keywords. Thank you.
   2. Visual Diagram of Process: The next section should include a visual diagram which lays out this process.
      1. See William or Sunny for this visual diagram. It can also be found in old reports.
   3. Our Target Websites: Then, we need to include a section called “Our Target Websites” which includes a short description followed by a table of the URLs that we scraped with the size of each website. Here is an example of a good description.
      1. Below, we have listed the websites which we have analyzed for you in order to determine the most frequently used keywords in this industry. These websites are also listed with the number of pages within that website’s domain name. This can give you a good idea of which websites are the biggest and which websites are relatively small. However, keep in mind that some pages, depending on their formats, could not be analyzed, so some websites (especially those near the bottom) may appear smaller than they actually are. We believe that the pages scraped from these websites form a very comprehensive image of the online content in this industry, allowing us to provide a very accurate and effective keyword analysis and ultimately perform search engine optimization.
   4. Keyword Lists: Then, include the full list of keywords that we found.
   5. In addition to this report, we need to send the client an Excel file in which they can write 20-30 of their own keywords that they find important. This Excel file can simply have a highlighted area directing them to include these words. After they fill out this Excel file, they should send the Excel file back to us for further analysis. To make the instructions clearer to the client, include a short set of directions in the email that we send to them. Here is a good example of instructions:
      1. In this step, we wrote a script, which indexed all of the words on the competitors' websites. After removing the useless words, we counted the total frequency, the percentage of all web pages the word occurred on, and the average number of occurrences per page in order to determine the most relevant words. The higher the words are positioned in the list provided in the report, the more frequently they are being used by the competitors.  
         Now, these competitor keywords are just an initial indication of the most effective keywords, so together with a list of your words, we are going to measure each word's Quality Score and Price Per Click (measurements by Google and Baidu that reflect the overall relevance and web traffic that a word garners). With this data, we can see which words carry the most potential, and which words should be used in your title, content or meta data in order to make Miffy Milk Powder highly ranked in the Search Engines.  
         So we are asking if you can deliver us words, separately for English and Chinese (20-30 for each), so we can execute step 2.  Please include these words in the attached Excel sheet, and deliver it back to us.
5. Once the Client returns their recommendations of keywords, we can use this in conjunction with the list that we created to determine which words are the MOST important words and which are less important. The way we are going to do this is by using Google Adwords and Baidu to determine each words quality.
   1. Google Adwords
      1. Create a Google Adwords account (this should be easy if you already have a Gmail account). It will likely ask you for information about your company. Just make up random information as we do not have the rights to actually make an account for our client.
      2. Import our list of words.
      3. Import the list provided by our clients
      4. In addition to the words that we important, we want to start thinking of clever combinations of words that might increase the specificity and quality of these words. We can even create short phrases. There is a section where you can manually add more keywords. Looking at the list that we already have, use your best judgement to combine words together to make relevant phrases. For example, if our list already has the keywords “milk powder” and “fresh”, then we could combine these to make “fresh milk powder”. Furthermore, if we also have “children”, we could combine all of these to create the short phrase “fresh milk powder for children”. Therefore, after adding these extra words, we should have somewhere between 400-500 words per language on Google Adwords.
      5. Google Adwords will automatically calculate the Cost Per Click (CPC) and the Quality Score. These are good indications of how important Google thinks these words are. Finally, export this information to an Excel file to include in the final report.
   2. Baidu Paid Search
      1. Baidu can be more difficult to use, so I recommend getting a Chinese person to help you.
      2. Make an account.
      3. Follow the same steps as with Google Adwords.
      4. Export the words to an Excel file to include in the final report.
   3. Putting it all together. We now need to include this in a final report for our clients, telling them once and for all which words are the best. Using the Quality Score, determine a cut-off for what you consider a high quality and low quality word. In the past, we have used a Quality Score of 7 as a the cut-off for good words. So, only including words with a Quality Score above this cut-off, we can give a list of the roughly 100 most important words in each language. Sorting by Quality Score, for each word also include its Baidu CPC for reference.
   4. The final report should be in the following format
      1. Introduction:
         1. Based on our Competitor Keyword Analysis, your specific word recommendations, and finally the Keyword Analysis and Report that we have included below, we have generated a series of recommendations that will help you optimize your website’s search engine results. Paying for advertising on Google and Baidu can be an expensive endeavor, but our recommendations included in this report can help you achieve similar results by strategically structuring the content of your web pages.
      2. Our Method: Briefly describe the method that we used.
         1. Google’s Adwords and Baidu Paid Search are useful services that allow companies to target advertisements towards certain keyword searches. These services use complex algorithms and extensive data resources to generate a price and determine the quality of these keywords and phrases. From the list of words that we generated from our Competitor Keyword Analysis and your recommendations, we use these two factors, price and quality, to help us determine which keywords and phrases are considered relevant by Google, Baidu, and other search engines. As a result, we can optimize your search engine results by focusing on only high quality keywords and phrases. Below, you will find a list of keywords and phrases. Each item has attributes “First Page CPC” and “Quality Score”. The “First Page CPC” is an estimate of the “Cost Per Click” (CPC) you would need to pay in order for your listing to appear on the first page of search results. The “Quality Score” is Google’s measurement of your webpage’s relevance to a user who is searching using that keyword, taking into account estimated traffic as well as the likelihood that a user will click through to your advertisement. More specific words tend to have higher Quality Scores, and highly trafficked words tend to have higher First Page CPCs. Quality Scores and First Page CPCs are very good indications of the effectiveness that a keyword or phrase will have in improving your search engine results. By focusing on keywords or phrases that have higher Quality Scores and higher First Page CPCs, we can signal to Google and Baidu that your website contains relevant material. The table containing these keywords and phrases is listed below. It is sorted first by “Quality Score”, and then by “First Page CPC (Google) in RMB”, putting the more effective keywords and phrases towards the top. For your reference, we have also included the original Competitor Keyword Analysis to the end of this report.
      3. Our Recommendation: Make sure that you replace specific examples in this example paragraph with words and examples that are more relevant to the client.
         1. Based on this analysis, we have concluded that the keywords and phrases that have a Quality Score of 7 or higher are the most effective. While all of these keywords and phrases can be used to help optimize your search engine results, the most focus should be placed on those with a Quality Score of 7 or higher. While we emphasize Quality Score the most, it is also important to consider the Front Page CPC which is a measurement of the total web traffic and popularity that the keyword or phrase receives. There is no need to pay for Google or Baidu’s advertising services; you can improve you search engine results by merely keeping this information in mind when writing the content of your webpage. We recommend that within <h1> tags, <h2> tags, <title> tags, and any other HTML tag that emphasizes text, you include these keywords and phrases whenever possible. If you cannot include these keywords and phrases verbatim, minor variations (such as breaking up a phrase with a few words in between) will also be effective. For example, in your content, when using the word “pipe”, Google or Baidu will favor your webpage if you also include more specific words like “steel”, “welded”, or “ssaw” nearby or on the same line. As another example, on a page about pipe coatings, it would be very beneficial to include “CSA-Z245.21” wherever possible as Google and Baidu will heavily favor your website when this word is searched. In order to target a larger audience be sure to also include more general and popular words (higher First Page CPC), like “Grade B”. Of course, be sure to only do this when it is relevant to your content as Google and Baidu also have functions that will disfavor your webpage if it thinks you are “cheating” the system.
      4. Keyword Report: This should be a table with the 100 most important words in each language. This table should be in the following format:
         1. Word
         2. First Page CPC (Google)
         3. Quality Score (Google)
         4. First Page CPC (Baidu)
      5. In the Final Report, at the end also include any previous reports that have been sent to the client.

**That’s it!**