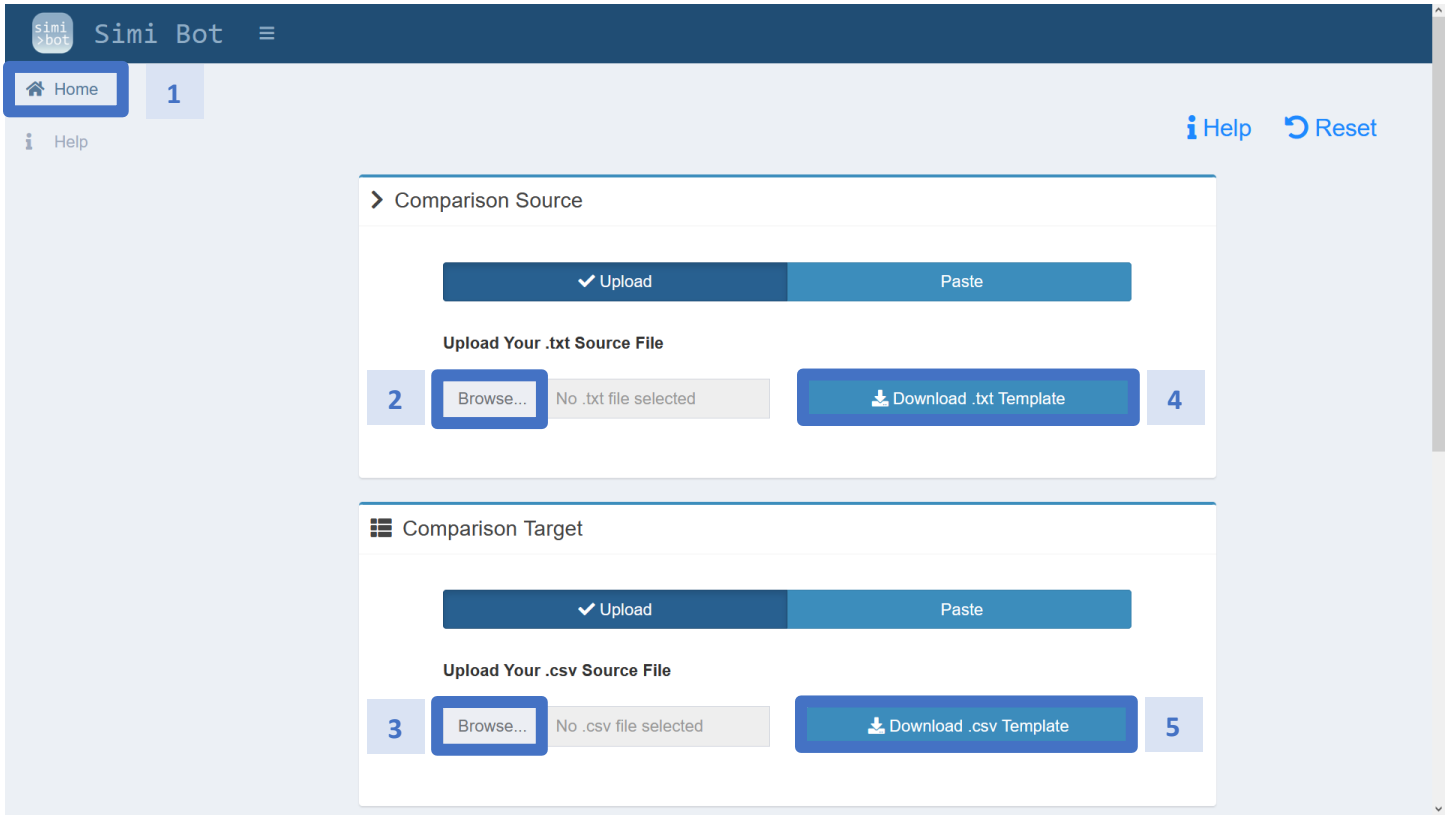
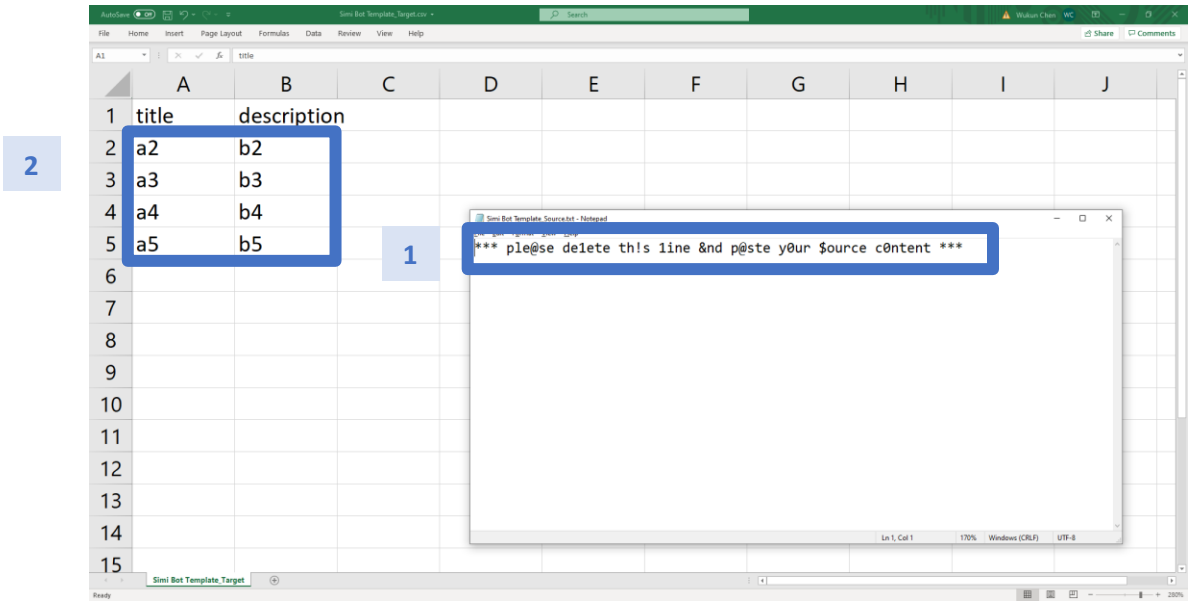
Simi Bot Tutorial

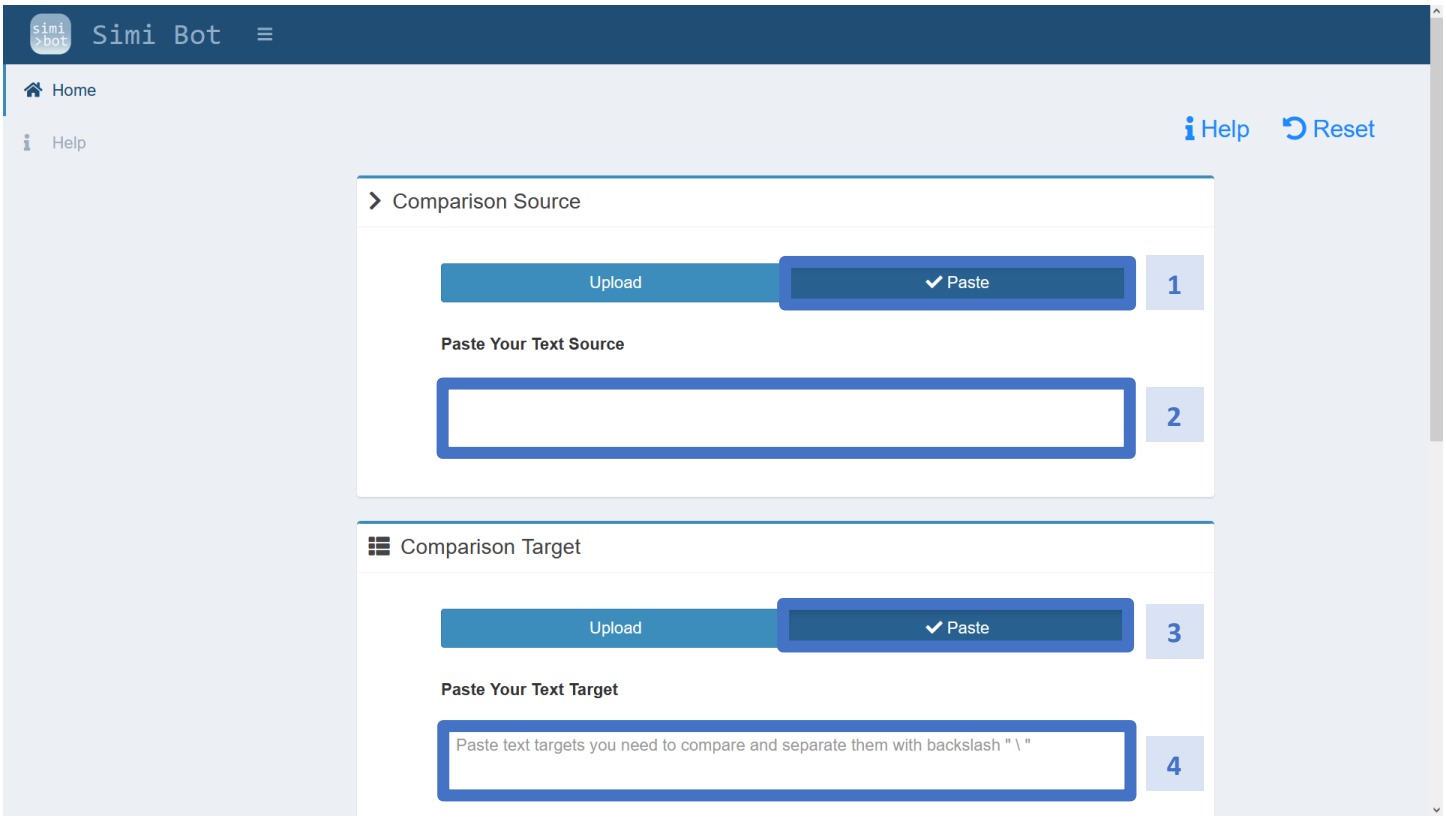
1. On the first tab **Home**, upload your source and target files. Download templates and modify them if you need. If you use your own csv target file, please make sure to input "title" in cell A1 and "description" in cell A2 as headers of data. The headers are critical to the following analysis.



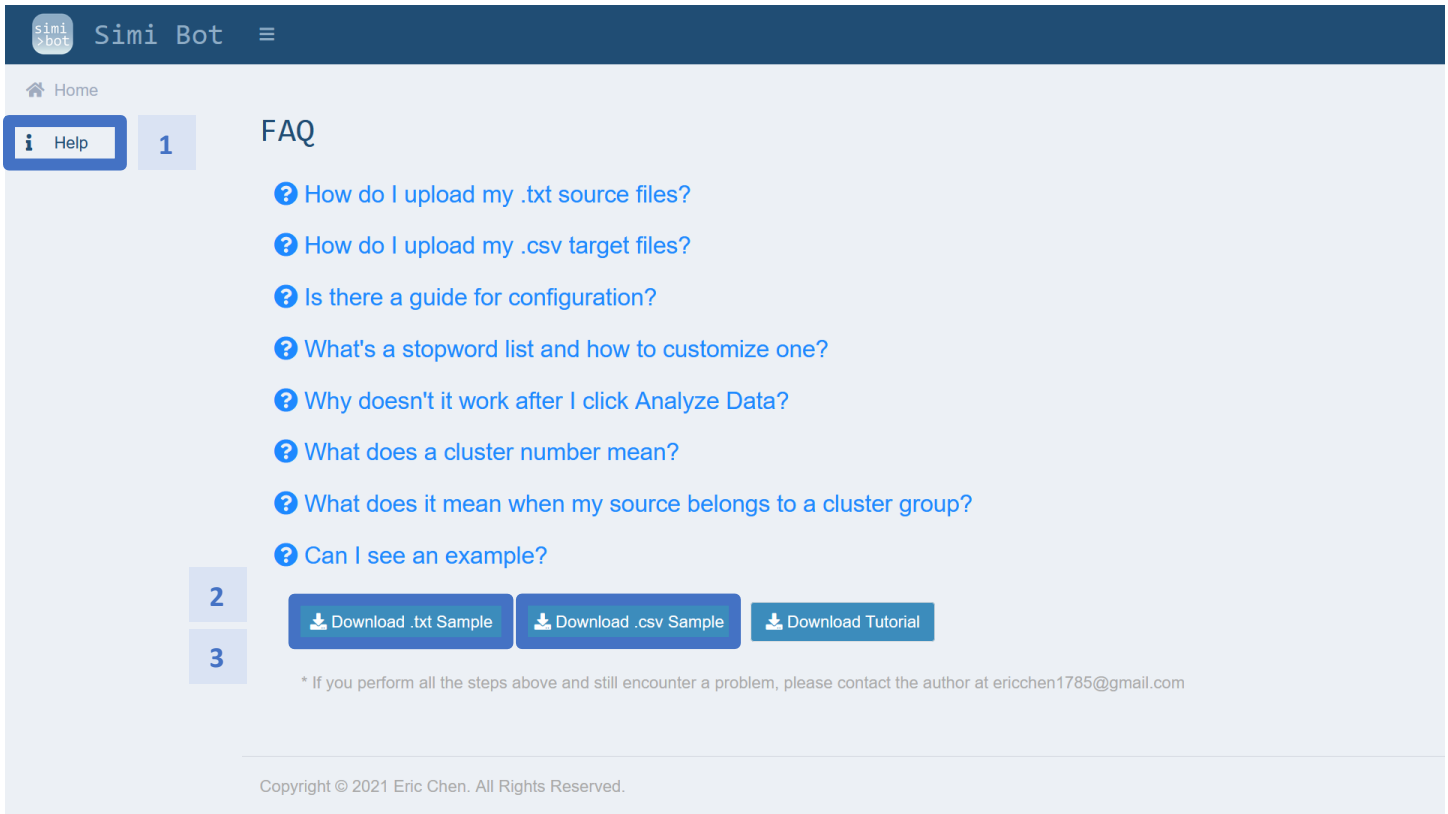
2. If you use the templates, open the templates *Simi Bot Template_Source.txt* and *Simi Bot Template_Target.csv*
   a. In *Simi Bot Template_Source.txt*, delete the placeholder line -> paste your text content -> save the file
   b. In *Simi Bot Template_Target.csv*, delete the placeholder contents in cell A2:B5 -> paste your text content in column A and B below the headers -> make sure your title and description are matched respectively -> save the file



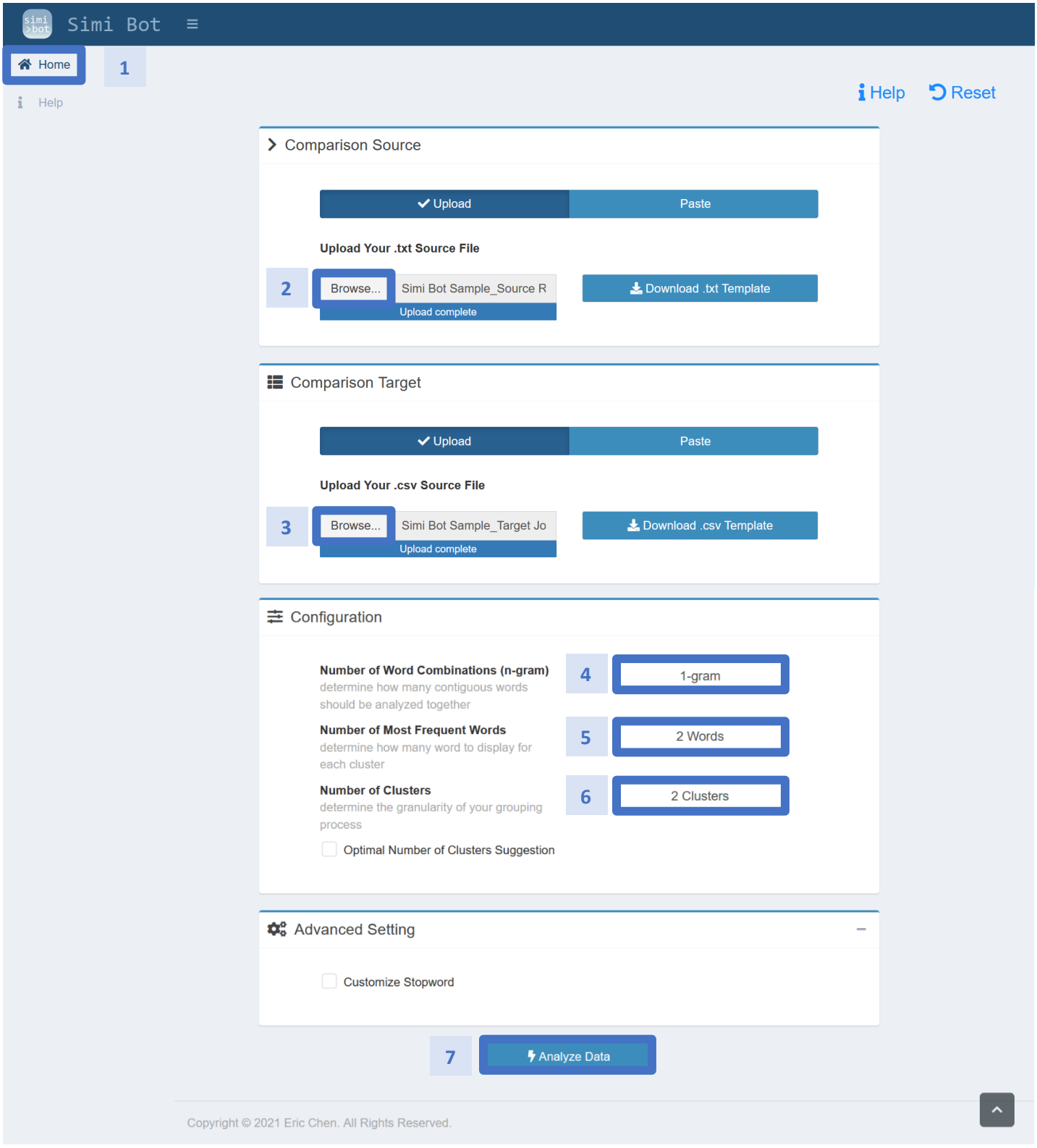
3. If you choose to paste your source and/or your target, please choose “Paste” in the accroding area. To separate every text target you paste, please use the backslash “\”.



4. On the fourth tab **Help**, download the samples *Simi Bot Sample_Source Resume .txt* and *Simi Bot Sample_Target Job Descriptions.csv*, which is the resume of the author of work and a list of 1000+ jobs. The following demonstration is based on these two files.



5. On the first tab **Home**, upload *Simi Bot Sample_Source Resume .txt* and *Simi Bot Sample_Target Job Descriptions.csv* -> configure 3 parameters -> click “Analyze Data”



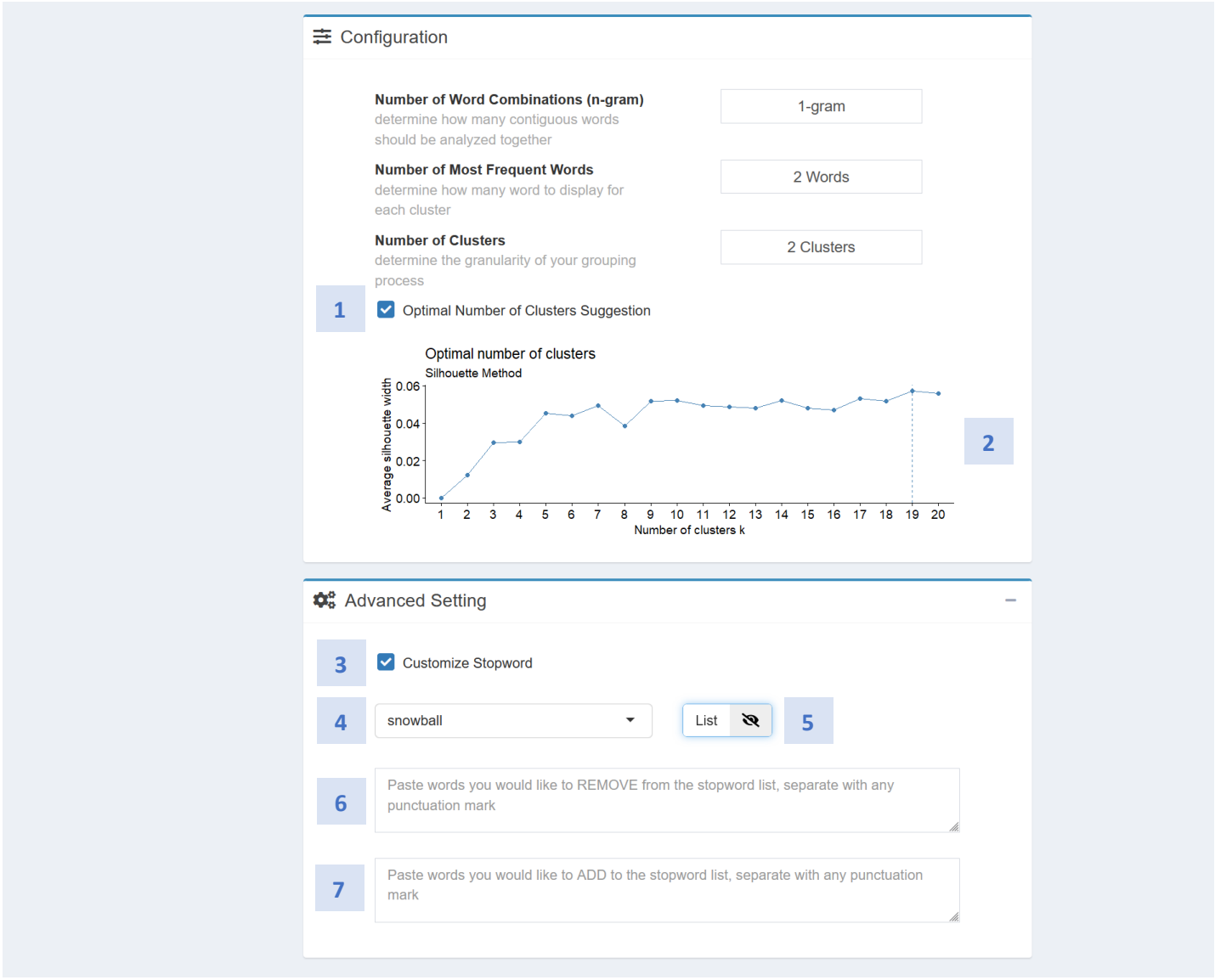
The *number of clusters* should not exceed the number of your targets. A practical advice is keeping it under half of the number of your targets.

If Simi Bot doesn’t start analyzing, please make sure you upload the right files.

6. If you have any advanced or customized need, Simi Bot provides suggestion for optimal number of clusters and stopword customization.

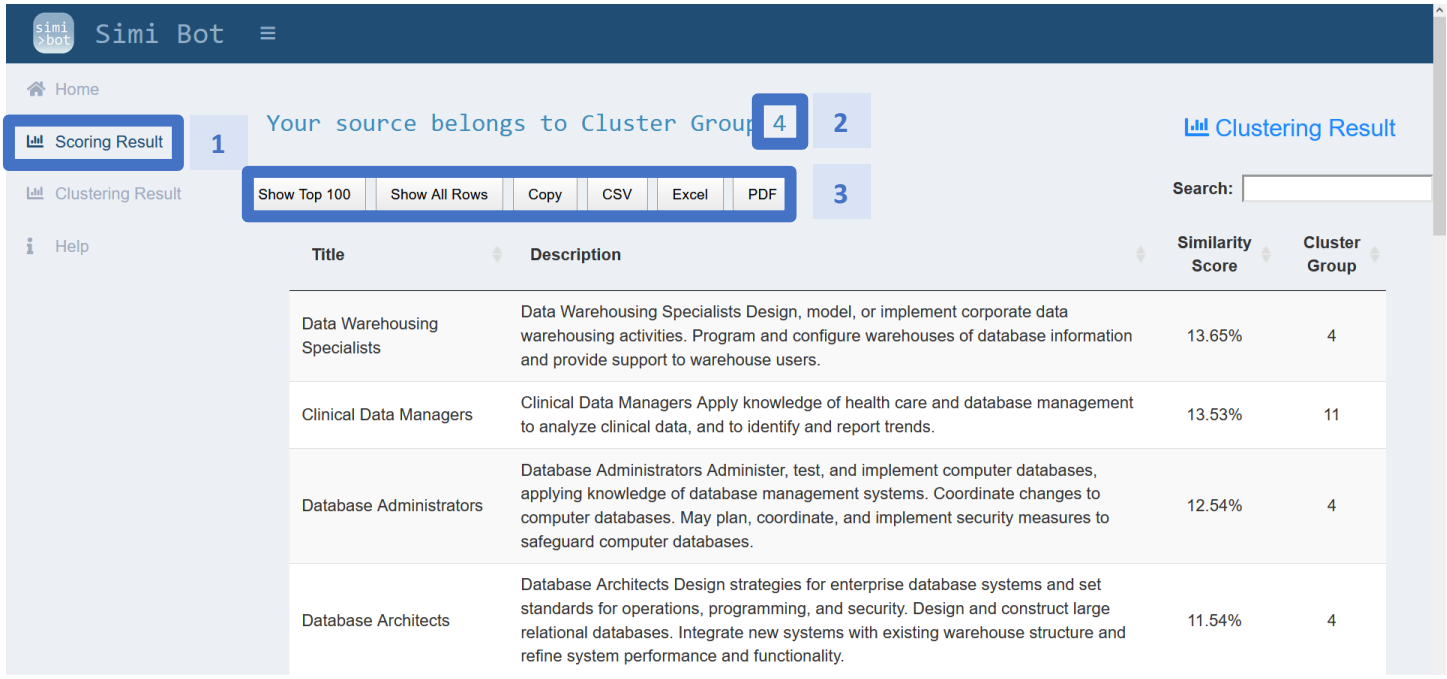
   The average silhouette width is computed and plotted for different number of clusters *k* from 1 to 20. The *k* with highest average silhouette width is marked with an auxiliary line. However, the *k* with highest average silhouette width isn't necessary be the optimal *k*. In the example below, the average silhouette width isn't significantly improved for *k* above 5. Like the *elbow method*, the optimal *k* is partially subjective. What we are looking for is a turning point where average silhouette width stop significantly improving. In this case, it could be 5, 3 or 7.

   Stop words are common words without concrete meaning. They could be the noise in text mining. As a part of the preprocessing, Simi Bot filters your text source and target with a stopword list (*snowball*). You can always choose another stopword list and view it. If there're any words that you want to compare but in the stopword list, you may remove them by pasting them in the first text area; if there're any words that you don't want to compare, you may add them by pasting them in the second text area. If an identical word is pasted into both text areas, it would first be removed then added to the stopword list.



In this tutorial, we'll skip the number of clusters optimization and stopword customization. We'll group the sample targets into 20 clusters and use 5 most frequent words to represent each cluster.

7. On the second tab **Scoring Result**, the first line tells you which cluster group you source belongs to. The following data table shows you the *title*, *description*, *similarity score*, and *cluster group number* of each target. The output is sorted in descending order of the *similarity score*. You could filter and sort the fields on the web, and/or download the result in your preferred format.



8. On the third tab **Clustering Result**, the data table shows you the *cluster group number*, *average similarity score*, *number of targets in group*, and *most frequent words in group* of each cluster group. The output is sorted in descending order of the *average similarity score*. You could filter and sort the fields on the web, and/or download the result in your preferred format.