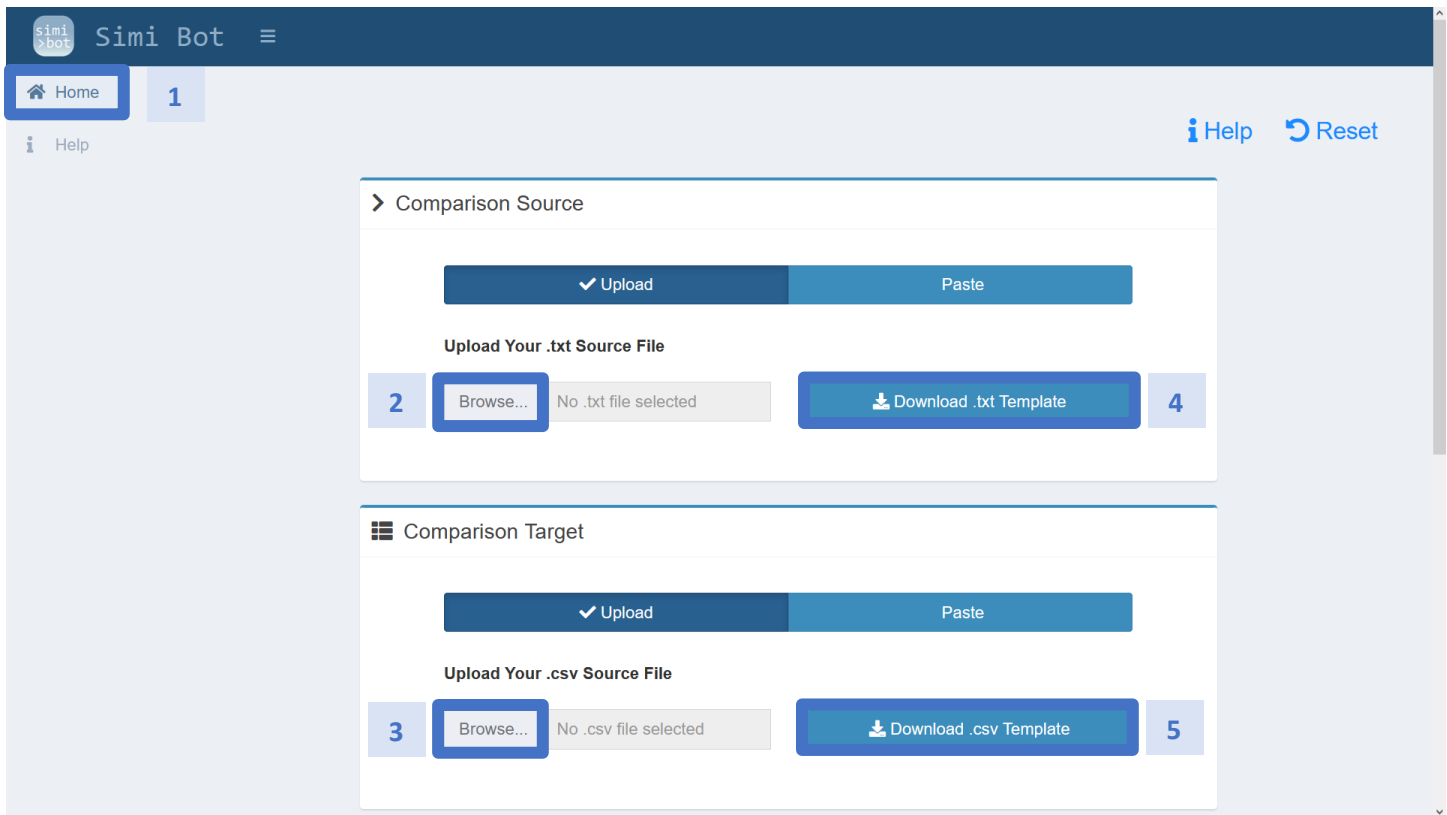
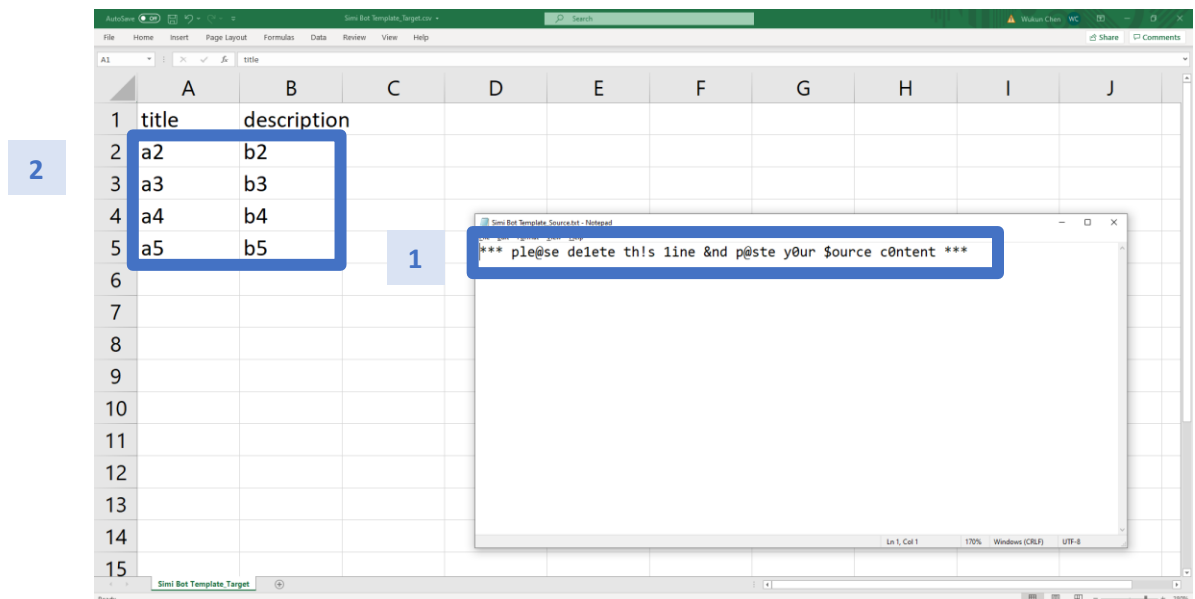


Simi Bot Tutorial

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 “\”.

The screenshot shows the Simi Bot web interface. At the top is a dark blue header with the 'simi bot' logo and a menu icon. Below the header is a light blue sidebar with 'Home' and 'Help' links. The main content area has two sections: 'Comparison Source' and 'Comparison Target'. Each section contains an 'Upload' button and a 'Paste' button (the 'Paste' button in the 'Comparison Source' section is highlighted with a blue box and has a '1' in a blue box next to it). Below each button pair is a text input field. The 'Comparison Source' input field is labeled 'Paste Your Text Source' and has a blue box around it with a '2' in a blue box next to it. The 'Comparison Target' input field is labeled 'Paste Your Text Target' and has a blue box around it with a '4' in a blue box next to it. The 'Comparison Target' input field contains the text 'Paste text targets you need to compare and separate them with 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.

The screenshot shows the Simi Bot web interface with the 'Help' tab selected. The 'Help' tab is highlighted with a blue box and has a '1' in a blue box next to it. The 'FAQ' section is visible, listing several questions. Below the FAQ is a section with three buttons: 'Download .txt Sample', 'Download .csv Sample', and 'Download Tutorial'. The 'Download .txt Sample' button is highlighted with a blue box and has a '2' in a blue box next to it. The 'Download .csv Sample' button is highlighted with a blue box and has a '3' in a blue box next to it. The 'Download Tutorial' button is highlighted with a blue box and has a '4' in a blue box next to it. The text input field contains the text 'Paste text targets you need to compare and separate them with backslash \"\"/>

* If you perform all the steps above and still encounter a problem, please contact the author at ericchen1785@gmail.com

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 screenshot displays the Simi Bot web application interface. At the top, there is a navigation bar with the 'simi bot' logo, the text 'Simi Bot', and a hamburger menu icon. Below this, a 'Home' tab is selected, indicated by a blue bar and the number '1'. To the right of the tab are 'Help' and 'Reset' links. The main content area is divided into three sections: 'Comparison Source', 'Comparison Target', and 'Configuration'. The 'Comparison Source' section has a '2' next to it and shows a 'Browse...' button selected, with a file named 'Simi Bot Sample_Source R' uploaded. The 'Comparison Target' section has a '3' next to it and shows a 'Browse...' button selected, with a file named 'Simi Bot Sample_Target Jo' uploaded. The 'Configuration' section has a '4' next to it and shows three input fields: 'Number of Word Combinations (n-gram)' set to '1-gram', 'Number of Most Frequent Words' set to '2 Words', and 'Number of Clusters' set to '2 Clusters'. Below these is an unchecked checkbox for 'Optimal Number of Clusters Suggestion'. At the bottom, there is an 'Advanced Setting' section with an unchecked checkbox for 'Customize Stopword'. A large blue button labeled 'Analyze Data' with a lightning bolt icon is at the bottom right, preceded by a '7' in a blue box. The footer contains the copyright notice 'Copyright © 2021 Eric Chen. All Rights Reserved.' and a small upward arrow icon.

simi bot Simi Bot

Home 1

Help Reset

> Comparison Source

✓ Upload Paste

Upload Your .txt Source File

2 Browse... Simi Bot Sample_Source R Download .txt Template

Upload complete

Comparison Target

✓ Upload Paste

Upload Your .csv Source File

3 Browse... Simi Bot Sample_Target Jo Download .csv Template

Upload complete

Configuration

4 Number of Word Combinations (n-gram) determine how many contiguous words should be analyzed together 1-gram

5 Number of Most Frequent Words determine how many word to display for each cluster 2 Words

6 Number of Clusters determine the granularity of your grouping process 2 Clusters

☐ Optimal Number of Clusters Suggestion

Advanced Setting

☐ Customize Stopword

7 Analyze Data

Copyright © 2021 Eric Chen. All Rights Reserved.

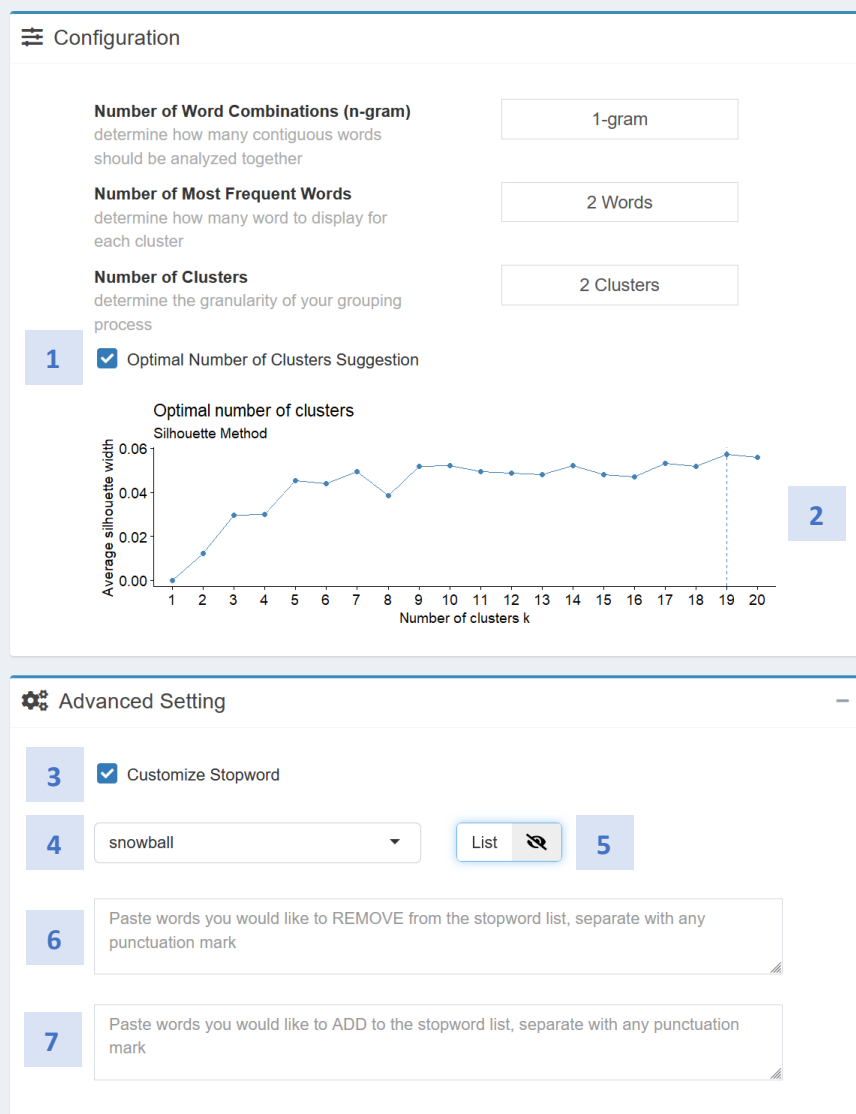
The *number of clusters* should not exceed the number of your targets. A practical advice is keeping it under half of the number of your targets.

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 stopwords 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 stopwords list (*snowball*). You can always choose another stopwords list and view it. If there're any words that you want to compare but in the stopwords 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 stopwords list.



In this tutorial, we'll skip the number of clusters optimization and stopwords 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.

Simi Bot

Home

Scoring Result

Clustering Result

Help

Your source belongs to Cluster Group 4

Show Top 100 Show All Rows Copy CSV Excel PDF

Search:

Title	Description	Similarity Score	Cluster Group
Data Warehousing Specialists	Data Warehousing Specialists Design, model, or implement corporate data warehousing activities. Program and configure warehouses of database information and provide support to warehouse users.	13.65%	4
Clinical Data Managers	Clinical Data Managers Apply knowledge of health care and database management to analyze clinical data, and to identify and report trends.	13.53%	11
Database Administrators	Database Administrators Administer, test, and implement computer databases, applying knowledge of database management systems. Coordinate changes to computer databases. May plan, coordinate, and implement security measures to safeguard computer databases.	12.54%	4
Database Architects	Database Architects Design strategies for enterprise database systems and set standards for operations, programming, and security. Design and construct large relational databases. Integrate new systems with existing warehouse structure and refine system performance and functionality.	11.54%	4

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.

Simi Bot

Home

Scoring Result

Clustering Result

Help

Scoring Result

Copy CSV Excel PDF

Search:

Cluster Group	Average Similarity Score	Number of Targets in Group	Most Frequent Words in Group
4	6.01%	31	computer, information, security, network, software
13	4.54%	24	financial, credit, loan, securities, individuals
12	2.63%	9	insurance, claims, company, policies, determine
5	2.42%	45	energy, control, production, plant, nuclear
3	2.12%	35	office, mail, managers, coordinate, direct
11	1.77%	80	health, care, medical, patients, clinical
6	1.76%	20	aircraft, air, weapons, flight, operations
10	1.60%	40	engineering, electrical, equipment, test, design
15	1.57%	10	food, workers, serving, preparation, processing
1	1.56%	396	services, provide, patrons, animals, merchandise
9	1.27%	22	teachers, sciences, listed, separately, social