Text Analysis

Process for text analysis

Text analysis is the process of using computer systems to read and understand human-written text for business insights. Text analysis software can independently classify, sort, and extract information from text to identify patterns, relationships, sentiments, and other actionable knowledge.

Packages being used for our project.

* NumPy
* Pandas
* OS
* JSON
* Matplotlib
* Seaborn
* Tqdm
* Jieba
* Time
* Sklearn

We need to download the font to understand and analyse Chinese language

!wget https://github.com/adobe-fonts/source-han-sans/raw/release/SubsetOTF/SourceHanSansCN.zip

!unzip -j "SourceHanSansCN.zip" "SourceHanSansCN/SourceHanSansCN-Regular.otf" -d "."

!rm SourceHanSansCN.zip

!ls

Loading the font from the downloaded files

import matplotlib.font\_manager as fm

font\_path = './SourceHanSansCN-Regular.otf'

prop = fm.FontProperties(fname=font\_path)

NOTE\*

Chinese does not have flexionary forms and also does not use spaces between ideograms to mark separate words. In the same time, some concepts are using 2 ore more ideograms in a sequence. The reader will cut in mind during reading the sequences of ideograms in groups, corresponding to different concepts, based on context. We will use **jieba** library to separate the ideograms in groups.

Function to get the most frequently used word

stopwords = set(STOPWORDS)

def show\_wordcloud(data, font\_path=font\_path, title = None):

wordcloud = WordCloud(

background\_color='white',

stopwords=stopwords,

font\_path=font\_path,

max\_words=50,

max\_font\_size=40,

scale=5,

random\_state=1

).generate(str(data))

fig = plt.figure(1, figsize=(10,10))

plt.axis('off')

if title:

prop = fm.FontProperties(fname=font\_path)

fig.suptitle(title, fontsize=40, fontproperties=prop)

fig.subplots\_adjust(top=2.3)

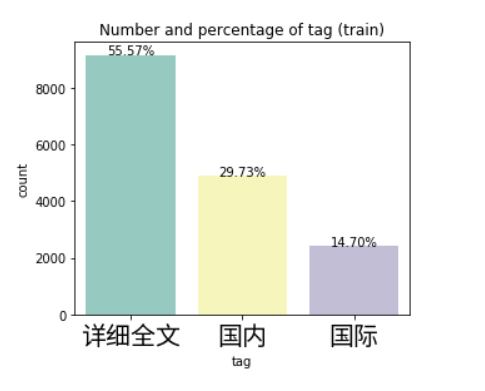
plt.imshow(wordcloud)

plt.show()

Splitting the data into sub parts i.e. Train data and Test data

train\_df, test\_df = train\_test\_split(data\_df, test\_size = 0.2, random\_state = 42)

This is the graphical representation of the split data.



A graph with different colored bars

Description automatically generated

Reference

Kaggle(sample Data and code for Chinese text) <https://www.kaggle.com/code/gpreda/chinese-text-classification?scriptVersionId=27601143&cellId=2>

Basics of text analysis

<https://www.youtube.com/watch?v=_rB7ZjamPwE&pp=ygUcY2hpbmVzZSB0ZXh0IGFuYWx5c2lzIHB5dGhvbg%3D%3D>

To understand Chinese language and the logics of how Chinese language works. Git hub repos

<https://github.com/dhchenx/cn-sentiment-measures>

<https://github.com/ChihoLeung/RoBERTa_Emotion_Classification>

<https://github.com/imsanjoykb/Saici.ai-NLP>