In [4]:

```
import pandas as pd
import numpy as np
import sklearn
import nltk
import scipy
import re
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
```

In [5]:

```
import langdetect as ld
import re
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
from nltk import word_tokenize, FreqDist
import matplotlib.pyplot as plt
from nltk.corpus import stopwords
import random
from sklearn.model_selection import train_test_split
from sklearn import svm
import sklearn.metrics as m
from sklearn.tree import DecisionTreeClassifier, plot_tree
from sklearn.model_selection import GridSearchCV
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.pipeline import Pipeline
import numpy as np
from sklearn.utils import resample
from plotnine import *
```

In [6]:

```
from bertopic import BERTopic
```

In [7]:

```
biometric = pd.read_excel(r"C:\Users\sahil\OneDrive\Desktop\pythontry\biometric.xlsx")        #loading data
```

In [8]:

```
biometric
```

Out[8]:

	Publication Number	Title (English)	Abstract	First Claim	Assignee/Applicant	Inventor	Priority Date - Earliest	Publication Date	INPADOC Family Members	IPC Current Full	...
0	AU2022203027A1	Implementation of biometric authentication	The present disclosure relates generally to im...	1. A method, comprising: at an electronic devi...	Apple Inc.	Os, Marcel Van | Abbasian, Reza | Anton, Peter...	2022-05-05	2022-05-26	NaN	G06F002132 | G06K000900	...
1	KR2022061930A	encryption module using finger scan and contro...	The present invention relates to an encryption...	1. An encryption module provided in a device c...	Park Jin,KR | SIM GYOOCHAN,KR | 박진,KR | 심규찬,KR	Park Jin | SIM, GYOOCHAN | 박진 | 심규찬	2017-07-03	2022-05-13	KR2022061930A | KR2019004019A	G06F002132 | G05B002302 | G06F002145 | G06F002...	...
2	CN114511427A	NaN	本申请公开了一种项目工地的安全教育监管方法、装置及系统，通过服务器获取安全教育终端发送的身份...	1.一种项目工地的安全教育监管方法，其特征在于，应用于服务器，所述方法包括：\n获取安全教育...	四川省大数据中心,成都市,四川省,610000,CN	顾红松 | 赵启斌 | 唐为之 | 王思航 | 覃子凌	2022-04-21	2022-05-17	CN114511427A	G06Q005020 | G06F00162455 | G06F002132 | G06F0...	...
3	CN114510703A	NaN	本发明涉及物联网技术领域，具体的公开了一种物联网设备的身份标识认证方法及通信系统，所述的物联...	1.一种物联网设备的身份标识通信系统，其特征在于，所述的物联网设备的身份标识通信系统包括语音...	尚班(北京)网络科技有限公司,北京市,北京市,100000,CN	李光顺	2022-04-20	2022-05-17	CN114510703A	G06F002144 | G06F002132 | G06F002145 | G16Y004050	...
4	CN114511917A	NaN	本发明实施例提供了一种图像数据的处理方法及装置、电子设备、存储介质，方法包括：获取针对目标人...	1.一种图像数据的处理方法，其特征在于，所述方法包括：\n获取针对目标人脸所采集的目标图像数...	北京美摄网络科技有限公司,北京市,北京市,100195,CN	张瑞全 | 尚盼龙 | 刘铁华 | 李振	2022-04-20	2022-05-17	CN114511917A	G06V004016 | G06F002132	...
...	...	...	...	...	...	...	...	...	...	...	...
28550	NL2011998C	NaN	NaN	NaN	NaN	NaN	2013-12-20	2015-06-26	NaN	NaN	...
28551	NO201300289A	NaN	NaN	NaN	NaN	NaN	2013-02-22	2014-08-25	NaN	NaN	...
28552	NO201300886A	NaN	NaN	NaN	NaN	NaN	2013-06-26	2014-06-29	NaN	NaN	...
28553	NO201301423A	NaN	NaN	NaN	NaN	NaN	2013-02-22	2014-08-25	NaN	NaN	...
28554	NL2006733C	NaN	NaN	NaN	NaN	NaN	2011-05-06	2012-11-08	NaN	NaN	...

28555 rows × 30 columns

In [9]:

```
biometric.describe()
```

Out[9]:

	Count of Cited Refs - Patent	Count of Cited Refs - Non-patent	Count of Citing Patents	Strategic Importance	Probability of Grant	Domain Influence	Combined Patent Impact	Probability of Restoration Post-Lapse	Probability of Early Lapse
count	28555.000000	28555.000000	28555.000000	28221.000000	22246.000000	28221.000000	28221.000000	69.000000	9222.000000
mean	6.038137	0.942042	1.404448	1.509830	66.812376	3.623173	3.223587	18.293333	27.942469
std	27.700650	10.508773	5.984700	2.156655	35.035114	6.251872	5.185330	13.533425	24.663635
min	0.000000	0.000000	0.000000	1.000000	0.000000	1.000000	1.000000	0.120000	0.000000
25%	0.000000	0.000000	0.000000	1.000000	36.450000	1.890000	1.710000	7.610000	7.362500
50%	3.000000	0.000000	0.000000	1.000000	70.440000	2.330000	2.070000	16.600000	14.500000
75%	6.000000	0.000000	1.000000	1.000000	100.000000	3.660000	3.140000	26.500000	51.600000
max	1353.000000	802.000000	232.000000	88.640000	100.000000	100.000000	95.710000	64.700000	89.800000

In [10]:

```
biometric.shape
```

Out[10]:

```
(28555, 30)
```

In [11]:

```
biometric.columns
```

Out[11]:

```
Index(['Publication Number', 'Title (English)', 'Abstract', 'First Claim',
       'Assignee/Applicant', 'Inventor', 'Priority Date - Earliest',
       'Publication Date', 'INPADOC Family Members', 'IPC Current Full',
       'Independent Claims', 'INPADOC Legal Status', 'Dead/Alive',
       'CPC - Current', 'Claims Count', 'Count of Cited Refs - Patent',
       'Count of Cited Refs - Non-patent', 'Count of Citing Patents',
       'Priority Country/Region', 'Application Date', 'Title - DWPI',
       'Abstract (English)', 'Maintenance Status (US)',
       'INPADOC Family Members Dead/Alive', 'Strategic Importance',
       'Probability of Grant', 'Domain Influence', 'Combined Patent Impact',
       'Probability of Restoration Post-Lapse', 'Probability of Early Lapse'],
      dtype='object')
```

In [12]:

```
biometric.isnull().sum()
```

Out[12]:

```
Publication Number                         0
Title (English)                          848
Abstract                                2096
First Claim                             1878
Assignee/Applicant                       724
Inventor                                1453
Priority Date - Earliest                   0
Publication Date                           0
INPADOC Family Members                  3192
IPC Current Full                        1129
Independent Claims                      2158
INPADOC Legal Status                    4854
Dead/Alive                                 0
CPC - Current                           8512
Claims Count                            1962
Count of Cited Refs - Patent               0
Count of Cited Refs - Non-patent           0
Count of Citing Patents                    0
Priority Country/Region                  379
Application Date                           0
Title - DWPI                            1251
Abstract (English)                      2409
Maintenance Status (US)                28323
INPADOC Family Members Dead/Alive       3192
Strategic Importance                     334
Probability of Grant                    6309
Domain Influence                         334
Combined Patent Impact                   334
Probability of Restoration Post-Lapse  28486
Probability of Early Lapse             19333
dtype: int64
```

In [19]:

```
# Removing Nan / nan values from Dataset
docs = list(biometric.loc[:, "Title (English)"].values)
biometric["Title (English)"].fillna('NO DATA', inplace=True)
```

In [20]:

```
biometric
```

Out[20]:

	Publication Number	Title (English)	Abstract	First Claim	Assignee/Applicant	Inventor	Priority Date - Earliest	Publication Date	INPADOC Family Members	IPC Current Full	...
0	AU2022203027A1	Implementation of biometric authentication	The present disclosure relates generally to im...	1. A method, comprising: at an electronic devi...	Apple Inc.	Os, Marcel Van | Abbasian, Reza | Anton, Peter...	2022-05-05	2022-05-26	NaN	G06F002132 | G06K000900	...
1	KR2022061930A	encryption module using finger scan and contro...	The present invention relates to an encryption...	1. An encryption module provided in a device c...	Park Jin,KR | SIM GYOOCHAN,KR | 박진,KR | 심규찬,KR	Park Jin | SIM, GYOOCHAN | 박진 | 심규찬	2017-07-03	2022-05-13	KR2022061930A | KR2019004019A	G06F002132 | G05B002302 | G06F002145 | G06F002...	...
2	CN114511427A	NO DATA	本申请公开了一种项目工地的安全教育监管方法、装置及系统，通过服务器获取安全教育终端发送的身份...	1.一种项目工地的安全教育监管方法，其特征在于，应用于服务器，所述方法包括：\n获取安全教育...	四川省大数据中心,成都市,四川省,610000,CN	顾红松 | 赵启斌 | 唐为之 | 王思航 | 覃子凌	2022-04-21	2022-05-17	CN114511427A	G06Q005020 | G06F00162455 | G06F002132 | G06F0...	...
3	CN114510703A	NO DATA	本发明涉及物联网技术领域，具体的公开了一种物联网设备的身份标识认证方法及通信系统，所述的物联...	1.一种物联网设备的身份标识通信系统，其特征在于，所述的物联网设备的身份标识通信系统包括语音...	尚班(北京)网络科技有限公司,北京市,北京市,100000,CN	李光顺	2022-04-20	2022-05-17	CN114510703A	G06F002144 | G06F002132 | G06F002145 | G16Y004050	...
4	CN114511917A	NO DATA	本发明实施例提供了一种图像数据的处理方法及装置、电子设备、存储介质，方法包括：获取针对目标人...	1.一种图像数据的处理方法，其特征在于，所述方法包括：\n获取针对目标人脸所采集的目标图像数...	北京美摄网络科技有限公司,北京市,北京市,100195,CN	张瑞全 | 尚盼龙 | 刘铁华 | 李振	2022-04-20	2022-05-17	CN114511917A	G06V004016 | G06F002132	...
...	...	...	...	...	...	...	...	...	...	...	...
28550	NL2011998C	NO DATA	NaN	NaN	NaN	NaN	2013-12-20	2015-06-26	NaN	NaN	...
28551	NO201300289A	NO DATA	NaN	NaN	NaN	NaN	2013-02-22	2014-08-25	NaN	NaN	...
28552	NO201300886A	NO DATA	NaN	NaN	NaN	NaN	2013-06-26	2014-06-29	NaN	NaN	...
28553	NO201301423A	NO DATA	NaN	NaN	NaN	NaN	2013-02-22	2014-08-25	NaN	NaN	...
28554	NL2006733C	NO DATA	NaN	NaN	NaN	NaN	2011-05-06	2012-11-08	NaN	NaN	...

28555 rows × 31 columns

In [21]:

```
import re
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

stop_words = stopwords.words('english')

def clean_text(x):
  x = str(x)
  x = x.lower()
  x = re.sub(r'#[A-Za-z0-9]*', ' ', x)
  x = re.sub(r'https*://.*', ' ', x)
  x = re.sub(r'@[A-Za-z0-9]+', ' ', x)
  tokens = word_tokenize(x)
  x = ' '.join([w for w in tokens if not w.lower() in stop_words])
  x = re.sub(r'[%s]' % re.escape('!"#$%&\()*+,-./:;<=>?@[\\]^_`{|}~“…”’'), ' ', x)
  x = re.sub(r'\d+', ' ', x)
  x = re.sub(r'\n+', ' ', x)
  x = re.sub(r'\s{2,}', ' ', x)
  return x


biometric['clean_text'] = biometric['First Claim'].apply(clean_text)
biometric.head()
```

Out[21]:

	Publication Number	Title (English)	Abstract	First Claim	Assignee/Applicant	Inventor	Priority Date - Earliest	Publication Date	INPADOC Family Members	IPC Current Full	...	(
0	AU2022203027A1	Implementation of biometric authentication	The present disclosure relates generally to im...	1. A method, comprising: at an electronic devi...	Apple Inc.	Os, Marcel Van | Abbasian, Reza | Anton, Peter...	2022-05-05	2022-05-26	NaN	G06F002132 | G06K000900	...	The d ge
1	KR2022061930A	encryption module using finger scan and contro...	The present invention relates to an encryption...	1. An encryption module provided in a device c...	Park Jin,KR | SIM GYOOCHAN,KR | 박진,KR | 심규찬,KR	Park Jin | SIM, GYOOCHAN | 박진 | 심규찬	2017-07-03	2022-05-13	KR2022061930A | KR2019004019A	G06F002132 | G05B002302 | G06F002145 | G06F002...	...	The enc
2	CN114511427A	NO DATA	本申请公开了一种项目工地的安全教育监管方法、装置及系统，通过服务器获取安全教育终端发送的身份...	1.一种项目工地的安全教育监管方法，其特征在于，应用于服务器，所述方法包括：\n获取安全教育...	四川省大数据中心,成都市,四川省,610000,CN	顾红松 | 赵启斌 | 唐为之 | 王思航 | 覃子凌	2022-04-21	2022-05-17	CN114511427A	G06Q005020 | G06F00162455 | G06F002132 | G06F0...	...	
3	CN114510703A	NO DATA	本发明涉及物联网技术领域，具体的公开了一种物联网设备的身份标识认证方法及通信系统，所述的物联...	1.一种物联网设备的身份标识通信系统，其特征在于，所述的物联网设备的身份标识通信系统包括语音...	尚班(北京)网络科技有限公司,北京市,北京市,100000,CN	李光顺	2022-04-20	2022-05-17	CN114510703A	G06F002144 | G06F002132 | G06F002145 | G16Y004050	...	
4	CN114511917A	NO DATA	本发明实施例提供了一种图像数据的处理方法及装置、电子设备、存储介质，方法包括：获取针对目标人...	1.一种图像数据的处理方法，其特征在于，所述方法包括：\n获取针对目标人脸所采集的目标图像数...	北京美摄网络科技有限公司,北京市,北京市,100195,CN	张瑞全 | 尚盼龙 | 刘铁华 | 李振	2022-04-20	2022-05-17	CN114511917A	G06V004016 | G06F002132	...	

5 rows × 31 columns

In [22]:

```
combi = biometric.append(biometric, ignore_index=True)
combi.shape
```

Out[22]:

```
(57110, 31)
```

In [23]:

```
# Creating Topics
model = BERTopic(language="english")
topics, probs = model.fit_transform(docs)
```

In [24]:

```
# Count Of topics
freq = model.get_topic_info()
print("Number of topics: {}".format( len(freq)))
freq.head()
```

```
Number of topics: 487
```

Out[24]:

	Topic	Count	Name
0	-1	8696	-1_face_access_management_biometric
1	0	855	0_no_data_verified_omitted
2	1	370	1_fingerprint_terminal_mobile_identification
3	2	334	2_iris_recognition_irisbased_recognizing
4	3	212	3_biological_characteristic_feature_biological...

Above table has 3 main columns which includes information about all the 465 topics in descending order of topics Count.

In [25]:

```
model.get_topic_freq().head()
```

Out[25]:

	Topic	Count
0	-1	8696
1	0	855
2	1	370
3	2	334
4	3	212

In [26]:

```
model.get_topic(49)[:10]
```

Out[26]:

```
[('fingerprint', 0.019033753810566167),
 ('electronic', 0.016654933596296086),
 ('recognizing', 0.015662043491238262),
 ('circuit', 0.012458468117969357),
 ('detecting', 0.01229620312601419),
 ('assembly', 0.010746108227533024),
 ('drag', 0.00974331199192471),
 ('documentor', 0.00974331199192471),
 ('originals', 0.00974331199192471),
 ('identification', 0.009419504742697926)]
```

In [27]:

```
model.get_topic_info()
```

Out[27]:

	Topic	Count	Name
0	-1	8696	-1_face_access_management_biometric
1	0	855	0_no_data_verified_omitted
2	1	370	1_fingerprint_terminal_mobile_identification
3	2	334	2_iris_recognition_irisbased_recognizing
4	3	212	3_biological_characteristic_feature_biological...
...	...	...	...
482	481	10	481_financial_calculation_machine_service
483	482	10	482_biological_measures_granting_early
484	483	10	483_users_authenticating_airblow_regarding
485	484	10	484_metal_sleeve_protective_recognizer
486	485	10	485_module_display_indisplay_substrate

487 rows × 3 columns

In [28]:

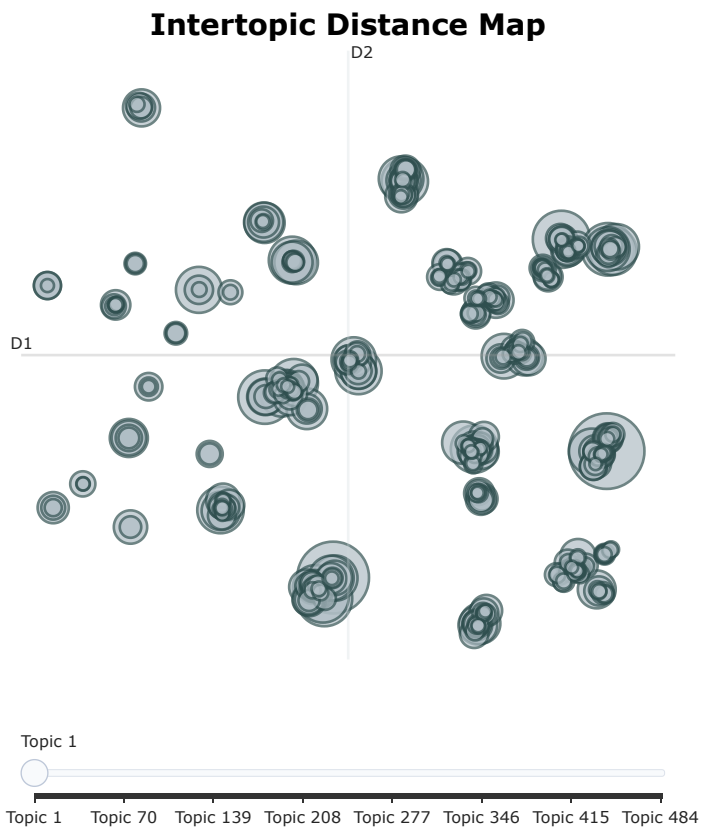
```
# Words for the first topic
topic_one = freq.iloc[0]["Topic"]
model.get_topic(topic_one) # Show the words and their c-TF-IDF scores
```

Out[28]:

```
[('face', 0.002325021552177429),
 ('access', 0.0023089890438906563),
 ('management', 0.002236970695283228),
 ('biometric', 0.0022292622138849845),
 ('to', 0.0022175611178084326),
 ('for', 0.0022075596093947473),
 ('system', 0.0021319206624910336),
 ('on', 0.0021270482620007985),
 ('based', 0.0021207124151201235),
 ('of', 0.002090661920155392)]
```

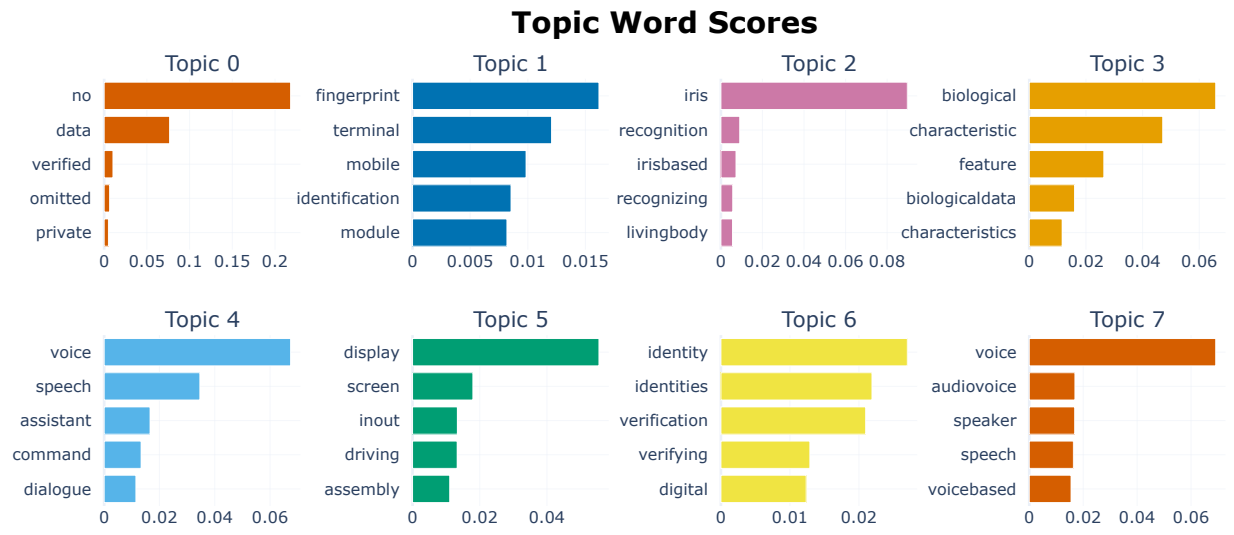
In [29]:

```
model.visualize_topics()
```



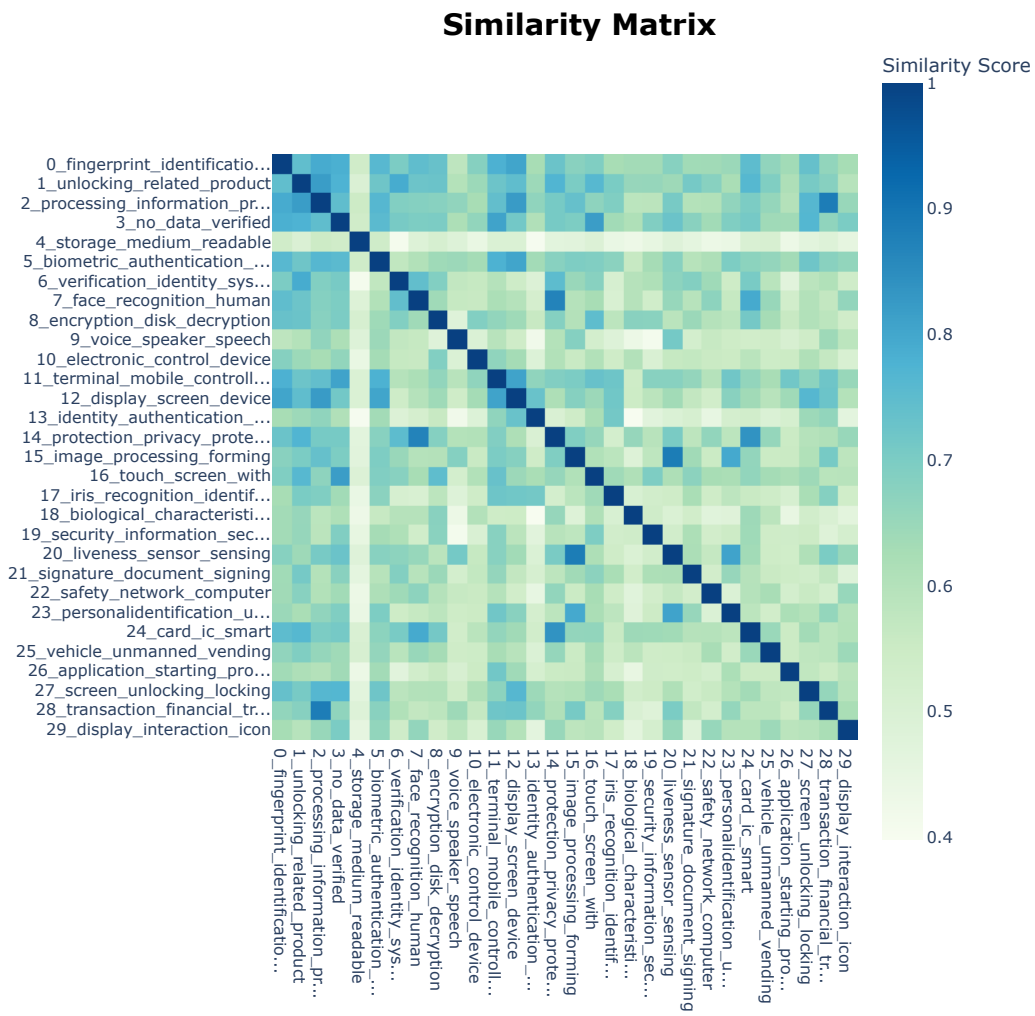
In [30]:

```
# Topics can be visualized in the form of barchart out of the c-TF-IDF score
# (TF-IDF (term frequency-inverse document frequency) is a statistical measure that
#  evaluates how relevant a word is to a document in a collection of documents)
```



In [33]:

```
model.visualize_heatmap()
```



In [34]:

```
# REDUCING THE TOPICS => by passing argument nr_topics with a number of topics you want
new_topics, new_probs = model.reduce_topics(docs, topics,
probs, nr_topics=30)
```

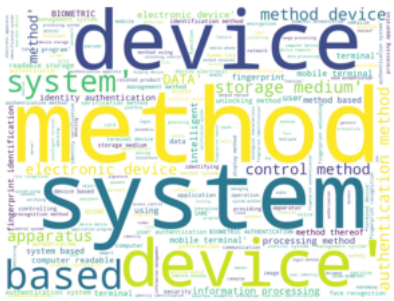
In [35]:

```
docs
```

```
 'Human-computer interaction method, device, device and storage medium',
 'APP fingerprint login method based on artificial intelligence and related device',
 'E-AUTHENTICATION SYSTEM USING EYE BLINKING AND DEEP LEARNING.',
 'A LOCKING UNIT AND A METHOD TO OPERATE A DEVICE',
 'Identity authentication method, device, electronic device and medium',
 'RECOMMENDED METHOD, DEVICE AND EQUIPMENT FOR LOGGING IN AFTER UNINSTALLING AND REINSTALLING AN APPLICATION',
 'AUTHENTICATOR APP FOR CONSENT ARCHITECTURE',
 'User authority management method and system',
 'AI screening identity recognition',
 'SMART WEARABLE DEVICE',
 'The control method of a server apparatus; a system; a server apparatus; a terminal, and a computer program',
 'BIOMETRIC AUTHENTICATION ON FORENSIC MEDICINE AND LEGAL PROCESS',
 'A multi-virus engine self-help checking and killing system and method',
 'NO DATA',
 'Data safety protection system based on internet',
 'SECURED FACE COVERING DEVICE',
 'NO DATA',
 'VERIFICATION METHOD AND VERIFICATION SYSTEM BASED ON BIOLOGICAL CHARACTERISTIC ID CHAIN, AND USER TERMINAL',
 'Computer operation auxiliary control device and system',
 'NO DATA',
```

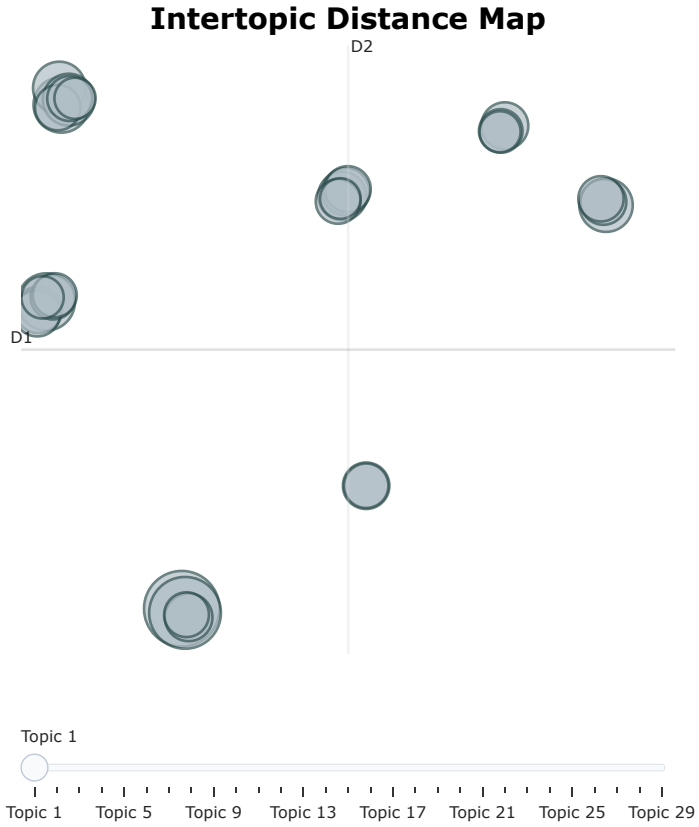
In [36]:

```
# Create stopword list :


stop_words = set(STOPWORDS)
stop_words.update(['ourselves', 'hers', 'between', 'yourself', 'but', 'again', 'there', 'about', 'once', 'during', 'out',
                   'very', 'having', 'with', 'they', 'own', 'an', 'be', 'some', 'for', 'do', 'its', 'yours', 'such', 'into',
                   'of', 'most', 'itself', 'other', 'off', 'is', 's', 'am', 'or', 'who', 'as', 'from', 'him', 'each', 'the',
                   'themselves', 'until', 'below', 'are', 'we', 'these', 'your', 'his', 'through', 'don', 'nor', 'me', 'were',
                   'her', 'more', 'himself', 'this', 'down', 'should', 'our', 'their', 'while', 'above', 'both', 'up', 'to',
                   'ours', 'had', 'she', 'all', 'no', 'when', 'at', 'any', 'before', 'them', 'same', 'and', 'been', 'have',
                   'in','will', 'on', 'does', 'yourselves', 'then', 'that', 'because', 'what', 'over', 'why', 'so',
                   'can', 'did', 'not','now', 'under', 'he', 'you', 'herself', 'has', 'just', 'where', 'too', 'only',
                   'myself', 'which', 'those', 'i','after', 'few', 'whom', 't', 'being', 'if', 'theirs', 'my',
                   'against', 'a', 'by', 'doing', 'it', 'how', 'further', 'was', 'here', 'than'])

# Generate a word cloud image
wordcloud = WordCloud(stopwords=stop_words, background_color="white", width=2000, height=1500).generate(str(docs))

# Display the generated image:
# the matplotlib way:
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```



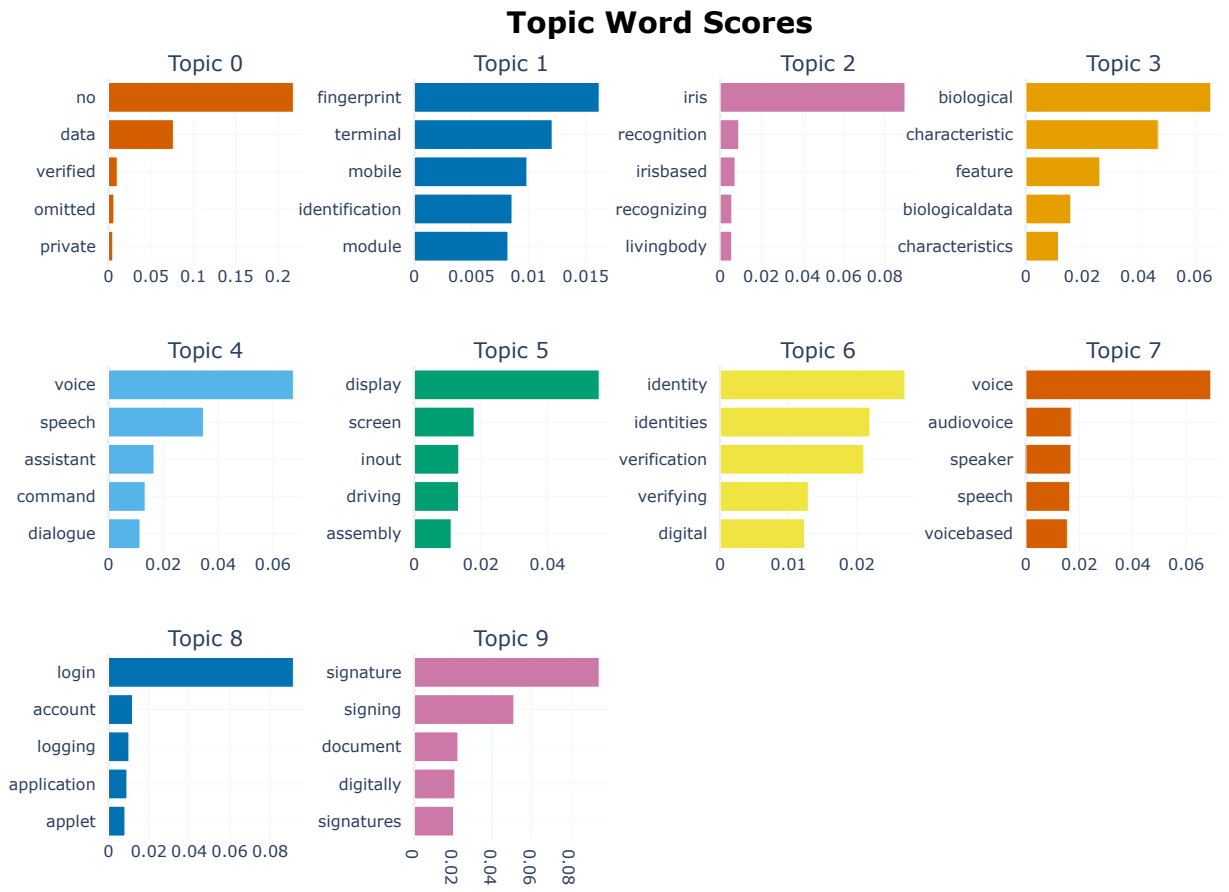
In [37]:

```
model.visualize_topics(top_n_topics=30)
```



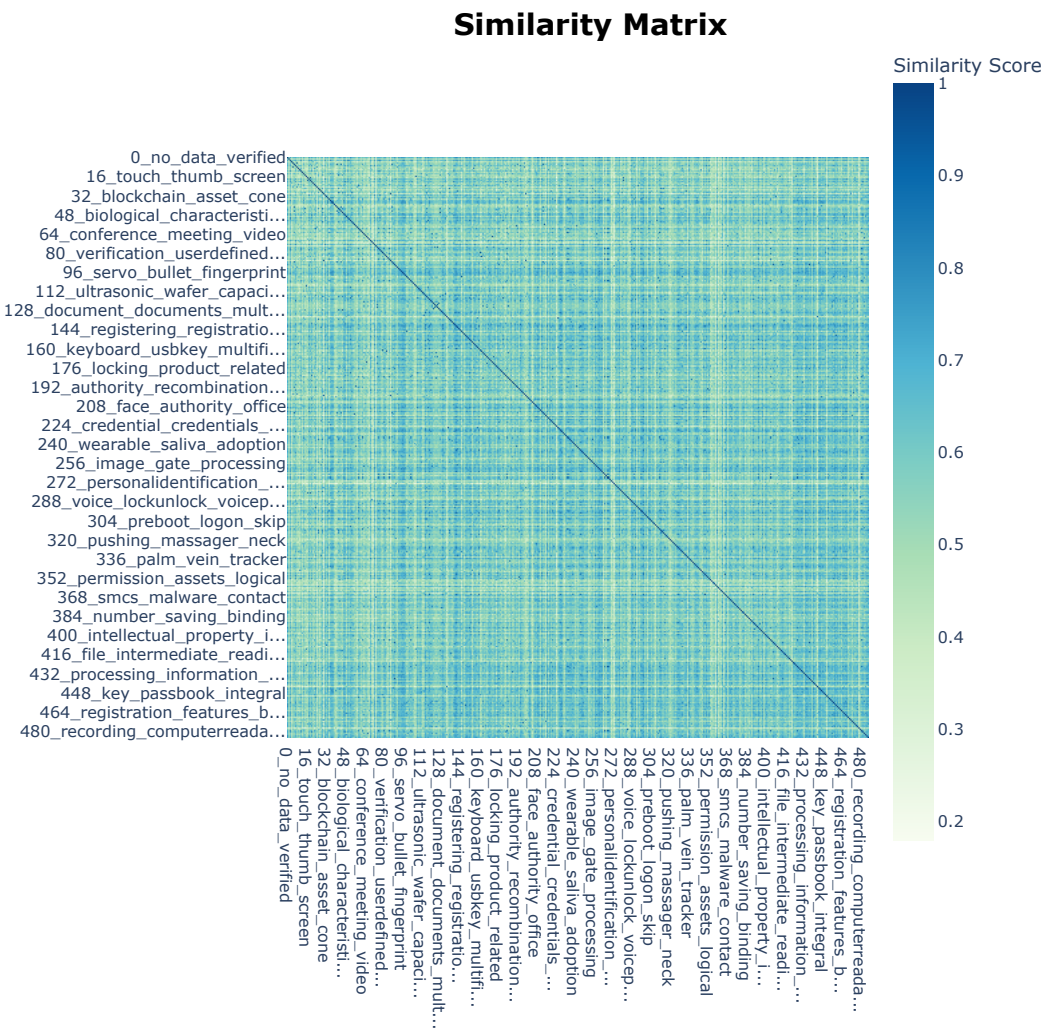
In [38]:

```
# Topics can be visualized in the form of barchart out of the c-TF-IDF score
# (TF-IDF (term frequency-inverse document frequency) is a statistical measure that
#  evaluates how relevant a word is to a document in a collection of documents)
model.visualize_barchart(top_n_topics=10)
```



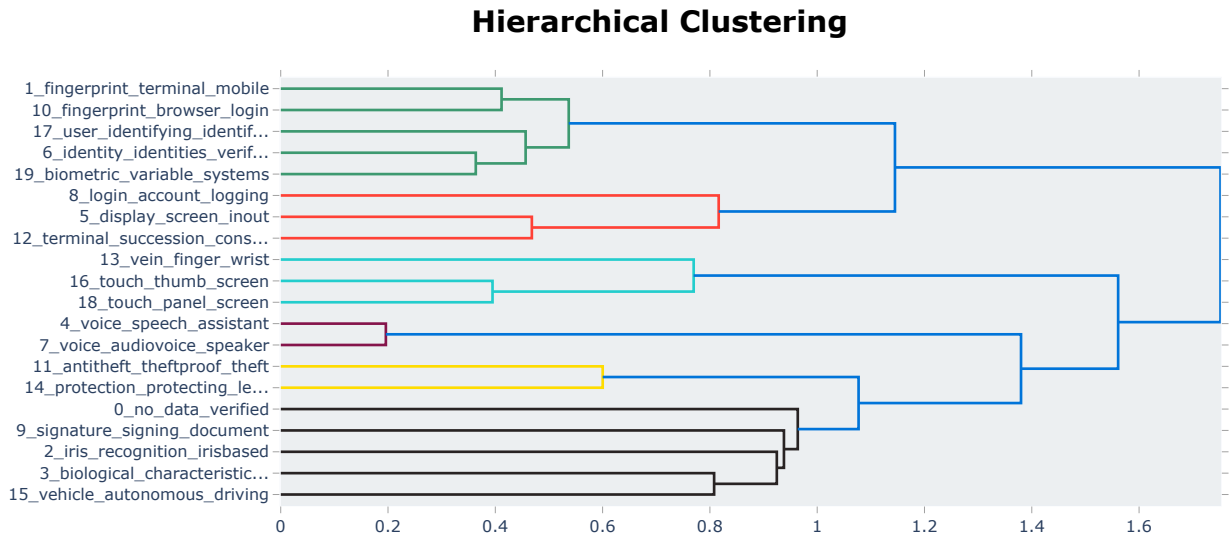
In [39]:

```
# We can visualize how similar certain topics are to each other
model.visualize_heatmap()
```



In [40]:

```
# Topics with the same colors have been grouped together which are similar one to another.
model.visualize_hierarchy(top_n_topics=20)
```



In [41]:

```
# Similar_topics contains the topics index from most similar to least similar.
# similarity contains the similarity scores in descending order.
similar_topic = input('Enter the name of word that you want to seach in topics => ')
similar_topics, similarity = model.find_topics(similar_topic, top_n = 3)
most_similar = similar_topics[0]
print("Most Similar Topic Info=\n {}".format(model.get_topic(most_similar)))
print("Similarity Score=> {}".format(similarity[0]))
print(similar_topics)
```

```
Enter the name of word that you want to seach in topics => eye
Most Similar Topic Info=
 [('eye', 0.10358731912455506), ('gaze', 0.05157348414122114), ('tracking', 0.03500238268619679), ('pupil', 0.0345169069655
2481), ('eyeball', 0.03137900633229529), ('movement', 0.02646420129327364), ('ocular', 0.01774499725361936), ('myopia', 0.0
16565715372948257), ('vision', 0.014325967817005875), ('detailed', 0.013165454088432886)]
Similarity Score=> 0.9267841817006135
[20, 293, 22]
```

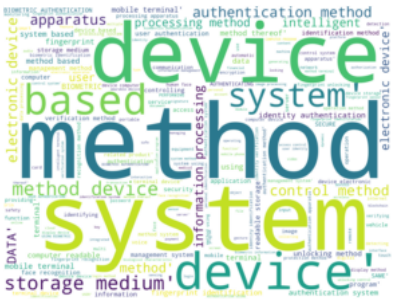
In [42]:

```
docs
```

```
 'Features management method, device, storage medium device and ',
 'intelligent work management system and method based on face recognition',
 'Shunting logistic distribution system and method',
 'Interface customizing method of doctor evaluation system and teacher terminal platform',
 'SYSTEM AND METHOD FOR FINANCIAL TRANSACTION VERIFICATION BASED ON A BIOMETRIC AUTHENTICATION',
 'A remote credit method, device, storage medium device and',
 'Integrated computer based on totem processor',
 'Campus data tracking and inquiring platform and inquiring method thereof',
 'Database information management method and system',
 'METHOD AND SERVER FOR PROVIDING SERVICE OF DISITAL SIGNATURE BASED ON FACE RECOGNITION',
 'safety method and system of safety',
 'Unlocking method based on Huarong Road, system, storage medium and computer device',
 'Mobile phone online loan information retrieval system',
 'A multi-angle face recognition comparison system',
 'METHOD AND SYSTEM FOR CHROMINANCE-BASED FACE LIVENESS DETECTION',
 'A control method of electric power tool, system, device and storage medium',
 'Storage method and system of radiological image data',
 'Self-adaptive identity real-time verification method, device, device and storage medium',
 'A risk level login processing method, system, storage medium device and',
 'Method and device for changing insurance contract data',
```

In [43]:

```
# Create stopword list :
stop_words = set(STOPWORDS)
stop_words.update(['ourselves', 'hers', 'between', 'yourself', 'but', 'again', 'there', 'about', 'once', 'during', 'out',
                   'very', 'having', 'with', 'they', 'own', 'an', 'be', 'some', 'for', 'do', 'its', 'yours', 'such', 'into',
                   'of', 'most', 'itself', 'other', 'off', 'is', 's', 'am', 'or', 'who', 'as', 'from', 'him', 'each', 'the',
                   'themselves', 'until', 'below', 'are', 'we', 'these', 'your', 'his', 'through', 'don', 'nor', 'me', 'were',
                   'her', 'more', 'himself', 'this', 'down', 'should', 'our', 'their', 'while', 'above', 'both', 'up', 'to',
                   'ours', 'had', 'she', 'all', 'no', 'when', 'at', 'any', 'before', 'them', 'same', 'and', 'been', 'have',
                   'in','will', 'on', 'does', 'yourselves', 'then', 'that', 'because', 'what', 'over', 'why', 'so',
                   'can', 'did', 'not','now', 'under', 'he', 'you', 'herself', 'has', 'just', 'where', 'too', 'only',
                   'myself', 'which', 'those', 'i','after', 'few', 'whom', 't', 'being', 'if', 'theirs', 'my',
                   'against', 'a', 'by', 'doing', 'it', 'how', 'further', 'was', 'here', 'than'])

# Generate a word cloud image
wordcloud = WordCloud(stopwords=stop_words, background_color="white", width=2000, height=1500).generate(str(docs))

# Display the generated image:
# the matplotlib way:
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```



In [45]:

```
# Stemming and Lemmatization
from nltk.stem.porter import PorterStemmer
from nltk.stem.wordnet import WordNetLemmatizer

porter = PorterStemmer()
wordnet = WordNetLemmatizer()

preprocessed_docs = []

for doc in stop_words:
    final_doc = []
    for word in doc:
        final_doc.append(porter.stem(word))
        final_doc.append(wordnet.lemmatize(word))

    preprocessed_docs.append(final_doc)

print(preprocessed_docs)
```

```
[['w', 'w', 'h', 'h', 'y', 'y'], ['t', 't', 'h', 'h', 'a', 'a', 't', 't'], ['o', 'o', 'n', 'n'], ['h', 'h', 'o', 'o', 'w',
'w'], ['s', 's', 'h', 'h', 'e', 'e', "'", "'", 's', 's'], ['t', 't', 'h', 'h', 'e', 'e', 'i', 'i', 'r', 'r', 's', 's'],
['w', 'w', 'h', 'h', 'o', 'o', 'm', 'm'], ['i', 'i', 't', 't', 's', 's', 'e', 'e', 'l', 'l', 'f', 'f'], ['a', 'a', 'r',
'r', 'e', 'e', 'n', 'n', "'", "'", 't', 't'], ['o', 'o', 'u', 'u', 'g', 'g', 'h', 'h', 't', 't'], ['w', 'w', 'h', 'h', 'a',
'a', 't', 't'], ['d', 'd', 'i', 'i', 'd', 'd', 'n', 'n', "'", "'", 't', 't'], ['y', 'y', 'o', 'o', 'u', 'u', 'r', 'r'],
['w', 'w', 'a', 'a', 's', 's', 'n', 'n', "'", "'", 't', 't'], ['y', 'y', 'o', 'o', 'u', 'u'], ['o', 'o', 'n', 'n', 'l',
'l', 'y', 'y'], ['i', 'i', "'", "'", 'l', 'l', 'l', 'l'], ['b', 'b', 'y', 'y'], ['a', 'a', 'b', 'b', 'o', 'o', 'u', 'u',
't', 't'], ['t', 't', 'h', 'h', 'e', 'e', 'n', 'n'], ['h', 'h', 'o', 'o', 'w', 'w', "'", "'", 's', 's'], ['o', 'o', 'u',
'u', 'r', 'r'], ['o', 'o', 'f', 'f'], ['a', 'a', 'l', 'l', 's', 's', 'o', 'o'], ['i', 'i', 'f', 'f'], ['d', 'd', 'u', 'u',
'r', 'r', 'i', 'i', 'n', 'n', 'g', 'g'], ['w', 'w', 'h', 'h', 'o', 'o', "'", "'", 's', 's'], ['h', 'h', 'a', 'a', 'v', 'v',
'e', 'e', 'n', 'n', "'", "'", 't', 't'], ['s', 's', 'o', 'o'], ['t', 't'], ['a', 'a'], ['y', 'y', 'o', 'o', 'u', 'u', "'",
"'", 'v', 'v', 'e', 'e'], ['a', 'a', 'f', 'f', 't', 't', 'e', 'e', 'r', 'r'], ['h', 'h', 'e', 'e'], ['l', 'l', 'e', 'e',
't', 't', "'", "'", 's', 's'], ['b', 'b', 'u', 'u', 't', 't'], ['i', 'i', "'", "'", 'd', 'd'], ['h', 'h', 'o', 'o', 'w',
'w', 'e', 'e', 'v', 'v', 'e', 'e', 'r', 'r'], ['i', 'i', 's', 's'], ['h', 'h', 'a', 'a', 'd', 'd'], ['f', 'f', 'r', 'r',
'o', 'o', 'm', 'm'], ['o', 'o', 'n', 'n', 'c', 'c', 'e', 'e'], ['h', 'h', 'e', 'e', "'", "'", 's', 's'], ['w', 'w', 'e',
'e', 'r', 'r', 'e', 'e'], ['g', 'g', 'e', 'e', 't', 't'], ['a', 'a', 'n', 'n'], ['m', 'm', 'e', 'e'], ['h', 'h', 'e', 'e',
'r', 'r', 'e', 'e'], ['a', 'a', 'n', 'n', 'y', 'y'], ['e', 'e', 'l', 'l', 's', 's', 'e', 'e'], ['t', 't', 'h', 'h', 'e',
'e', 'y', 'y', "'", "'", 'v', 'v', 'e', 'e'], ['h', 'h', 'a', 'a', 'd', 'd', 'n', 'n', "'", "'", 't', 't'], ['u', 'u', 'n',
'n', 't', 't', 'i', 'i', 'l', 'l'], ['i', 'i', "'", "'", 'm', 'm'], ['a', 'a', 'g', 'g', 'a', 'a', 'i', 'i', 'n', 'n', 's',
's', 't', 't'], ['n', 'n', 'o', 'o', 't', 't'], ['a', 'a', 'r', 'r', 'e', 'e'], ['t', 't', 'h', 'h', 'e', 'e', 'y', 'y',
"'", "'", 'l', 'l', 'l', 'l'], ['t', 't', 'h', 'h', 'i', 'i', 's', 's'], ['h', 'h', 'e', 'e', "'", "'", 'l', 'l', 'l',
'l'], ['b', 'b', 'e', 'e'], ['a', 'a', 't', 't'], ['i', 'i', 's', 's', 'n', 'n', "'", "'", 't', 't'], ['c', 'c', 'a', 'a',
'n', 'n', 'n', 'n', 'o', 'o', 't', 't'], ['w', 'w', 'h', 'h', 'i', 'i', 'c', 'c', 'h', 'h'], ['t', 't', 'h', 'h', 'e', 'e',
'i', 'i', 'r', 'r'], ['t', 't', 'h', 'h', 'e', 'e', 'y', 'y'], ['h', 'h', 'a', 'a', 's', 's', 'n', 'n', "'", "'", 't',
't'], ['y', 'y', 'o', 'o', 'u', 'u', "'", "'", 'l', 'l', 'l', 'l'], ['l', 'l', 'i', 'i', 'k', 'k', 'e', 'e'], ['b', 'b',
'e', 'e', 'i', 'i', 'n', 'n', 'g', 'g'], ['a', 'a', 's', 's'], ['e', 'e', 'a', 'a', 'c', 'c', 'h', 'h'], ['w', 'w', 'h',
'h', 'o', 'o'], ['f', 'f', 'u', 'u', 'r', 'r', 't', 't', 'h', 'h', 'e', 'e', 'r', 'r'], ['y', 'y', 'o', 'o', 'u', 'u', "'",
"'", 'r', 'r', 'e', 'e'], ['d', 'd', 'o', 'o', 'n', 'n'], ['s', 's', 'o', 'o', 'm', 'm', 'e', 'e'], ['w', 'w', 'e', 'e',
"'", "'", 'l', 'l', 'l', 'l'], ['h', 'h', 'a', 'a', 'v', 'v', 'e', 'e'], ['i', 'i'], ['o', 'o', 'w', 'w', 'n', 'n'], ['m',
'm', 'y', 'y'], ['n', 'n', 'o', 'o'], ['d', 'd', 'o', 'o'], ['i', 'i', 't', 't', "'", "'", 's', 's'], ['s', 's', 'h', 'h',
'o', 'o', 'u', 'u', 'l', 'l', 'd', 'd'], ['t', 't', 'h', 'h', 'e', 'e'], ['h', 'h', 'e', 'e', 'r', 'r', 's', 's', 'e', 'e',
'l', 'l', 'f', 'f'], ['s', 's', 'i', 'i', 'n', 'n', 'c', 'c', 'e', 'e'], ['b', 'b', 'e', 'e', 'l', 'l', 'o', 'o', 'w',
'w'], ['m', 'm', 'o', 'o', 's', 's', 't', 't'], ['y', 'y', 'o', 'o', 'u', 'u', 'r', 'r', 's', 's', 'e', 'e', 'l', 'l', 'v',
'v', 'e', 'e', 's', 's'], ['v', 'v', 'e', 'e', 'r', 'r', 'y', 'y'], ['d', 'd', 'o', 'o', 'w', 'w', 'n', 'n'], ['o', 'o',
'f', 'f', 'f', 'f'], ['s', 's'], ['r', 'r'], ['m', 'm', 'y', 'y', 's', 's', 'e', 'e', 'l', 'l', 'f', 'f'], ['w', 'w', 'e',
'e', "'", "'", 'd', 'd'], ['h', 'h', 't', 't', 't', 't', 'p', 'p'], ['t', 't', 'h', 'h', 'e', 'e', 'm', 'm'], ['h', 'h',
'i', 'i', 's', 's'], ['h', 'h', 'e', 'e', 'r', 'r', 'e', 'e', "'", "'", 's', 's'], ['s', 's', 'u', 'u', 'c', 'c', 'h',
'h'], ['t', 't', 'o', 'o', 'o', 'o'], ['w', 'w', 'h', 'h', 'e', 'e', 'r', 'r', 'e', 'e'], ['h', 'h', 'e', 'e', 'r', 'r'],
['d', 'd', 'i', 'i', 'd', 'd'], ['w', 'w', 'a', 'a', 's', 's'], ['i', 'i', 't', 't', 's', 's'], ['b', 'b', 'e', 'e', 'f',
'f', 'o', 'o', 'r', 'r', 'e', 'e'], ['n', 'n', 'o', 'o', 'r', 'r'], ['t', 't', 'h', 'h', 'a', 'a', 't', 't', "'", "'", 's',
's'], ['c', 'c', 'o', 'o', 'm', 'm'], ['c', 'c', 'o', 'o', 'u', 'u', 'l', 'l', 'd', 'd', 'n', 'n', "'", "'", 't', 't'],
['c', 'c', 'a', 'a', 'n', 'n', "'", "'", 't', 't'], ['t', 't', 'h', 'h', 'r', 'r', 'o', 'o', 'u', 'u', 'g', 'g', 'h', 'h'],
['n', 'n', 'o', 'o', 'w', 'w'], ['t', 't', 'h', 'h', 'e', 'e', 'y', 'y', "'", "'", 'd', 'd'], ['t', 't', 'h', 'h', 'e',
'e', 'y', 'y', "'", "'", 'r', 'r', 'e', 'e'], ['w', 'w', 'h', 'h', 'y', 'y', "'", "'", 's', 's'], ['s', 's', 'h', 'h', 'e',
'e'], ['b', 'b', 'e', 'e', 'c', 'c', 'a', 'a', 'u', 'u', 's', 's', 'e', 'e'], ['o', 'o', 't', 't', 'h', 'h', 'e', 'e', 'r',
'r'], ['o', 'o', 'v', 'v', 'e', 'e', 'r', 'r'], ['h', 'h', 'e', 'e', "'", "'", 'd', 'd'], ['j', 'j', 'u', 'u', 's', 's',
't', 't'], ['f', 'f', 'o', 'o', 'r', 'r'], ['s', 's', 'a', 'a', 'm', 'm', 'e', 'e'], ['b', 'b', 'e', 'e', 'e', 'e', 'n',
'n'], ['w', 'w', 'e', 'e'], ['o', 'o', 't', 't', 'h', 'h', 'e', 'e', 'r', 'r', 'w', 'w', 'i', 'i', 's', 's', 'e', 'e'],
['w', 'w', 'e', 'e', "'", "'", 'v', 'v', 'e', 'e'], ['d', 'd', 'o', 'o', 'i', 'i', 'n', 'n', 'g', 'g'], ['a', 'a', 'm',
'm'], ['e', 'e', 'v', 'v', 'e', 'e', 'r', 'r'], ['s', 's', 'h', 'h', 'e', 'e', "'", "'", 'l', 'l', 'l', 'l'], ['y', 'y',
'o', 'o', 'u', 'u', 'r', 'r', 's', 's', 'e', 'e', 'l', 'l', 'f', 'f'], ['m', 'm', 'o', 'o', 'r', 'r', 'e', 'e'], ['u', 'u',
'p', 'p'], ['a', 'a', 'l', 'l', 'l', 'l'], ['c', 'c', 'a', 'a', 'n', 'n'], ['i', 'i', 'n', 'n', 't', 't', 'o', 'o'], ['t',
't', 'h', 'h', 'e', 'e', 'm', 'm', 's', 's', 'e', 'e', 'l', 'l', 'v', 'v', 'e', 'e', 's', 's'], ['o', 'o', 'u', 'u', 'r',
'r', 's', 's'], ['t', 't', 'h', 'h', 'e', 'e', 's', 's', 'e', 'e'], ['a', 'a', 'n', 'n', 'd', 'd'], ['s', 's', 'h', 'h',
'o', 'o', 'u', 'u', 'l', 'l', 'd', 'd', 'n', 'n', "'", "'", 't', 't'], ['t', 't', 'h', 'h', 'e', 'e', 'r', 'r', 'e', 'e'],
['w', 'w', 'i', 'i', 'l', 'l', 'l', 'l'], ['w', 'w', 'o', 'o', 'u', 'u', 'l', 'l', 'd', 'd'], ['t', 't', 'h', 'h', 'a',
'a', 'n', 'n'], ['w', 'w', 'e', 'e', 'r', 'r', 'e', 'e', 'n', 'n', "'", "'", 't', 't'], ['t', 't', 'o', 'o'], ['k', 'k'],
['d', 'd', 'o', 'o', 'e', 'e', 's', 's', 'n', 'n', "'", "'", 't', 't'], ['b', 'b', 'e', 'e', 't', 't', 'w', 'w', 'e', 'e',
'e', 'e', 'n', 'n'], ['w', 'w', 'o', 'o', 'u', 'u', 'l', 'l', 'd', 'd', 'n', 'n', "'", "'", 't', 't'], ['d', 'd', 'o', 'o',
'n', 'n', "'", "'", 't', 't'], ['h', 'h', 'i', 'i', 'm', 'm'], ['c', 'c', 'o', 'o', 'u', 'u', 'l', 'l', 'd', 'd'], ['y',
'y', 'o', 'o', 'u', 'u', 'r', 'r', 's', 's'], ['b', 'b', 'o', 'o', 't', 't', 'h', 'h'], ['o', 'o', 'r', 'r'], ['w', 'w',
'i', 'i', 't', 't', 'h', 'h'], ['t', 't', 'h', 'h', 'o', 'o', 's', 's', 'e', 'e'], ['w', 'w', 'h', 'h', 'e', 'e', 'n', 'n',
"'", "'", 's', 's'], ['s', 's', 'h', 'h', 'a', 'a', 'l', 'l', 'l', 'l'], ['h', 'h', 'e', 'e', 'n', 'n', 'c', 'c', 'e',
'e'], ['m', 'm', 'u', 'u', 's', 's', 't', 't', 'n', 'n', "'", "'", 't', 't'], ['i', 'i', "'", "'", 'v', 'v', 'e', 'e'],
['a', 'a', 'b', 'b', 'o', 'o', 'v', 'v', 'e', 'e'], ['h', 'h', 'a', 'a', 'v', 'v', 'i', 'i', 'n', 'n', 'g', 'g'], ['i',
'i', 't', 't'], ['w', 'w', 'h', 'h', 'e', 'e', 'r', 'r', 'e', 'e', "'", "'", 's', 's'], ['w', 'w', 'h', 'h', 'a', 'a', 't',
't', "'", "'", 's', 's'], ['i', 'i', 'n', 'n'], ['o', 'o', 'u', 'u', 't', 't'], ['t', 't', 'h', 'h', 'e', 'e', 'r', 'r',
'e', 'e', "'", "'", 's', 's'], ['h', 'h', 'i', 'i', 'm', 'm', 's', 's', 'e', 'e', 'l', 'l', 'f', 'f'], ['w', 'w', 'h', 'h',
'e', 'e', 'n', 'n'], ['w', 'w', 'w', 'w', 'w', 'w'], ['o', 'o', 'u', 'u', 'r', 'r', 's', 's', 'e', 'e', 'l', 'l', 'v', 'v',
'e', 'e', 's', 's'], ['s', 's', 'h', 'h', 'a', 'a', 'n', 'n', "'", "'", 't', 't'], ['w', 'w', 'o', 'o', 'n', 'n', "'", "'",
't', 't'], ['w', 'w', 'e', 'e', "'", "'", 'r', 'r', 'e', 'e'], ['h', 'h', 'e', 'e', 'r', 'r', 's', 's'], ['y', 'y', 'o',
'o', 'u', 'u', "'", "'", 'd', 'd'], ['t', 't', 'h', 'h', 'e', 'e', 'r', 'r', 'e', 'e', 'f', 'f', 'o', 'o', 'r', 'r', 'e',
'e'], ['a', 'a', 'g', 'g', 'a', 'a', 'i', 'i', 'n', 'n'], ['f', 'f', 'e', 'e', 'w', 'w'], ['d', 'd', 'o', 'o', 'e', 'e',
's', 's'], ['s', 's', 'h', 'h', 'e', 'e', "'", "'", 'd', 'd'], ['u', 'u', 'n', 'n', 'd', 'd', 'e', 'e', 'r', 'r'], ['w',
'w', 'h', 'h', 'i', 'i', 'l', 'l', 'e', 'e'], ['h', 'h', 'a', 'a', 's', 's']]
```