DSI Exercise 03:

Assignment 1: Apply Discretization

Discretize the attribute “Overall rank” into the Top 10, Top 11-50, and Top 51-100 of the happiest countries in the world in 2019. Leave all other attributes as is (you can apply settings to a selection of attributes at once). Look at the results and compare it to the original data set (you can use the “Data Table” widget for that). Then use the “Select Rows” widget to select the Top 10 and look at the re- sults. Make screenshots of the complete workflow, the discretization settings, and the resulting data.

Answer:
The screen shots are given below:

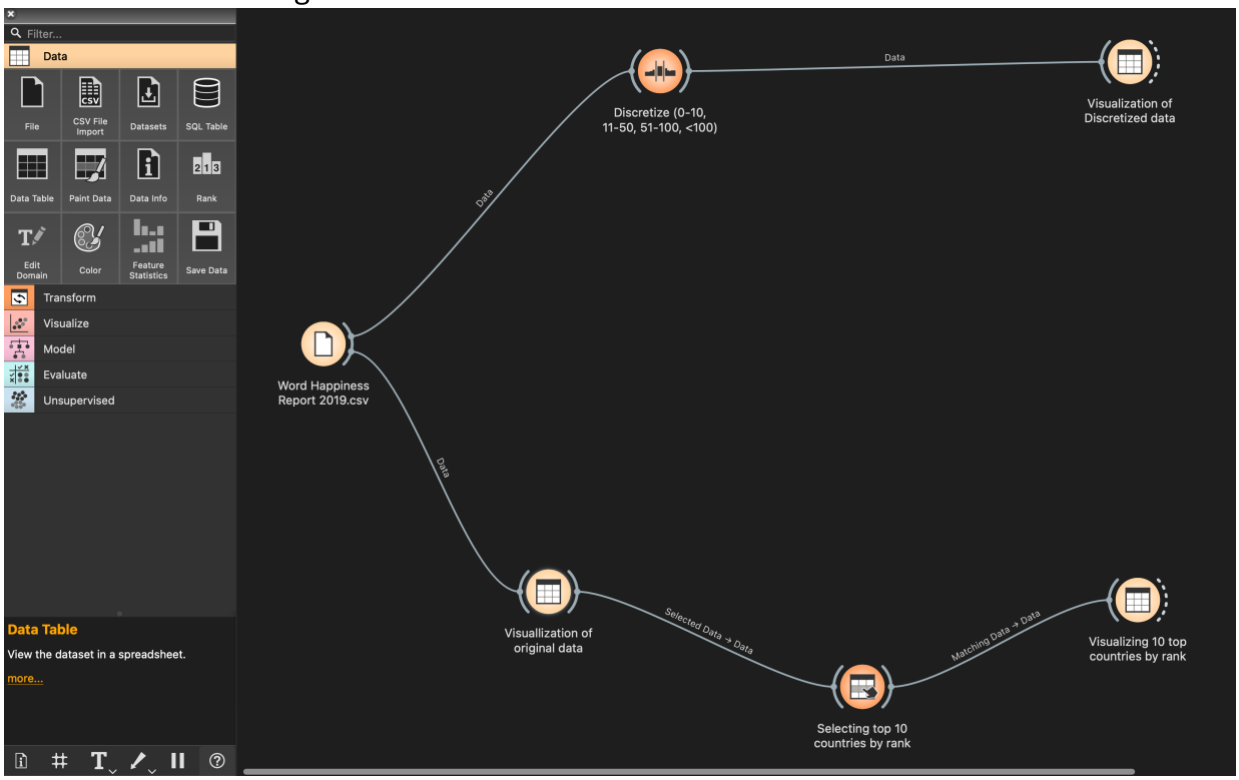


Figure 1Complete workflow of discretizing 'Overall rank' attribute

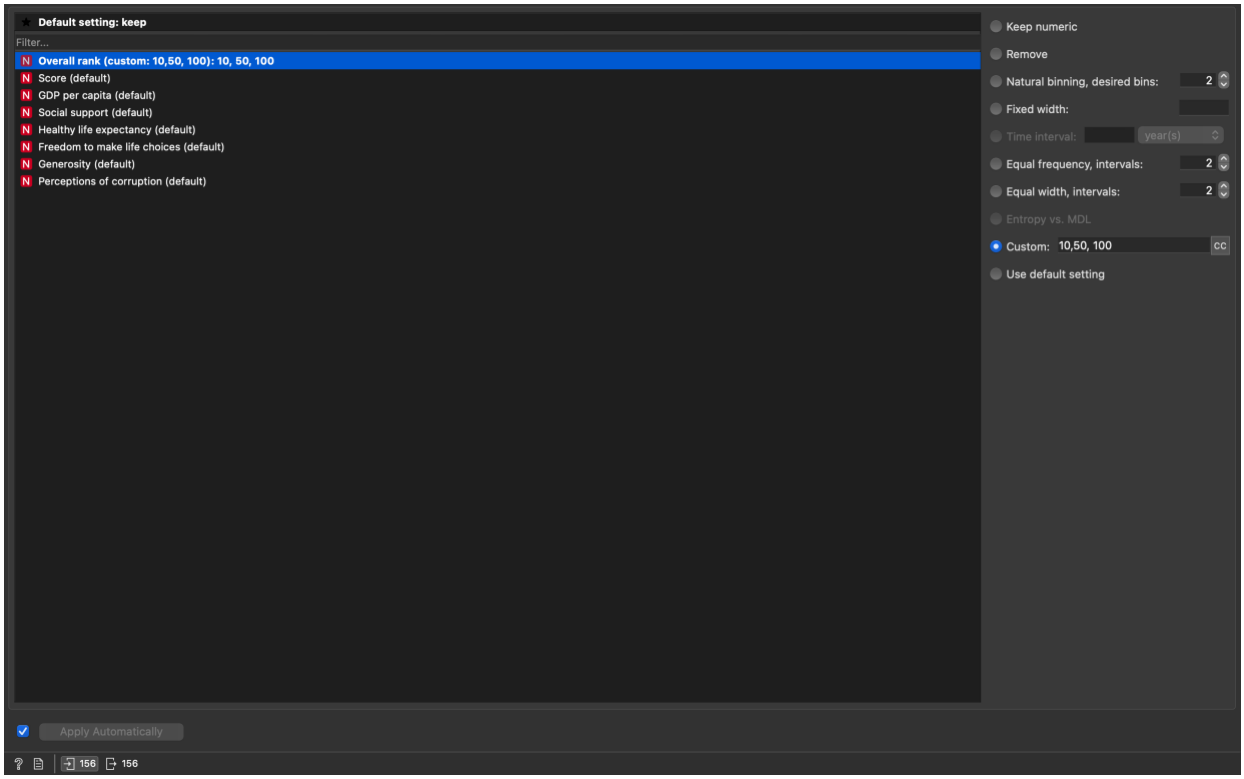
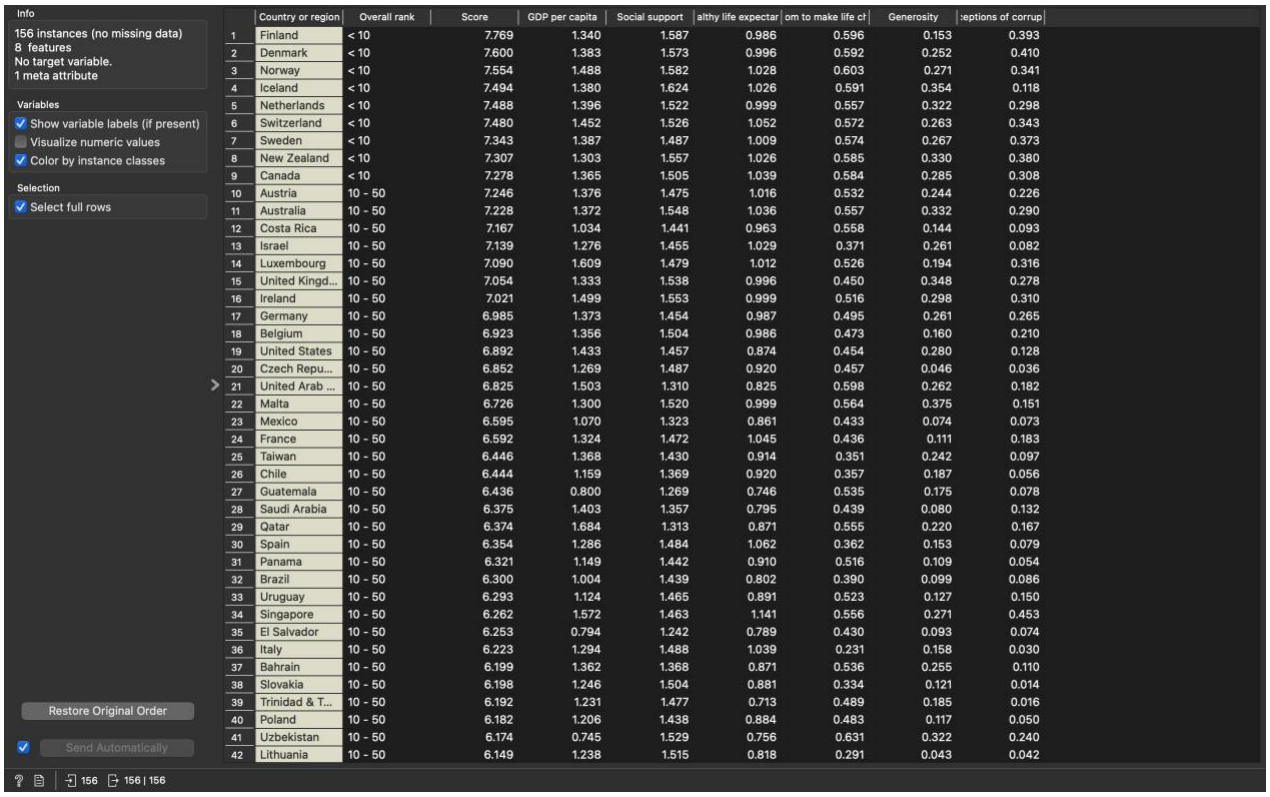
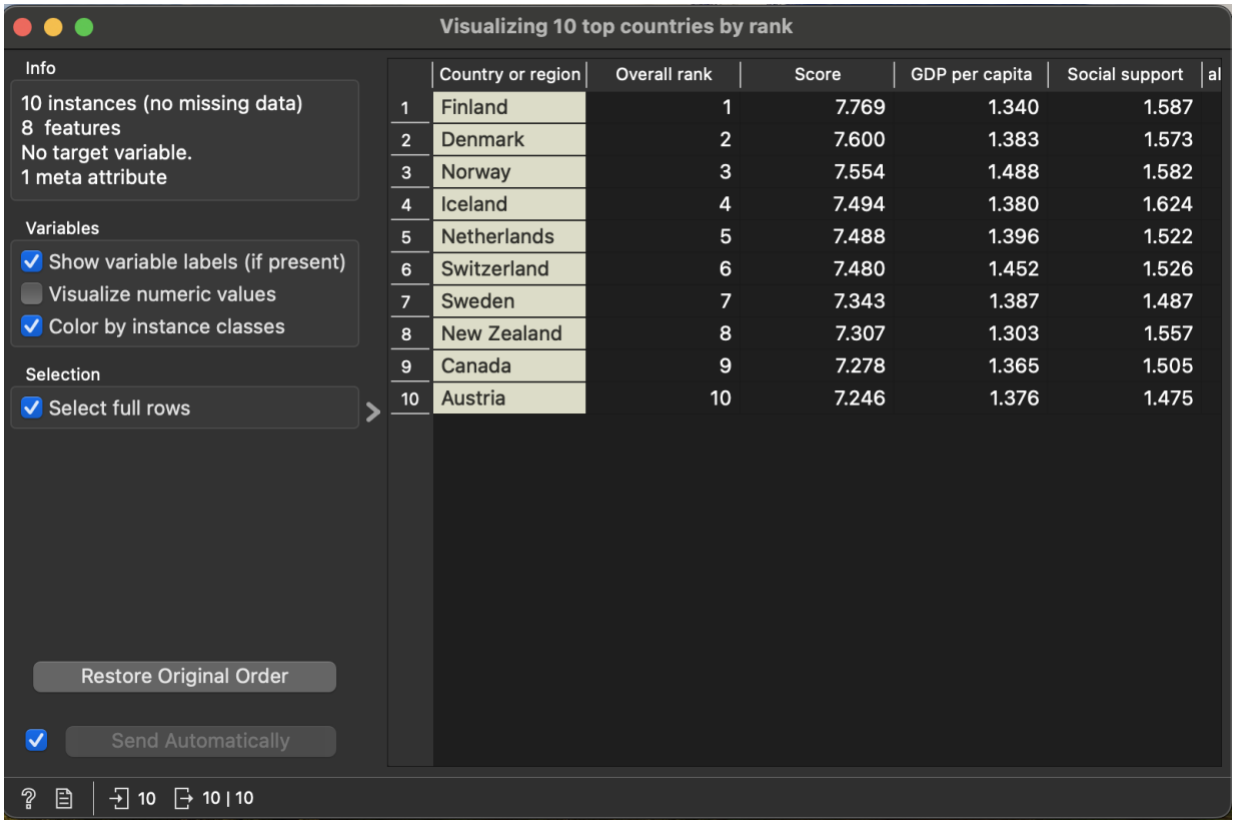


Figure 2 Discretization settings



	Country or region	Overall rank	Score	GDP per capita	Social support	althy life expectar	om to make life ch	Generosity	ceptions of corrup
1	Finland	< 10	7.769	1.340	1.587	0.986	0.596	0.153	0.393
2	Denmark	< 10	7.600	1.383	1.573	0.996	0.592	0.252	0.410
3	Norway	< 10	7.554	1.488	1.582	1.028	0.603	0.271	0.341
4	Iceland	< 10	7.494	1.380	1.624	1.026	0.591	0.354	0.118
5	Netherlands	< 10	7.488	1.396	1.522	0.999	0.557	0.322	0.298
6	Switzerland	< 10	7.480	1.452	1.526	1.052	0.572	0.263	0.343
7	Sweden	< 10	7.343	1.387	1.487	1.009	0.574	0.267	0.373
8	New Zealand	< 10	7.307	1.303	1.557	1.026	0.585	0.330	0.380
9	Canada	< 10	7.278	1.365	1.505	1.039	0.584	0.285	0.308
10	Austria	10 - 50	7.246	1.376	1.475	1.016	0.532	0.244	0.226
11	Australia	10 - 50	7.228	1.372	1.548	1.036	0.557	0.332	0.290
12	Costa Rica	10 - 50	7.167	1.034	1.441	0.963	0.558	0.144	0.093
13	Israel	10 - 50	7.139	1.276	1.455	1.029	0.371	0.261	0.082
14	Luxembourg	10 - 50	7.090	1.609	1.479	1.012	0.526	0.194	0.316
15	United Kingd...	10 - 50	7.054	1.333	1.538	0.996	0.450	0.348	0.278
16	Ireland	10 - 50	7.021	1.499	1.553	0.999	0.516	0.298	0.310
17	Germany	10 - 50	6.985	1.373	1.454	0.987	0.495	0.261	0.265
18	Belgium	10 - 50	6.923	1.356	1.504	0.986	0.473	0.160	0.210
19	United States	10 - 50	6.892	1.433	1.457	0.874	0.454	0.280	0.128
20	Czech Repu...	10 - 50	6.852	1.269	1.487	0.920	0.457	0.046	0.036
21	United Arab ...	10 - 50	6.825	1.503	1.310	0.825	0.598	0.262	0.182
22	Malta	10 - 50	6.726	1.300	1.520	0.999	0.564	0.375	0.151
23	Mexico	10 - 50	6.595	1.070	1.323	0.861	0.433	0.074	0.073
24	France	10 - 50	6.592	1.324	1.472	1.045	0.436	0.111	0.183
25	Taiwan	10 - 50	6.446	1.368	1.430	0.914	0.351	0.242	0.097
26	Chile	10 - 50	6.444	1.159	1.369	0.920	0.357	0.187	0.056
27	Guatemala	10 - 50	6.436	0.800	1.269	0.746	0.535	0.175	0.078
28	Saudi Arabia	10 - 50	6.375	1.403	1.357	0.795	0.439	0.080	0.132
29	Qatar	10 - 50	6.374	1.684	1.313	0.871	0.555	0.220	0.167
30	Spain	10 - 50	6.354	1.286	1.484	1.062	0.362	0.153	0.079
31	Panama	10 - 50	6.321	1.149	1.442	0.910	0.516	0.109	0.054
32	Brazil	10 - 50	6.300	1.004	1.439	0.802	0.390	0.099	0.086
33	Uruguay	10 - 50	6.293	1.124	1.465	0.891	0.523	0.127	0.150
34	Singapore	10 - 50	6.262	1.572	1.463	1.141	0.556	0.271	0.453
35	El Salvador	10 - 50	6.253	0.794	1.242	0.789	0.430	0.093	0.074
36	Italy	10 - 50	6.223	1.294	1.488	1.039	0.231	0.158	0.030
37	Bahrain	10 - 50	6.199	1.362	1.368	0.871	0.536	0.255	0.110
38	Slovakia	10 - 50	6.198	1.246	1.504	0.881	0.334	0.121	0.014
39	Trinidad & T...	10 - 50	6.192	1.231	1.477	0.713	0.489	0.185	0.016
40	Poland	10 - 50	6.182	1.206	1.438	0.884	0.483	0.117	0.050
41	Uzbekistan	10 - 50	6.174	0.745	1.529	0.756	0.631	0.322	0.240
42	Lithuania	10 - 50	6.149	1.238	1.515	0.818	0.291	0.043	0.042

Figure 3 Full resulting data after discretization



	Country or region	Overall rank	Score	GDP per capita	Social support	al
1	Finland	1	7.769	1.340	1.587	
2	Denmark	2	7.600	1.383	1.573	
3	Norway	3	7.554	1.488	1.582	
4	Iceland	4	7.494	1.380	1.624	
5	Netherlands	5	7.488	1.396	1.522	
6	Switzerland	6	7.480	1.452	1.526	
7	Sweden	7	7.343	1.387	1.487	
8	New Zealand	8	7.307	1.303	1.557	
9	Canada	9	7.278	1.365	1.505	
10	Austria	10	7.246	1.376	1.475	

Figure 4 Top 10 countries by 'Overall rank'

Assignment 2: Pre-processing of Text Data

1. First, remove all preprocessing steps except “Transformation” and transform the corpus into lowercase. This avoids to have multiple versions (“Large” and “large”) of the same word. If you checked the box “Apply Automatically” on the lower left, the changes are directly ap- plied. You can use the “Word Cloud” widget to observe the effect of the transformation. Please note that also punctuation symbols are also counted. Save the word cloud clicking on the disk symbol.

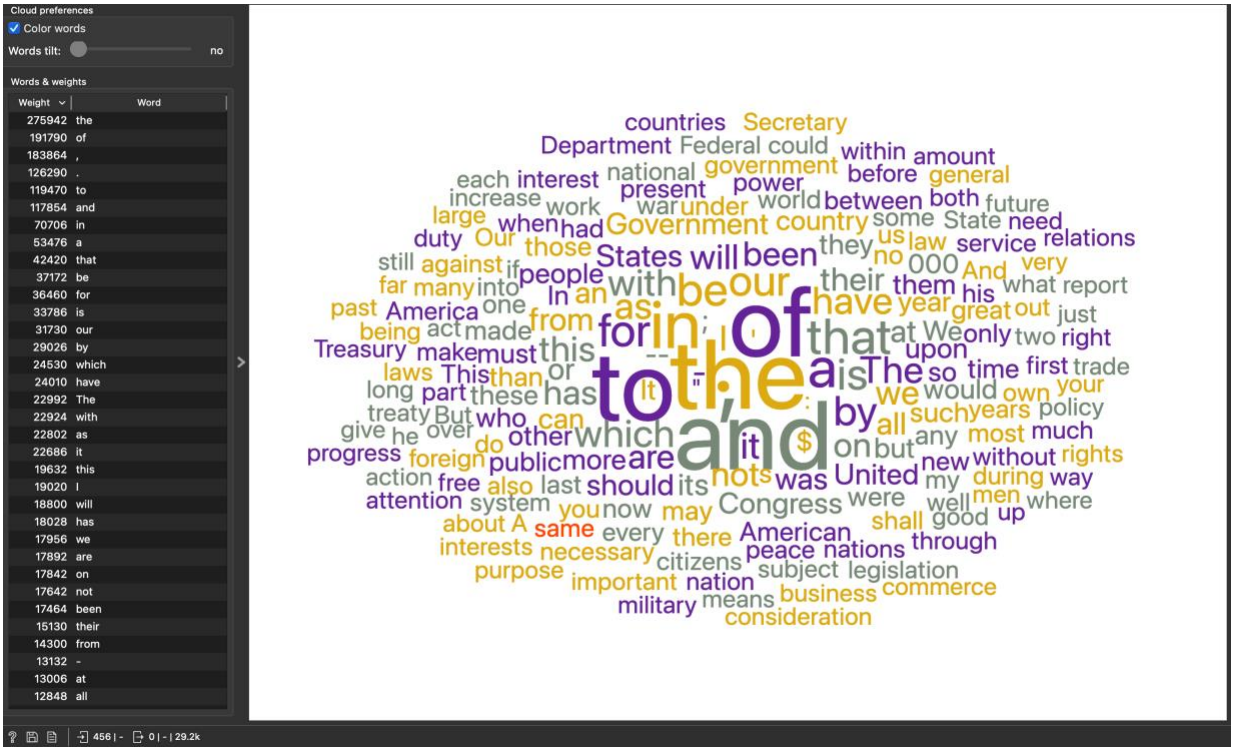


Figure 5 Original Word Cloud

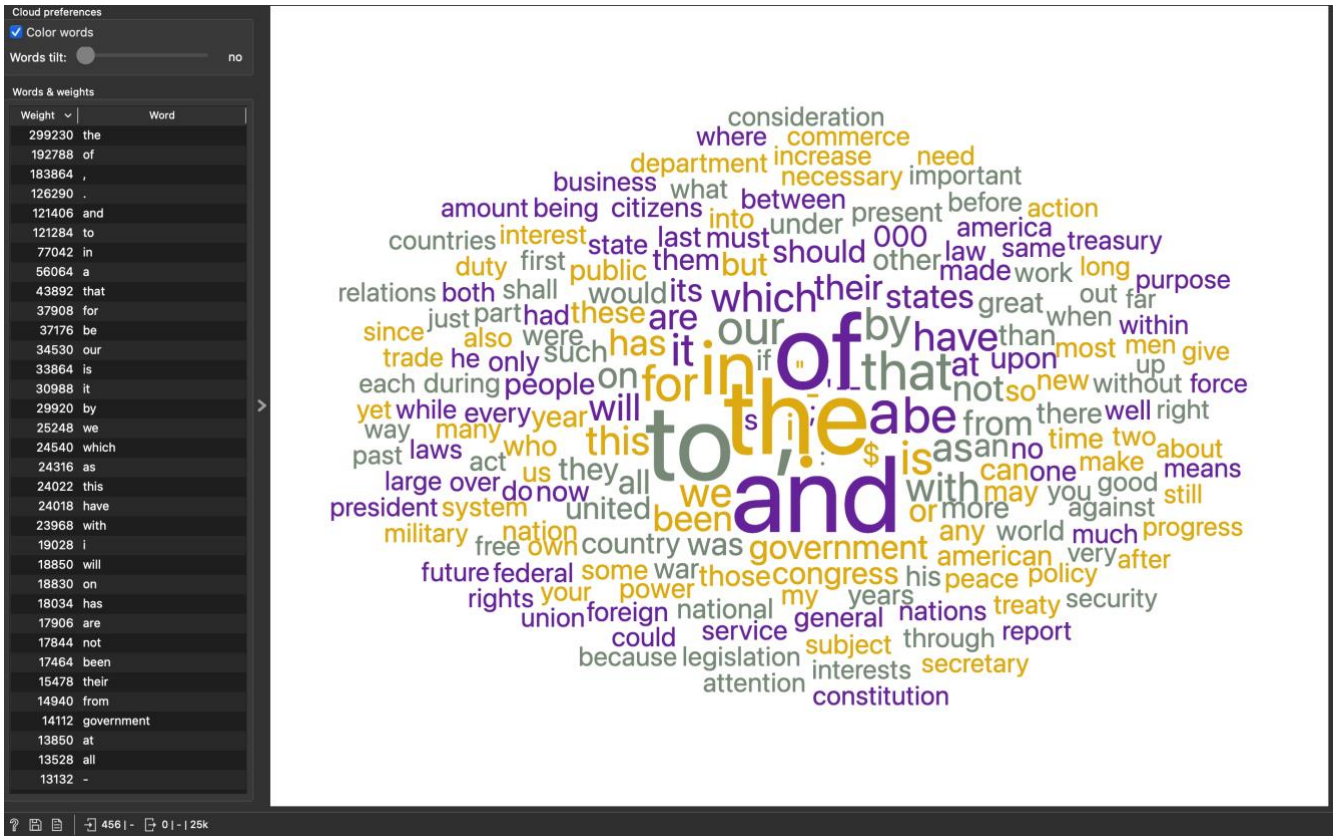


Figure 6 After 'Transformation' to lowercase

2. Now apply “Tokenization”. This is used to break the text into smaller pieces like sentences, words, ... Please have a look at the aforementioned documentation to understand the differ- ent tokenization methods. Use Regular Expressions to split the text by words without keep- ing punctuation. This is quite a common way to break down text. What are the top 3 words? Save the word cloud clicking on the disk symbol.

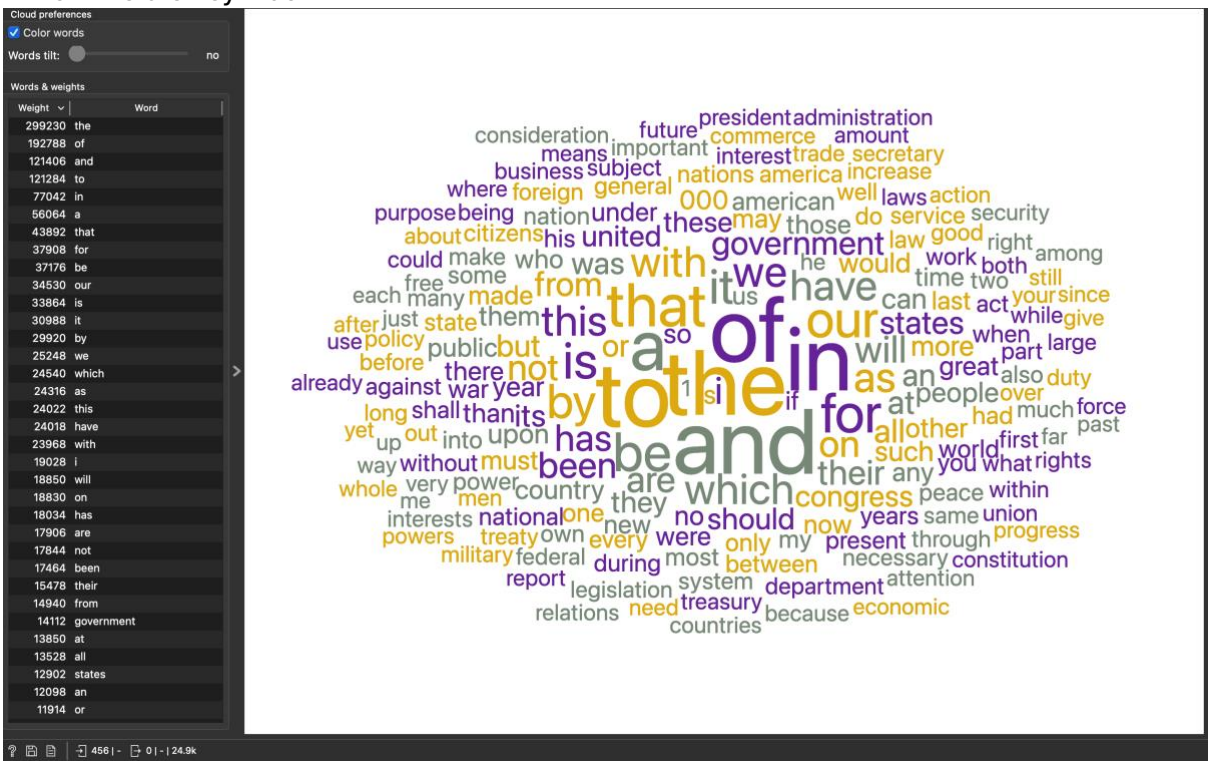


Figure 7 After tokenization

Fromt the image, we can see that ‘the’, ‘of’, and ‘and’ are the top three words.

3. Now filter the stopwords. What are the top 3 words now? Save the word cloud clicking on the disk symbol.

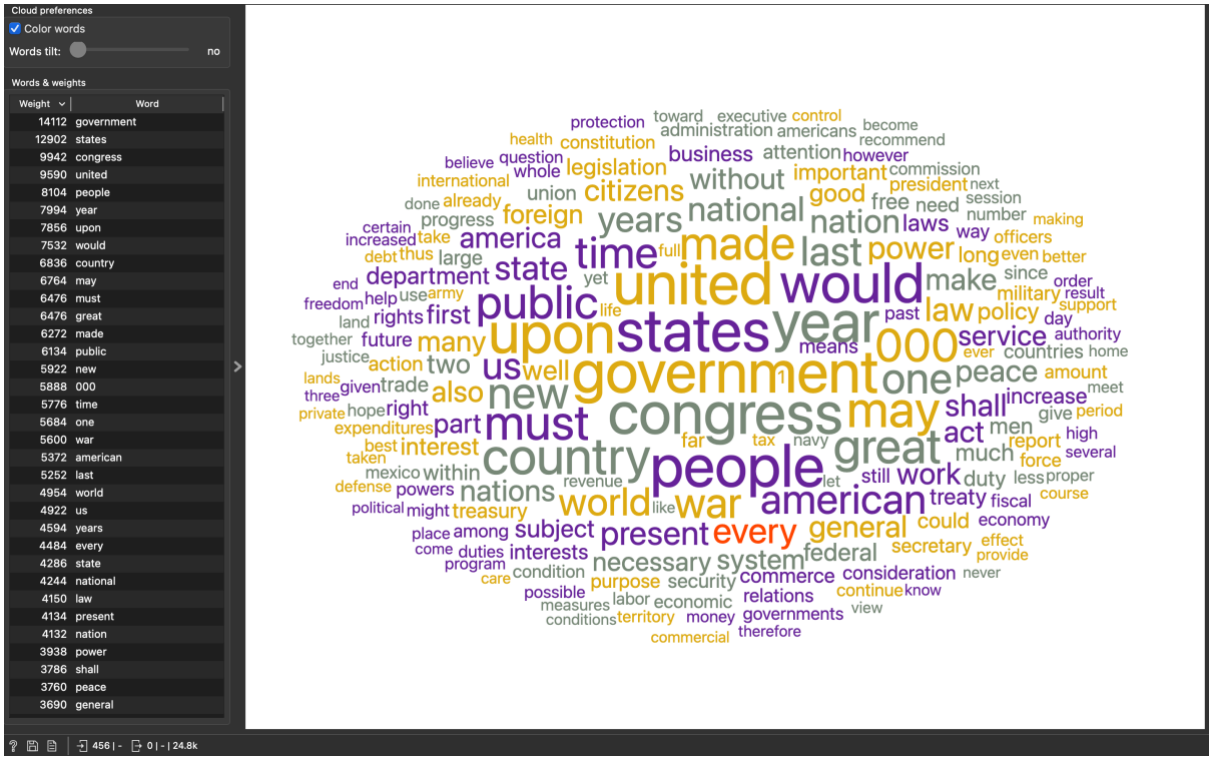


Figure 8 After filtering the stopwords.

After filtering, we can see that the 'government', 'states', and 'congress' are the top three words.

4. Apply the standard normalization (Porter Stemmer). This does, to put it simple, convert words to their base form, like e.g. "the boy's cars are different colors" à "the boy car be differ color". What do you observe?

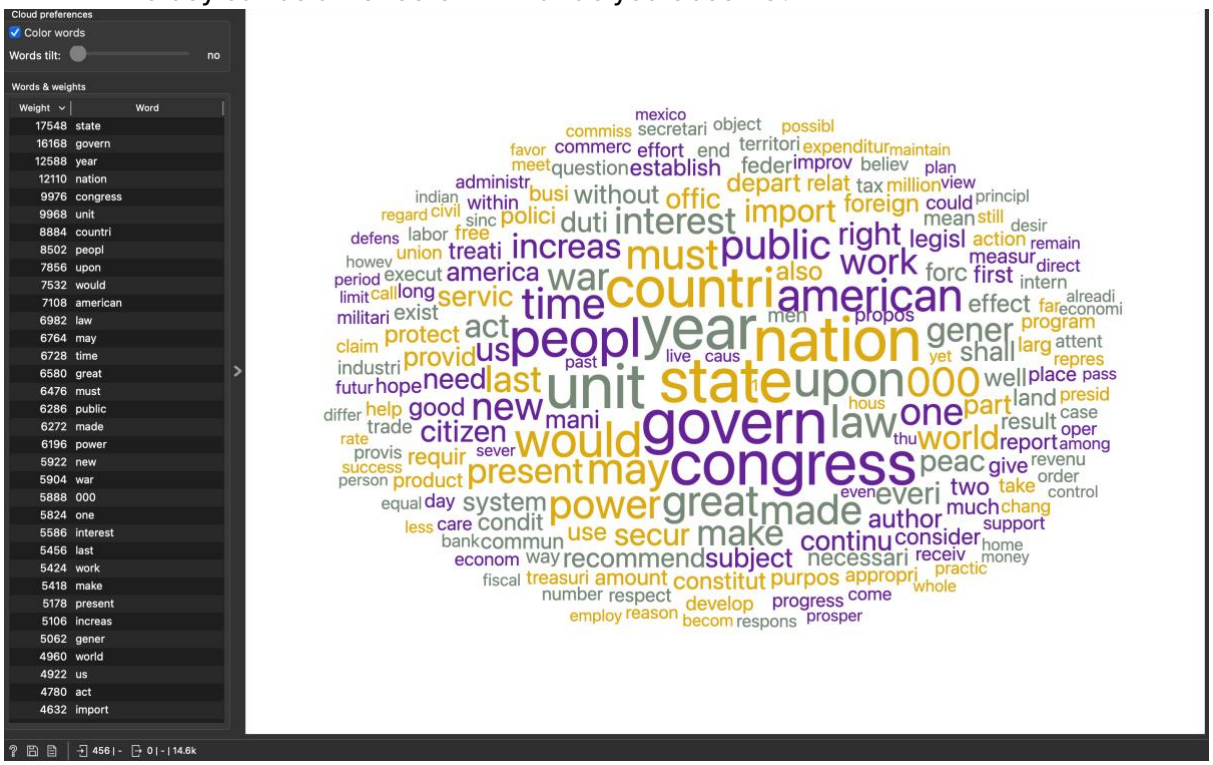


Figure 9 After normalization

After applying standard normalization (Porter Stemmer), we can see that the ‘state’, ‘govern’, and ‘year’ are the top three words. Before normalization, there was a full word ‘government’ in the top position. However, after normalization, it drops down to second position and only base form ‘govern’ is remained.

The complete workflow and preprocessing settings are also shown below.

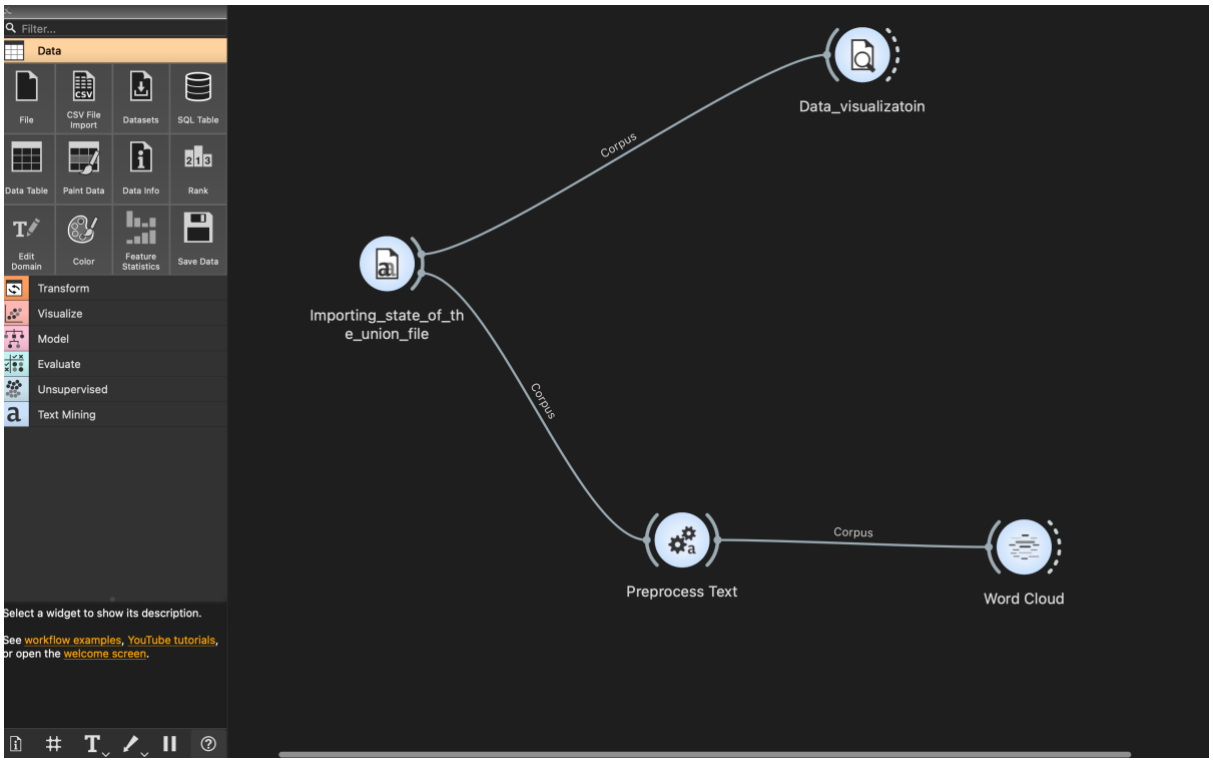


Figure 10 Complete workflow

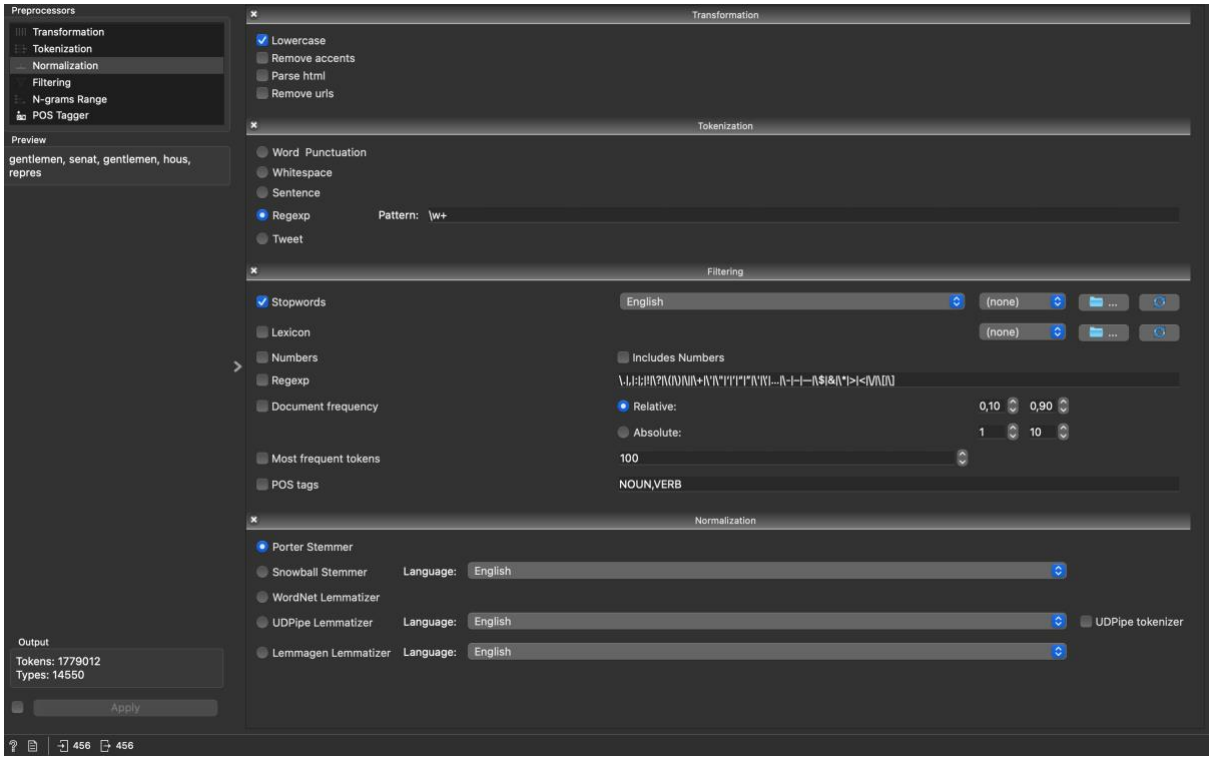


Figure 11 Preprocessing steps and settings