ECNU

基于Twitter的评论爬取分析
Trump2024美国总统大选成功的概率

POWERPOINT DESIGN

罗文琦 刘子阳

2024.12.6

CONTENTS

01

项目概述

Project Overview

项目背景

分析Twitter评论数据，

预测Trump2024年美国总统大选成功的概率。

项目目标

Step1

爬取Twitter

关于Trump和Harris

大选的推文、评论内容。

Step2

数据预处理

正则表达式+dropna

Step3

Spacy模型

关键词抓取

Vader模型

情感分析

Step4

LSTM深度学习

推文的情绪和热度

预测大选结果

Step5

可视化

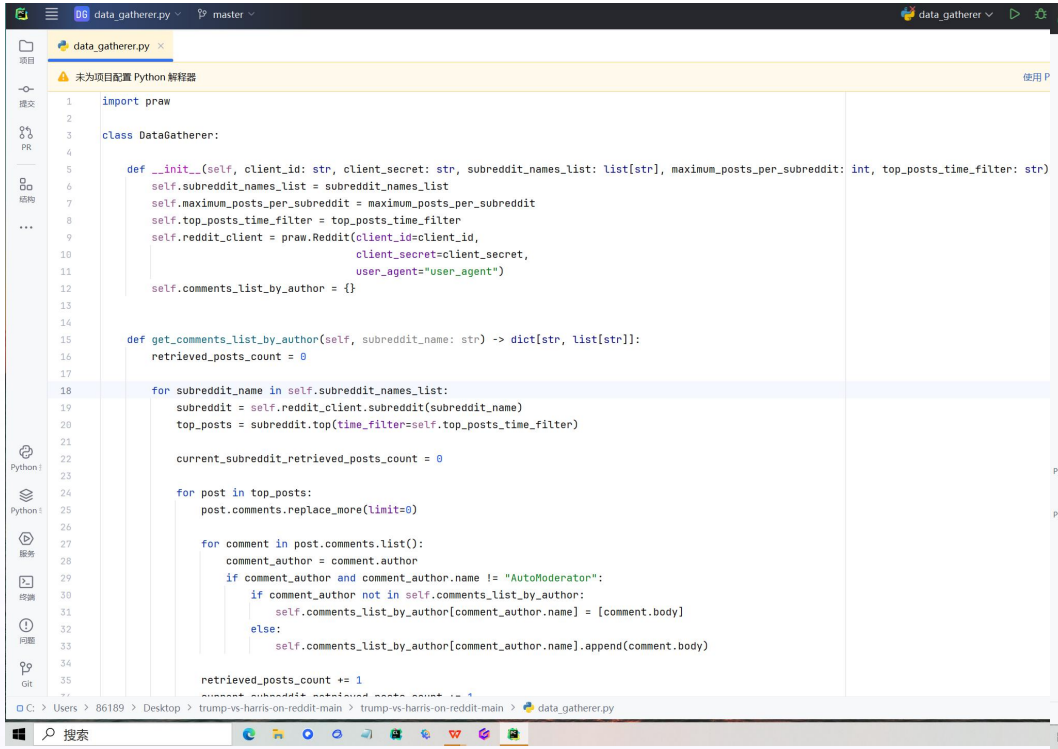
数据大屏+词云+聚类

02

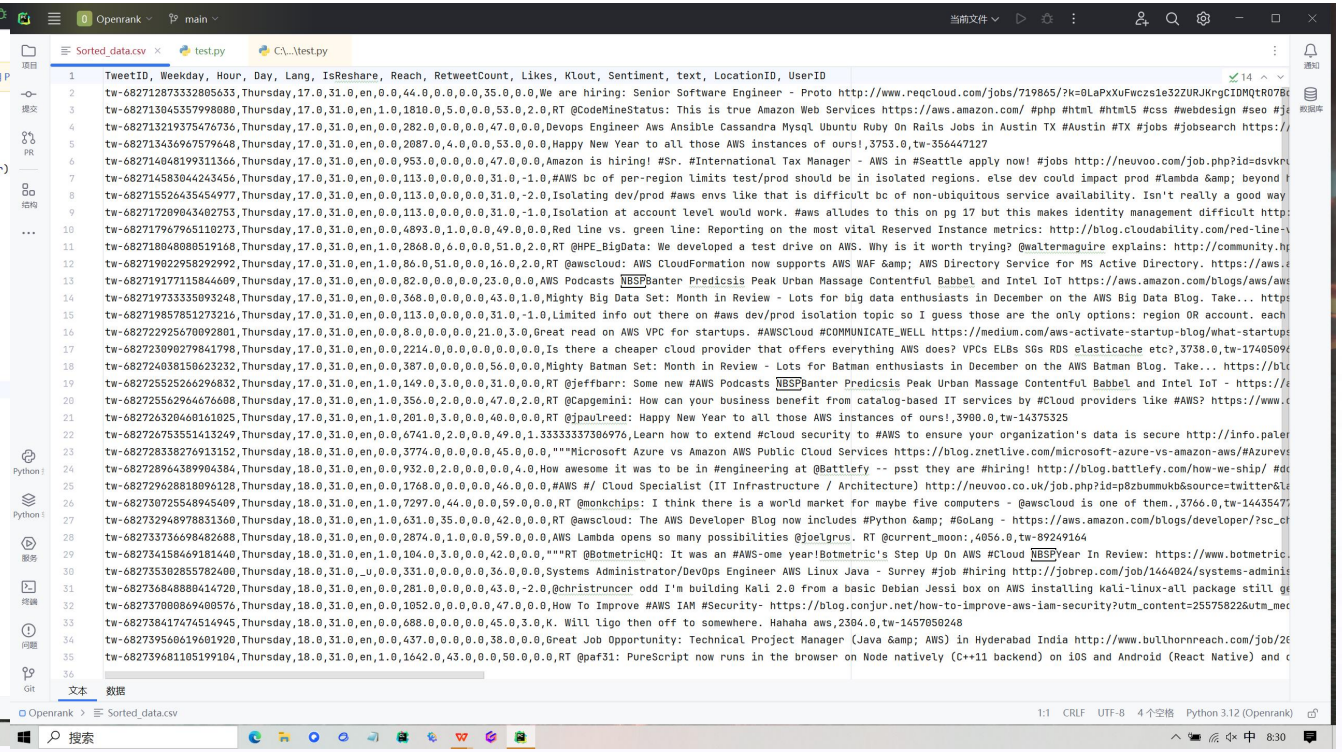
数据获取

Data Acquistion

数据获取



```python
import praw

class DataGatherer:

    def __init__(self, client_id: str, client_secret: str, subreddit_names_list: list[str], maximum_posts_per_subreddit: int, top_posts_time_filter: str):
        self.subreddit_names_list = subreddit_names_list
        self.maximum_posts_per_subreddit = maximum_posts_per_subreddit
        self.top_posts_time_filter = top_posts_time_filter
        self.reddit_client = praw.Reddit(client_id=client_id,
                                         client_secret=client_secret,
                                         user_agent="user_agent")
        self.comments_list_by_author = {}


    def get_comments_list_by_author(self, subreddit_name: str) -> dict[str, list[str]]:
        retrieved_posts_count = 0

        for subreddit_name in self.subreddit_names_list:
            subreddit = self.reddit_client.subreddit(subreddit_name)
            top_posts = subreddit.top(time_filter=self.top_posts_time_filter)

            current_subreddit_retrieved_posts_count = 0

            for post in top_posts:
                post.comments.replace_more(limit=0)

                for comment in post.comments.list():
                    comment_author = comment.author
                    if comment_author and comment_author.name != "AutoModerator":
                        if comment_author not in self.comments_list_by_author:
                            self.comments_list_by_author[comment_author.name] = [comment.body]
                        else:
                            self.comments_list_by_author[comment_author.name].append(comment.body)

                retrieved_posts_count += 1
```



```
TweetID, Weekday, Hour, Day, Lang, IsReshare, Reach, RetweetCount, Likes, Klout, Sentiment, text, LocationID, UserID
tw-682712873332805633,Thursday,17.0,31.0,en,0.0,44.0,0.0,0.0,35.0,0.0,We are hiring: Senior Software Engineer - Proto http://www.reqcloud.com/jobs/719865/?k=0LaPxXuFwczs1e32ZURJKrgCIDMQtRO78
tw-682713045357998080,Thursday,17.0,31.0,en,1.0,1810.0,5.0,0.0,53.0,2.0,RT @CodeMineStatus: This is true Amazon Web Services https://aws.amazon.com/ #php #html #html5 #css #webdesign #seo #
tw-682713219375476736,Thursday,17.0,31.0,en,0.0,282.0,0.0,0.0,47.0,0.0,Devops Engineer Aws Ansible Cassandra Mysql Ubuntu Ruby On Rails Jobs in Austin TX #Austin #TX #jobs #jobsearch https:/
tw-682713436965579648,Thursday,17.0,31.0,en,0.0,2087.0,4.0,0.0,53.0,0.0,Happy New Year to all those AWS instances of ours!,3753.0,tw-356447127
tw-682714048199311366,Thursday,17.0,31.0,en,0.0,953.0,0.0,0.0,47.0,0.0,Amazon is hiring! #Sr. #International Tax Manager - AWS in #Seattle apply now! #jobs http://neuvoo.com/job.php?id=dsvkr
tw-682714583044243456,Thursday,17.0,31.0,en,0.0,113.0,0.0,0.0,31.0,-1.0,#AWS bc of per-region limits test/prod should be in isolated regions. else dev could impact prod #lambda &amp;amp; beyond
tw-682715526435454977,Thursday,17.0,31.0,en,0.0,113.0,0.0,0.0,31.0,-2.0,Isolating dev/prod #aws envs like that is difficult bc of non-ubiquitous service availability. Isn't really a good way
tw-682717209043402753,Thursday,17.0,31.0,en,0.0,113.0,0.0,0.0,31.0,-1.0,Isolation at account level would work. #aws alludes to this on pg 17 but this makes identity management difficult http:
tw-682717967965110273,Thursday,17.0,31.0,en,0.0,4893.0,1.0,0.0,49.0,0.0,Red line vs. green line: Reporting on the most vital Reserved Instance metrics: http://blog.cloudability.com/red-line-v
tw-682718048080519168,Thursday,17.0,31.0,en,1.0,2868.0,6.0,0.0,51.0,2.0,RT @HPE_BigData: We developed a test drive on AWS. Why is it worth trying? @waltermaguire explains: http://community.hp
tw-682719022958292992,Thursday,17.0,31.0,en,1.0,86.0,51.0,0.0,16.0,2.0,RT @awscloud: AWS CloudFormation now supports AWS WAF &amp;amp; AWS Directory Service for MS Active Directory. https://aws.a
tw-682719177115844609,Thursday,17.0,31.0,en,0.0,82.0,0.0,0.0,23.0,0.0,AWS Podcasts NBSPBanter Predicsis Peak Urban Massage Contentful Babbel and Intel IoT https://aws.amazon.com/blogs/aws/aws
tw-682719733335093248,Thursday,17.0,31.0,en,0.0,368.0,0.0,0.0,43.0,1.0,Mighty Big Data Set: Month in Review - Lots for big data enthusiasts in December on the AWS Big Data Blog. Take... https
tw-682719857851273216,Thursday,17.0,31.0,en,0.0,113.0,0.0,0.0,31.0,-1.0,Limited info out there on #aws dev/prod isolation topic so I guess those are the only options: region OR account. each
tw-682722925670092801,Thursday,17.0,31.0,en,0.0,8.0,0.0,0.0,21.0,3.0,Great read on AWS VPC for startups. #AWSCloud #COMMUNICATE_WELL https://medium.com/aws-activate-startup-blog/what-startups
tw-682723090279841798,Thursday,17.0,31.0,en,0.0,2214.0,0.0,0.0,0.0,0.0,Is there a cheaper cloud provider that offers everything AWS does? VPCs ELBs SGs RDS elasticache etc?,3738.0,tw-1740509
tw-682724038150623232,Thursday,17.0,31.0,en,0.0,387.0,0.0,0.0,56.0,0.0,Mighty Batman Set: Month in Review - Lots for Batman enthusiasts in December on the AWS Batman Blog. Take... https://blo
tw-682725525266296832,Thursday,17.0,31.0,en,1.0,149.0,3.0,0.0,31.0,0.0,RT @jeffbarr: Some new #AWS Podcasts NBSPBanter Predicsis Peak Urban Massage Contentful Babbel and Intel IoT - https://a
tw-682725562964676608,Thursday,17.0,31.0,en,1.0,356.0,2.0,0.0,47.0,2.0,RT @Capgemini: How can your business benefit from catalog-based IT services by #Cloud providers like #AWS? https://www.c
tw-682726320460161025,Thursday,17.0,31.0,en,1.0,201.0,3.0,0.0,40.0,0.0,RT @jpaulreed: Happy New Year to all those AWS instances of ours!,3900.0,tw-14375325
tw-682726753551413249,Thursday,17.0,31.0,en,0.0,6741.0,2.0,0.0,49.0,1.33333337306976,Learn how to extend #cloud security to #AWS to ensure your organization's data is secure http://info.paler
tw-682728338276913152,Thursday,18.0,31.0,en,0.0,3774.0,0.0,0.0,45.0,0.0,"""Microsoft Azure vs Amazon AWS Public Cloud Services https://blog.znetlive.com/microsoft-azure-vs-amazon-aws/#Azurevs
tw-682728964389904384,Thursday,18.0,31.0,en,0.0,932.0,2.0,0.0,0.0,4.0,How awesome it was to be in #engineering at @Battlefy -- psst they are #hiring! http://blog.battlefy.com/how-we-ship/ #d
tw-682729628818096128,Thursday,18.0,31.0,en,0.0,1768.0,0.0,0.0,46.0,0.0,#AWS #/ Cloud Specialist (IT Infrastructure / Architecture) http://neuvoo.co.uk/job.php?id=p8zbummukb&source=twitter&la
tw-682730725548945409,Thursday,18.0,31.0,en,1.0,7297.0,44.0,0.0,59.0,0.0,RT @monkchips: I think there is a world market for maybe five computers - @awscloud is one of them.,3766.0,tw-14435477
tw-682732948978831360,Thursday,18.0,31.0,en,1.0,631.0,35.0,0.0,42.0,0.0,RT @awscloud: The AWS Developer Blog now includes #Python &amp;amp; #GoLang - https://aws.amazon.com/blogs/developer/?sc_ch
tw-682733736698482688,Thursday,18.0,31.0,en,0.0,2874.0,1.0,0.0,59.0,0.0,AWS Lambda opens so many possibilities @joelgrus. RT @current_moon:,4056.0,tw-89249164
tw-682734158469181440,Thursday,18.0,31.0,en,1.0,104.0,3.0,0.0,42.0,0.0,"""RT @BotmetricHQ: It was an #AWS-ome year!Botmetric's Step Up On AWS #Cloud NBSPYear In Review: https://www.botmetric.
tw-682735302855782400,Thursday,18.0,31.0,_u,0.0,331.0,0.0,0.0,36.0,0.0,Systems Administrator/DevOps Engineer AWS Linux Java - Surrey #job #hiring http://jobrep.com/job/1464024/systems-adminis
tw-682736848880414720,Thursday,18.0,31.0,en,0.0,281.0,0.0,0.0,43.0,-2.0,@christruncer odd I'm building Kali 2.0 from a basic Debian Jessi box on AWS installing kali-linux-all package still g
tw-682737008869400576,Thursday,18.0,31.0,en,0.0,1052.0,0.0,0.0,47.0,0.0,How To Improve #AWS IAM #Security- https://blog.conjur.net/how-to-improve-aws-iam-security?utm_content=25575822&utm_me
tw-682738417474514945,Thursday,18.0,31.0,en,0.0,688.0,0.0,0.0,45.0,3.0,K. Will ligo then off to somewhere. Hahaha aws,2304.0,tw-1457050248
tw-682739560619601920,Thursday,18.0,31.0,en,0.0,437.0,0.0,0.0,38.0,0.0,Great Job Opportunity: Technical Project Manager (Java &amp;amp; AWS) in Hyderabad India http://www.bullhornreach.com/job/2
tw-682739681105199104,Thursday,18.0,31.0,en,1.0,1642.0,43.0,0.0,50.0,0.0,RT @paf31: PureScript now runs in the browser on Node natively (C++11 backend) on iOS and Android (React Native) and
```

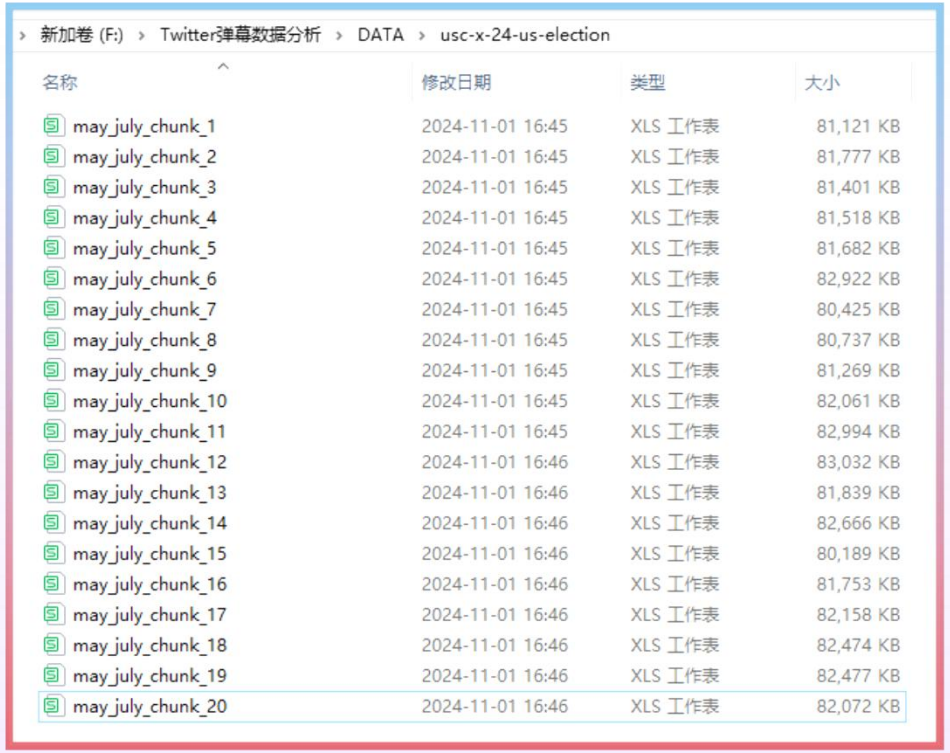
技术层面

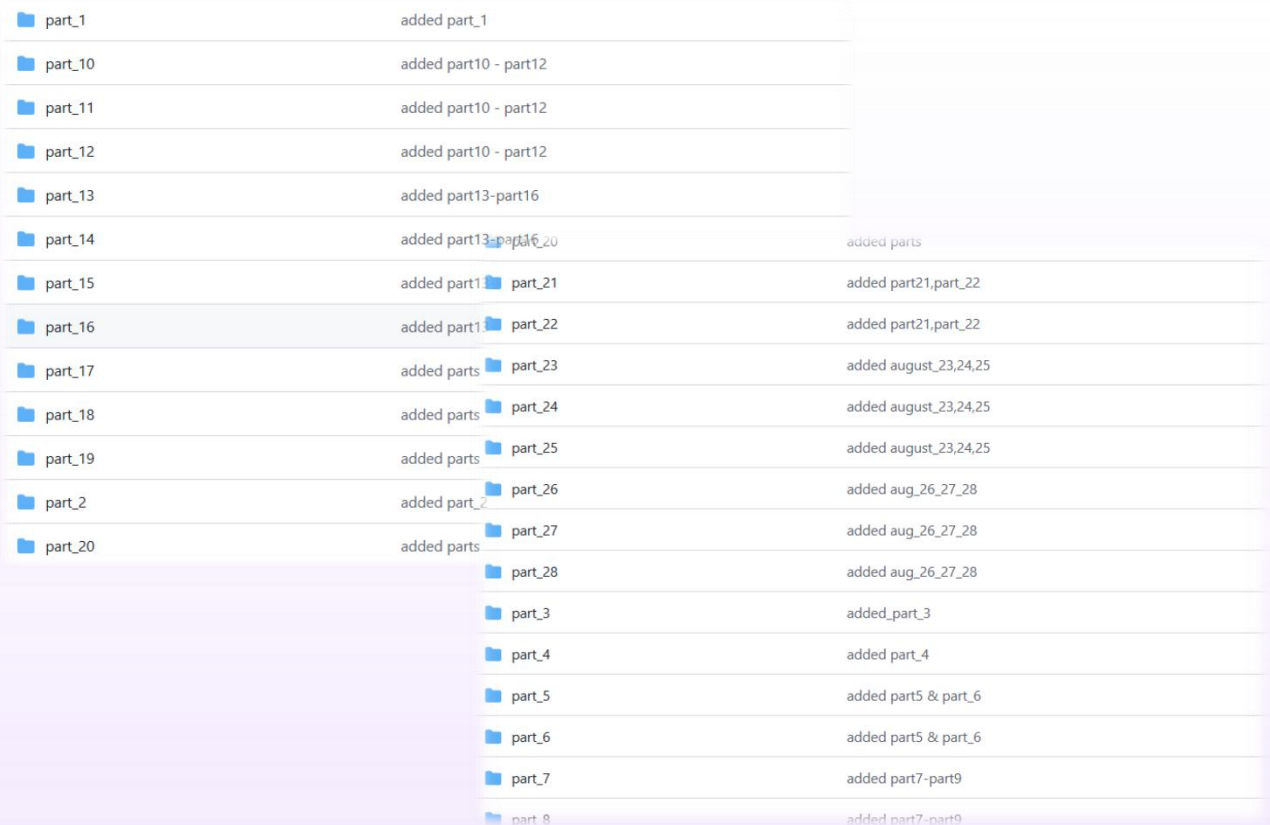
使用python和spyder作为爬取工具。

评论爬取

获取推文的评论，通过Tweet ID和API请求。

分批分页处理，应对Twitter对评论数据的爬取限制。





数据拼接

构建爬取数据的集合，将爬取数据的csv文件进行合并

同样使用python进行分批分类处理。

Openrank > test.py

Sorted_data.csv | output.csv | test.py | tweets_data.csv | C:\...\test.py

```python
import pandas as pd

# 定义文件路径列表
file_paths = [
    r'C:\Users\86189\Desktop\may_july_chunk_1.csv',
    r'C:\Users\86189\Desktop\may_july_chunk_2.csv',
    r'C:\Users\86189\Desktop\may_july_chunk_3.csv',
    r'C:\Users\86189\Desktop\may_july_chunk_4.csv',
    r'C:\Users\86189\Desktop\may_july_chunk_5.csv',
    r'C:\Users\86189\Desktop\may_july_chunk_6.csv',
    r'C:\Users\86189\Desktop\may_july_chunk_7.csv',
    r'C:\Users\86189\Desktop\may_july_chunk_8.csv',
    r'C:\Users\86189\Desktop\may_july_chunk_9.csv',
    r'C:\Users\86189\Desktop\may_july_chunk_10.csv'
]

# 读取第一个CSV文件
df = pd.read_csv(file_paths[0], low_memory=False)

# 读取其余的CSV文件并拼接
for i in range(1, len(file_paths)):
    df_i = pd.read_csv(file_paths[i], low_memory=False)
    # 拼接并去除重复的行
    df = pd.concat([df, df_i]).drop_duplicates()

# 保留需要的列（根据需求进行修改）
# 这里保留所有列（如果只想保留部分列，可以通过 df.iloc[:, :] 来选择）
df = df.iloc[:, :]  # 如果需要特定列，请修改这一行

# 保存拼接后的CSV文件
output_file = r'C:\Users\86189\Desktop\csv_merge.csv'

# 使用 utf-8 编码保存文件，或者使用 'ignore' 来忽略无法编码的字符
df.to_csv(output_file, index=False, encoding='utf-8')  # 使用utf-8编码
```

merge.py

```python
import pandas as pd

# 创建一个空的DataFrame用来存储合并的结果
merged_df = pd.DataFrame()

# 定义每次读取的块大小
chunksize = 10000
# 循环遍历每个文件，逐块读取并合并
for i in range(1, 21):
    file_name = f'{i}.csv'
    for chunk in pd.read_csv(file_name, chunksize=chunksize):
        # 如果是第二个及以后文件，去掉第一行（列名），假设第一个文件的列名是正确的且要保留
        if i > 1 and chunk.index[0] == 0:
            chunk = chunk.iloc[1:]
        # 将当前块的数据合并到最终的DataFrame中
        merged_df = pd.concat( objs: [merged_df, chunk], ignore_index=True)
# 保存最终的合并结果
merged_df.to_csv( path_or_buf: 'all_merged_file.csv', index=False)
```

03

数据处理

Data Processing

关键词提取

01 数据筛选

02 使用spacy提取评论中的关键词。

03 做出词云图进行可视化

数据清理后的部分结果

test

merge.py | data pre-processed.py

```python
import pandas as pd
import re

# 读取数据
df = pd.read_csv('pre_data.csv')

# 保留需要的列
df = df[['rawContent', 'replyCount', 'retweetCount', 'likeCount', 'quoteCount']]


# 清理文本数据
def clean_text(text):
    # 如果文本是缺失值或非字符串类型，则返回空字符串
    if not isinstance(text, str):
        return ''
    # 去除URL
    text = re.sub( pattern: r'http\S+|www\S+', repl: '', text)
    # 允许标点符号和中文字符
    text = re.sub( pattern: r'[^\w\s.,!?;:]', repl: '', text)  # 保留字母、数字、空格、常见标点符号
    # 如果文本为空，返回空字符串
    return text.strip()


df['cleaned_content'] = df['rawContent'].apply(lambda x: clean_text(x))
# 删除包含缺失值的行
df.dropna(inplace=True)

# 输出处理后的数据（可选）
df.to_csv('cleaned_data.csv', index=False)

# 你可以查看数据结构
print(df.head())

# 检查清理后的数据数量
print(f"Number of rows after cleaning: {len(df)}")
```

Python Packages

Git | Python Packages | TODO | Problems | Terminal | Services

利用正则化表达式去除URL并保留字母、数字、空格和常见符号

利用dropna函数，对不符合条件的内容做一个清洗

	rawContent	cleaned_content
2	@lukepbeasley I cant imagine anyo	lukepbeasley I cant imagine anyo
3	Voters can also sway me away from	Voters can also sway me away fro
4	@PoodleHead57 @BobOnderMO Can you	PoodleHead57 BobOnderMO Can you
5	@Morning_Joe @JoeNBC The fact rem	Morning_Joe JoeNBC The fact rema
6	@BidenHQ That's funny you're obvi	BidenHQ Thats funny youre obviou
7	#Internacional Las modificaciones introducidas por el grupo palestino HamÃ¡s en el acuerdo diseÃ±ado por el presidente de EEUU, Joe Biden,	Internacional Las modificaciones introducidas por el grupo palestino HamÃ¡s en el acuerdo diseÃ±ado por el presidente de EEUU, Joe Biden,
8	MAGA RAGES Over Hunter Biden Verd	MAGA RAGES Over Hunter Biden Ver
9	@harryjsisson There is no doubt i	harryjsisson There is no doubt i
10	@ShaunPMaca @MoneyTalkUS @byHeatherLong Hourly wages are irrelevant.	ShaunPMaca MoneyTalkUS byHeatherLong Hourly wages are irrelevant.
11	@nepolabo_maga ã ã ã ã â ª ã ã ¯ã ã ã ï¼ \â (â ò' â	nepolabo_maga

利用spacy进行关键词提取并过滤重要内容

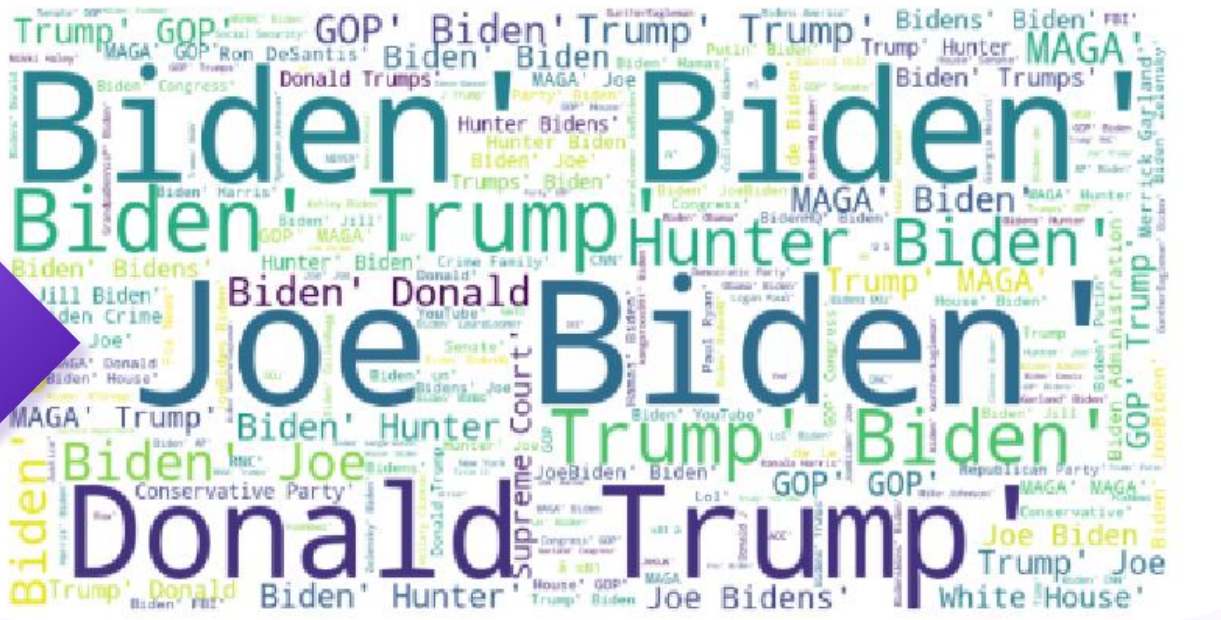
```python
# 加载spaCy模型
nlp = spacy.load('en_core_web_sm')

# 提取关键词（例如：人物、组织等）
def extract_keywords(text):  1 usage
    doc = nlp(text)
    keywords = [ent.text for ent in doc.ents if ent.label_ in ['PERSON', 'ORG']]
    return keywords

# 提取关键词
df['keywords'] = df['cleaned_content'].apply(extract_keywords)
```

cleaned_content	keywords
lukepbeasley I cant imagine anyone actual	['Donald']
Morning_Joe JoeNBC The fact remains that	['Joe Biden']
BidenHQ Thats funny youre obviously tryin	['BidenHQ']
Internacional Las modificaciones introduc	['Joe Biden', 'la retirada del', 'Franja de Gaza', 'semana de la implementaciÃ³n']
LadyMagaUSA nypost Lady Maga is here to s	['Lady Maga', 'HOORAY']
teenburger Fernand46357857 Terrorism is c	['Biden', 'Hamas', 'Biden']
MAGA hats are for morons racists	['MAGA']
BidenHQ Were still poorer than ever. We r	['Drop Biden', 'Trump']

数据可视化

```python
import pandas as pd
from wordcloud import WordCloud
import matplotlib.pyplot as plt

# 读取 Excel 文件
df = pd.read_excel('keywords_extracted.xlsx')  # 根据实际路径调整

# 假设你的 DataFrame 中有一列 'keywords'，它存储了提取的关键词
# 如果 'keywords' 是字符串类型，每行是关键词的列表，你可能需要先将它们合并成一个长字符串
all_keywords = ' '.join(df['keywords'].apply(lambda x: ' '.join(x) if isinstance(x, list) else str(x)))

# 创建词云对象
wordcloud = WordCloud(width=800, height=400, background_color='white').generate(all_keywords)

# 显示词云
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()
```

可视化



按照关键词统计数据做出词云图

绘制评论情绪分布的聚类图

04

数据分析

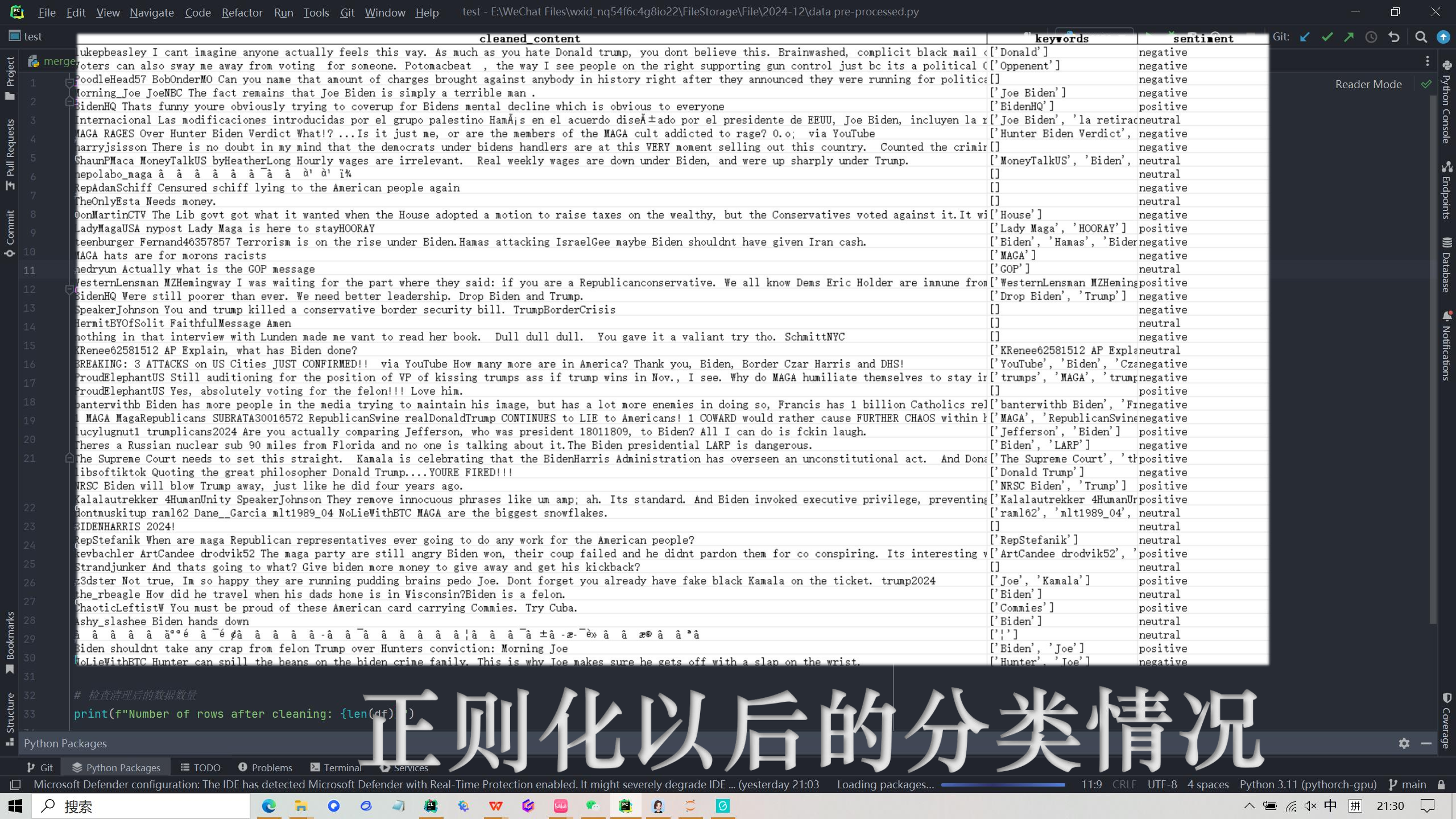
Data Analysis

情感分析

```python
# 初始化 VADER 情感分析器
sia = SentimentIntensityAnalyzer()

# 定义情感分类函数
def get_sentiment(text):  1 usage
    # 确保文本是字符串
    if not isinstance(text, str):
        text = str(text)  # 将非字符串类型转换为字符串
    score = sia.polarity_scores(text)
    if score['compound'] ⩾ 0.05:
        return 'positive'
    elif score['compound'] ⩽ -0.05:
        return 'negative'
    else:
        return 'neutral'
```

01 正则化表达

02 使用情感分析模型VADER对评论的情绪进行分类进行情感分析，将评论分为积极（支持某候选人）、消极（反对候选人）和中立（未表态）



cleaned_content	keywords	sentiment
lukepbeasley I cant imagine anyone actually feels this way. As much as you hate Donald trump, you dont believe this. Brainwashed, complicit black mail	['Donald']	negative
oters can also sway me away from voting for someone. Potomacbeat , the way I see people on the right supporting gun control just bc its a political	['Oppenent']	negative
oodleHead57 BobOnderMO Can you name that amount of charges brought against anybody in history right after they announced they were running for politica	[]	negative
orning_Joe JoeNBC The fact remains that Joe Biden is simply a terrible man .	['Joe Biden']	negative
idenHQ Thats funny youre obviously trying to coverup for Bidens mental decline which is obvious to everyone	['BidenHQ']	positive
nternacional Las modificaciones introducidas por el grupo palestino HamÃ¡s en el acuerdo diseÃ±ado por el presidente de EEUU, Joe Biden, incluyen la r	['Joe Biden', 'la retirac	neutral
AGA RAGES Over Hunter Biden Verdict What!? ...Is it just me, or are the members of the MAGA cult addicted to rage? O.o; via YouTube	['Hunter Biden Verdict',	negative
arryjsisson There is no doubt in my mind that the democrats under bidens handlers are at this VERY moment selling out this country. Counted the crimir	[]	negative
haunPMaca MoneyTalkUS byHeatherLong Hourly wages are irrelevant. Real weekly wages are down under Biden, and were up sharply under Trump.	['MoneyTalkUS', 'Biden',	neutral
epolabo_maga â â â â â â ¯â â à' à' ï¼	[]	neutral
epAdamSchiff Censured schiff lying to the American people again	[]	negative
heOnlyEsta Needs money.	[]	neutral
oonMartinCTV The Lib govt got what it wanted when the House adopted a motion to raise taxes on the wealthy, but the Conservatives voted against it.It wi	['House']	negative
adyMagaUSA nypost Lady Maga is here to stayHOORAY	['Lady Maga', 'HOORAY']	positive
eenburger Fernand46357857 Terrorism is on the rise under Biden.Hamas attacking IsraelGee maybe Biden shouldnt have given Iran cash.	['Biden', 'Hamas', 'Bider	negative
AGA hats are for morons racists	['MAGA']	negative
edryun Actually what is the GOP message	['GOP']	neutral
esternLensman MZHemingway I was waiting for the part where they said: if you are a Republicanconservative. We all know Dems Eric Holder are immune fron	['WesternLensman MZHeming	positive
idenHQ Were still poorer than ever. We need better leadership. Drop Biden and Trump.	['Drop Biden', 'Trump']	negative
peakerJohnson You and trump killed a conservative border security bill. TrumpBorderCrisis	[]	negative
ermitBYOfSolit FaithfulMessage Amen	[]	neutral
othing in that interview with Lunden made me want to read her book. Dull dull dull. You gave it a valiant try tho. SchmittNYC	[]	negative
Renee62581512 AP Explain, what has Biden done?	['KRenee62581512 AP Expla	neutral
REAKING: 3 ATTACKS on US Cities JUST CONFIRMED!! via YouTube How many more are in America? Thank you, Biden, Border Czar Harris and DHS!	['YouTube', 'Biden', 'Cza	negative
roudElephantUS Still auditioning for the position of VP of kissing trumps ass if trump wins in Nov., I see. Why do MAGA humiliate themselves to stay ir	['trumps', 'MAGA', 'trump	negative
roudElephantUS Yes, absolutely voting for the felon!!! Love him.	[]	positive
anterwithb Biden has more people in the media trying to maintain his image, but has a lot more enemies in doing so, Francis has 1 billion Catholics rel	['banterwithb Biden', 'Fr	negative
MAGA MagaRepublicans SUBRATA30016572 RepublicanSwine realDonaldTrump CONTINUES to LIE to Americans! 1 COWARD would rather cause FURTHER CHAOS within h	['MAGA', 'RepublicanSwine	negative
ucylugnut1 trumplicans2024 Are you actually comparing Jefferson, who was president 18011809, to Biden? All I can do is fckin laugh.	['Jefferson', 'Biden']	positive
heres a Russian nuclear sub 90 miles from Florida and no one is talking about it.The Biden presidential LARP is dangerous.	['Biden', 'LARP']	negative
he Supreme Court needs to set this straight. Kamala is celebrating that the BidenHarris Administration has overseen an unconstitutional act. And Dona	['The Supreme Court', 'th	positive
ibsoftiktok Quoting the great philosopher Donald Trump....YOURE FIRED!!!	['Donald Trump']	negative
RSC Biden will blow Trump away, just like he did four years ago.	['NRSC Biden', 'Trump']	positive
alalautrekker 4HumanUnity SpeakerJohnson They remove innocuous phrases like um amp; ah. Its standard. And Biden invoked executive privilege, preventing	['Kalalautrekker 4HumanUr	positive
ontmuskitup raml62 Dane__Garcia mlt1989_04 NoLieWithBTC MAGA are the biggest snowflakes.	['raml62', 'mlt1989_04',	neutral
IDENHARRIS 2024!	[]	neutral
epStefanik When are maga Republican representatives ever going to do any work for the American people?	['RepStefanik']	neutral
evbachler ArtCandee drodvik52 The maga party are still angry Biden won, their coup failed and he didnt pardon them for co conspiring. Its interesting v	['ArtCandee drodvik52', '	positive
trandjunker And thats going to what? Give biden more money to give away and get his kickback?	[]	neutral
3dster Not true, Im so happy they are running pudding brains pedo Joe. Dont forget you already have fake black Kamala on the ticket. trump2024	['Joe', 'Kamala']	positive
he_rbeagle How did he travel when his dads home is in Wisconsin?Biden is a felon.	['Biden']	neutral
haoticLeftistW You must be proud of these American card carrying Commies. Try Cuba.	['Commies']	positive
shy_slashee Biden hands down	['Biden']	neutral
â â â â â ä°°é â ¯é ¢â â â â â -â â ¯â â â â â â ¦â â â ¯â ±â -æ-¯è» â â æ® â â °â	['¦']	neutral
iden shouldnt take any crap from felon Trump over Hunters conviction: Morning Joe	['Biden', 'Joe']	positive
oLieWithBTC Hunter can spill the beans on the biden crime family. This is why Joe makes sure he gets off with a slap on the wrist.	['Hunter', 'Joe']	negative

```
# 检查清理后的数据数量
print(f"Number of rows after cleaning: {len(df)}")
```

正则化以后的分类情况

05

预测分析

Predictive Analysis

模型选择

使用机器学习模型Random Forest

使用历史选举数据和热度趋势作为训练集

趋势分析

利用Random Forest对不同候选人的支持情况进行深度学习并分类

通过训练数据学习情感和关键词的模式，预测不同候选人的支持情况。

```
# 将 'keywords' 和 'candidate' 转化为特征
vectorizer = TfidfVectorizer(max_features=5000)
X = vectorizer.fit_transform(df['keywords'])  # 使用 keywords 列进行特征提取

# 使用 candidate 列作为标签，分别为 Trump、Harris 或 Biden
y = df['candidate'].apply(lambda x: 1 if x == 'Trump' else (2 if x == 'Harris' else 0))  # Trump:1

# 划分训练集和测试集
X_train, X_test, y_train, y_test = train_test_split( *arrays: X, y, test_size=0.2, random_state=42)

# 训练 Random Forest 模型
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
```

```python
# 训练 Random Forest 模型
model = RandomForestClassifier(n_estimators=100, random_state=
model.fit(X_train, y_train)

# 获取 Trump 和 Harris 的预测概率
probabilities = model.predict_proba(X_test)

# 获取每个样本是 Trump（类别 1）或 Harris（类别 2）的概率
prob_trump = probabilities[:, 1]  # Trump 类别的预测概率
prob_harris = probabilities[:, 2]  # Harris 类别的预测概率

# 计算 Trump 和 Harris 获胜的支持比例
trump_support = np.mean(prob_trump > prob_harris) * 100
harris_support = np.mean(prob_harris > prob_trump) * 100

print(f"Trump Support: {trump_support:.2f}%")
print(f"Harris Support: {harris_support:.2f}%")
```

```
Trump Support: 62.44%
Harris Support: 0.92%
```

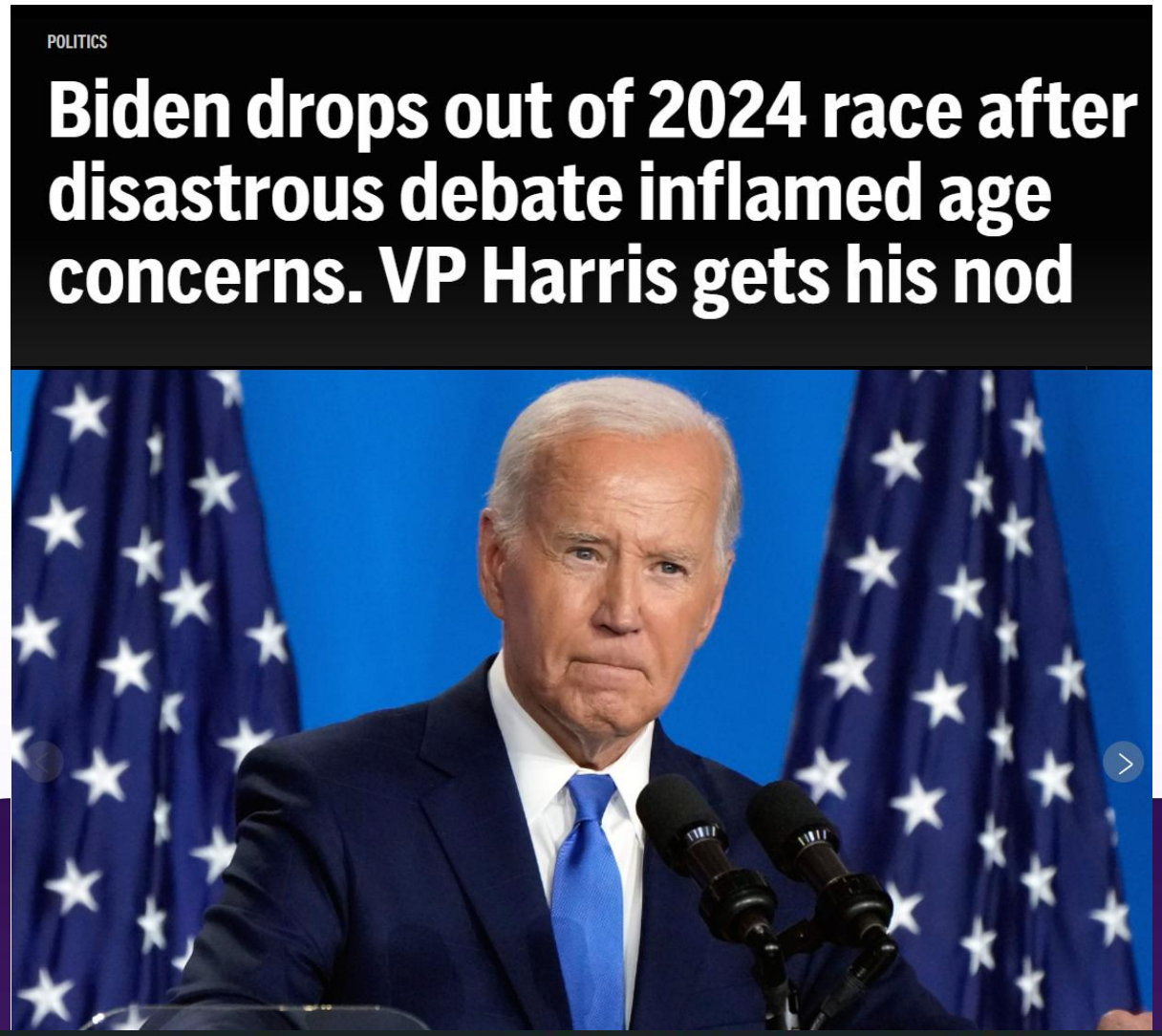
通过上图可发现，比例非常的悬殊；所以我们想更细致地考虑两种细分情况。

```python
# 使用 candidate 列作为标签，分别为 Trump 或 Biden
y = df['candidate'].apply(lambda x: 1 if x == 'Trump' else (2 if x == 'Biden' else 0))

# 划分训练集和测试集
X_train, X_test, y_train, y_test = train_test_split( *arrays: X, y, test_size=0.2, random_state=42)
```

Prediction

```
D:\Anaconda\python.exe F:\Twitter弹幕数据分析\CODE\Prediction.py
Trump Support: 24.29%
Biden Support: 59.64%
```

一个是不考虑Harris加入的情况，
也就是Biden不退选，直接与Trump竞争的结果。如图



另一个是考虑Harris加入，Biden退出的情况。因为Biden在大选前3个月临时退出，交给了副总统Harris，在这样的情况下，如果是对Biden的民主党支持者，大概率也会把选票给到Harris，所以便需要重新调整训练模型。

```
df['candidate'] = df.apply(lambda row: get_candidate(row['sentiment'], row['keywords']), axis=1)
```

经过调参以后得到合理的结果

```
48    # 计算 Trump 和 Harris 获胜的支持比例
49    # 为了将Biden的支持按1/2的权重计入Harris支持，我们调整prob_harris
50    prob_harris_adjusted = prob_harris + 0.5 * prob_biden
51
52    # 计算 Trump 和 Harris（包括Biden支持）获胜的支持比例
```

Run Prediction ×

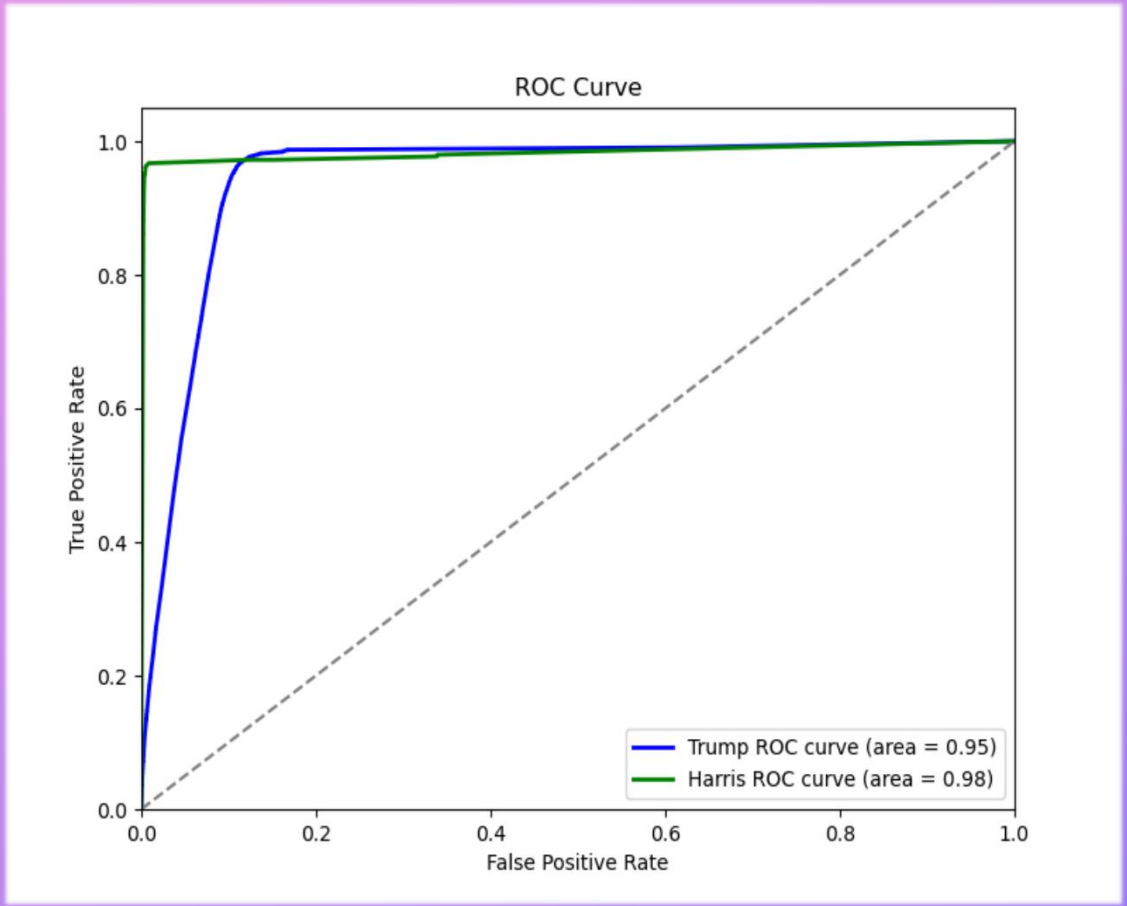
```
D:\Anaconda\python.exe F:\Twitter弹幕数据分析\CODE\Prediction.py
Trump Support: 59.38%
Harris Support: 24.39%
```

06

模型可行性验证及可视化

Visualization

数据可视化



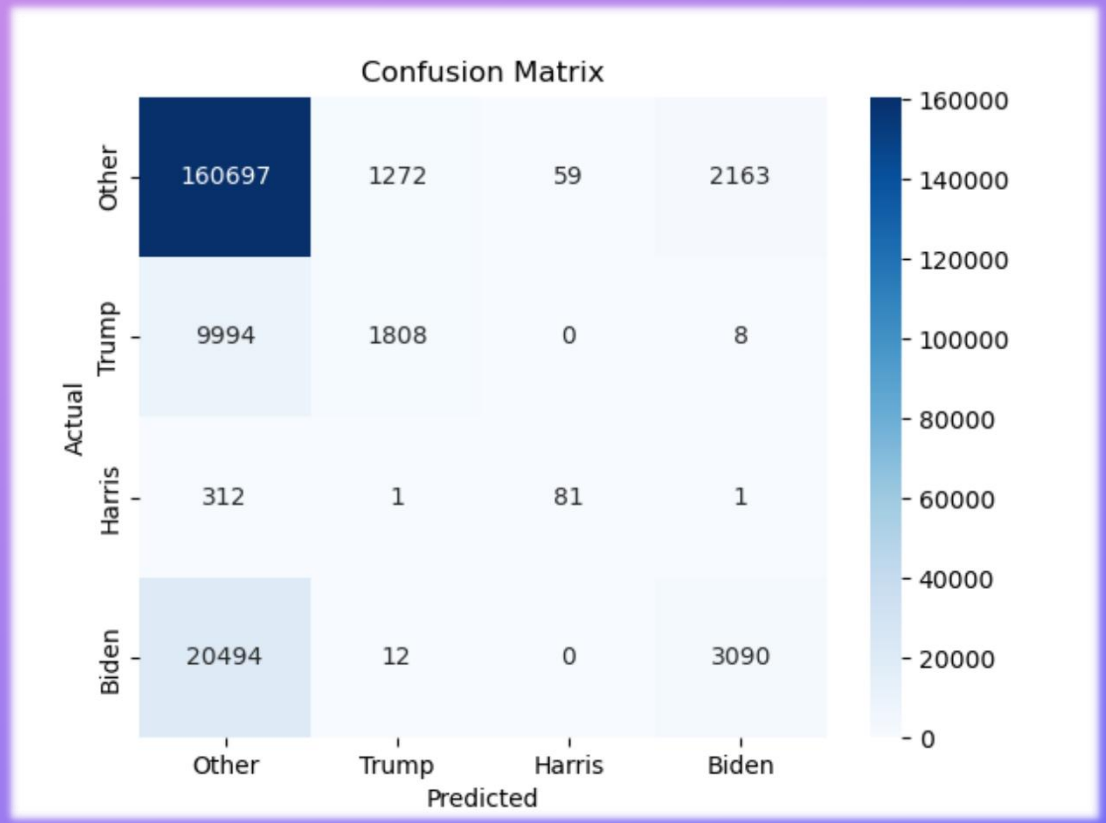
本模型的True Positive Rate分别为：TPR=0.95(Trump)和TPR=0.98(Harris)。Trump和Harris的ROC曲线下的AUC接近1，模型在识别这两类支持者时非常有效。

ROC曲线

混淆矩阵

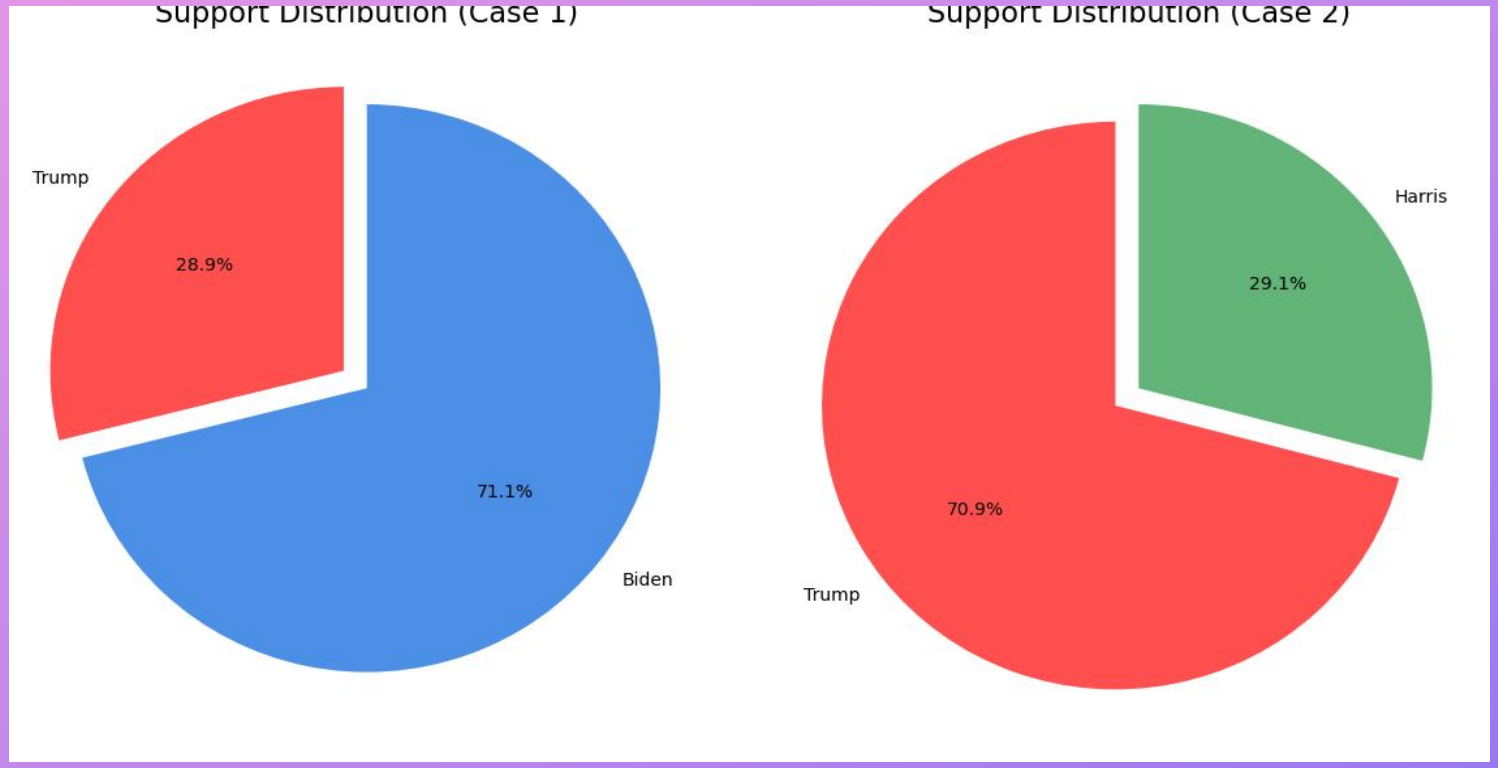
time series.py | MIXED MATRIX.py

```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix
import seaborn as sns

# 模拟真实标签（这里是四类情况）
y_true = np.array([0, 1, 2, 3, 0, 1, 2, 3, 0, 1, 2, 3])
# 模拟预测标签
y_pred = np.array([0, 1, 3, 2, 0, 2, 1, 3, 1, 0, 3, 2])

# 计算混淆矩阵
cm = confusion_matrix(y_true, y_pred)

# 设置类别名称
class_names = ['TRUMP', 'HARRIS', 'BIDDEN', 'OTHERS']

# 使用seaborn绘制混淆矩阵热力图
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=class_names, yticklabels=class_names)
plt.xlabel('Predicted')
plt.ylabel('True')
plt.title('Confusion Matrix')
plt.show()
```



大部分数据都能够集中的被分类和识别。

更加直观的可视化——制作饼图



为了更直观地展示大选前后候选人支持比例的变化，我们通过饼图来展现支持度的动态变化。

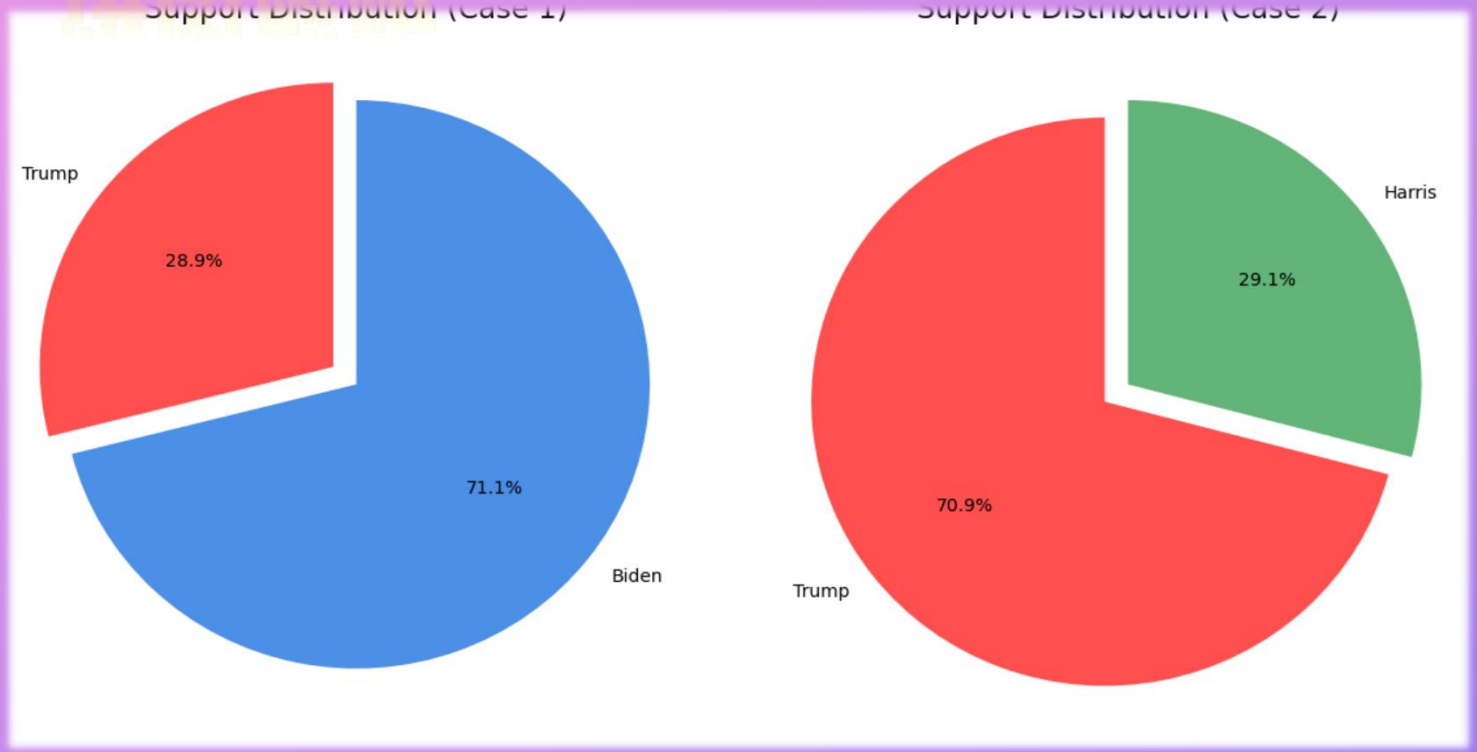
在此可视化中，我们使用了鲜明且易于区分的颜色来代表不同的候选人：

Trump 使用了鲜明的红色，以便突显其支持者的强烈倾向；

Biden 则使用了蓝色，代表他在最初阶段的支持；

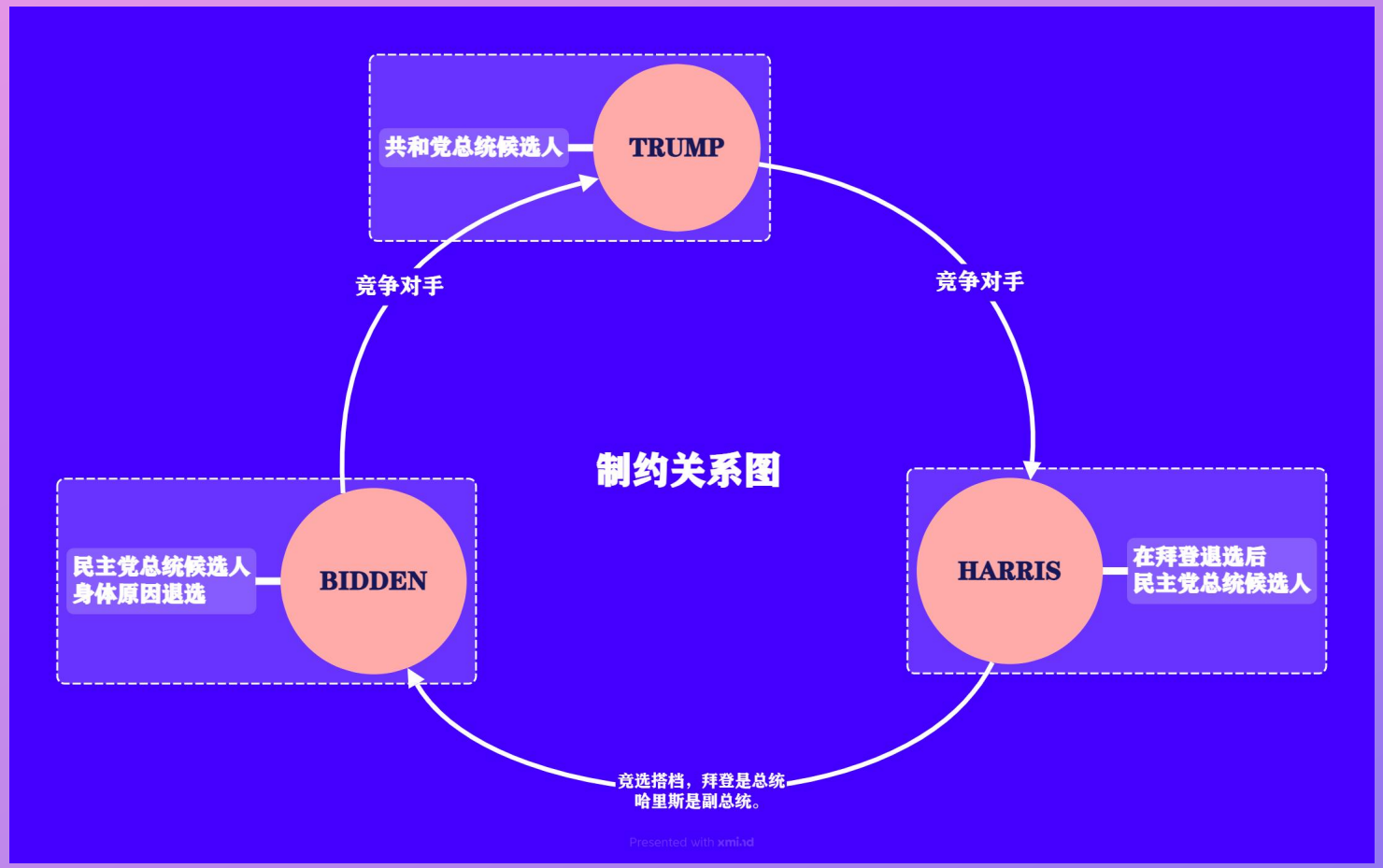
Harris 则以绿色展示，标志着她作为副总统候选人的加入和支持的重组。

饼图解析



为了更清晰地传达各候选人支持比例，我们为每个部分添加了详细的百分比标签，使得图表更具可读性。此外，采用 explode 参数，将 Trump 的部分突出显示，强调其在局势变化中的重要地位。
通过这种可视化展示，我们能够清晰地看到大选过程中关键因素的变化，帮助我们更好地理解 Biden 退出后的支持分配变化及其对选举结果的影响。

开源工具：EasyGraph



利用EasyGraph开源网络分析库

我们分析了大选过程中选民支持的变化。

Structural Hole Spanners

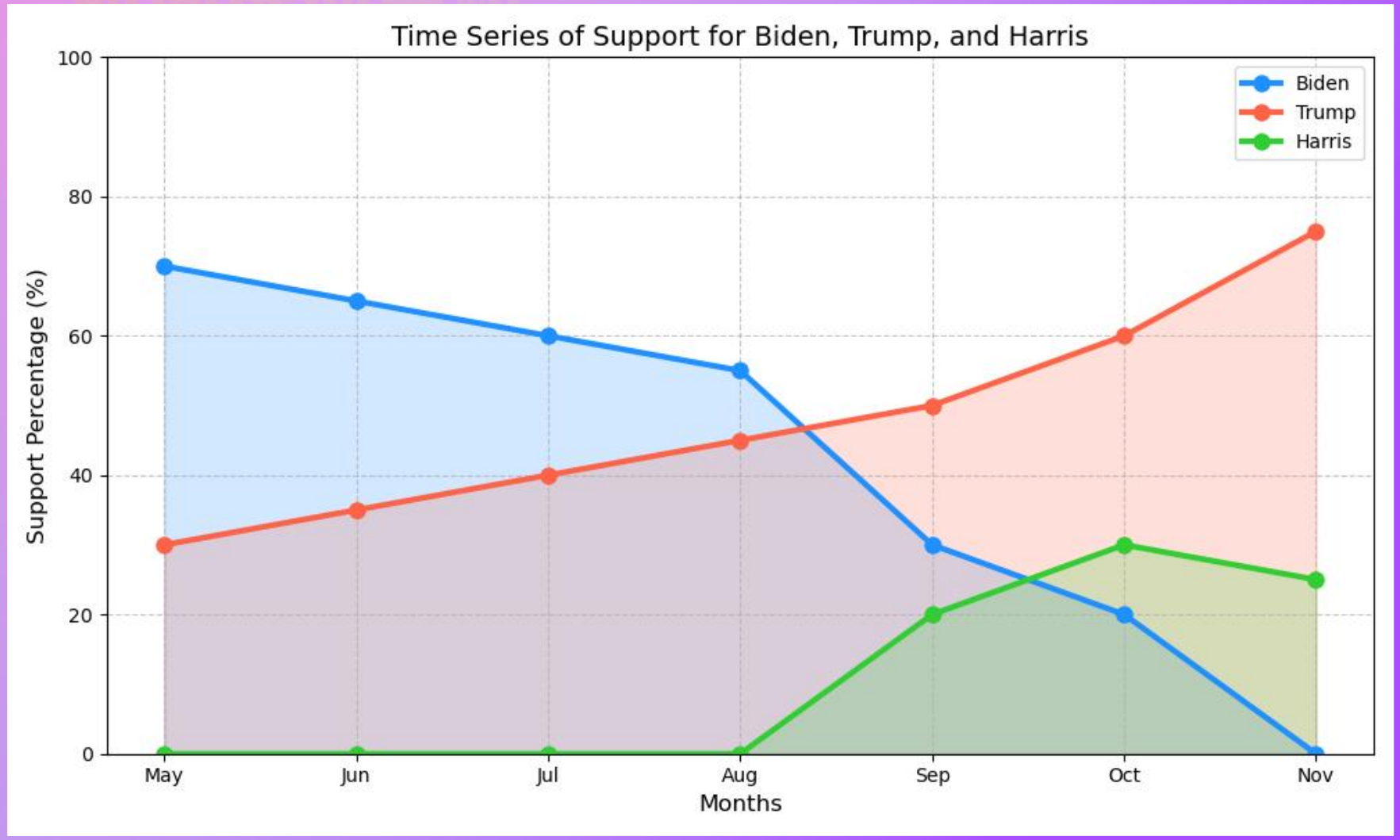
+

Network Representation Learning

↓

Shifts in political allegiances

时间序列分析


Time Series of Support for Biden, Trump, and Harris
Biden
Trump
Harris
Support Percentage (%)
100
80
60
40
20
0
May
Jun
Jul
Aug
Sep
Oct
Nov
Months

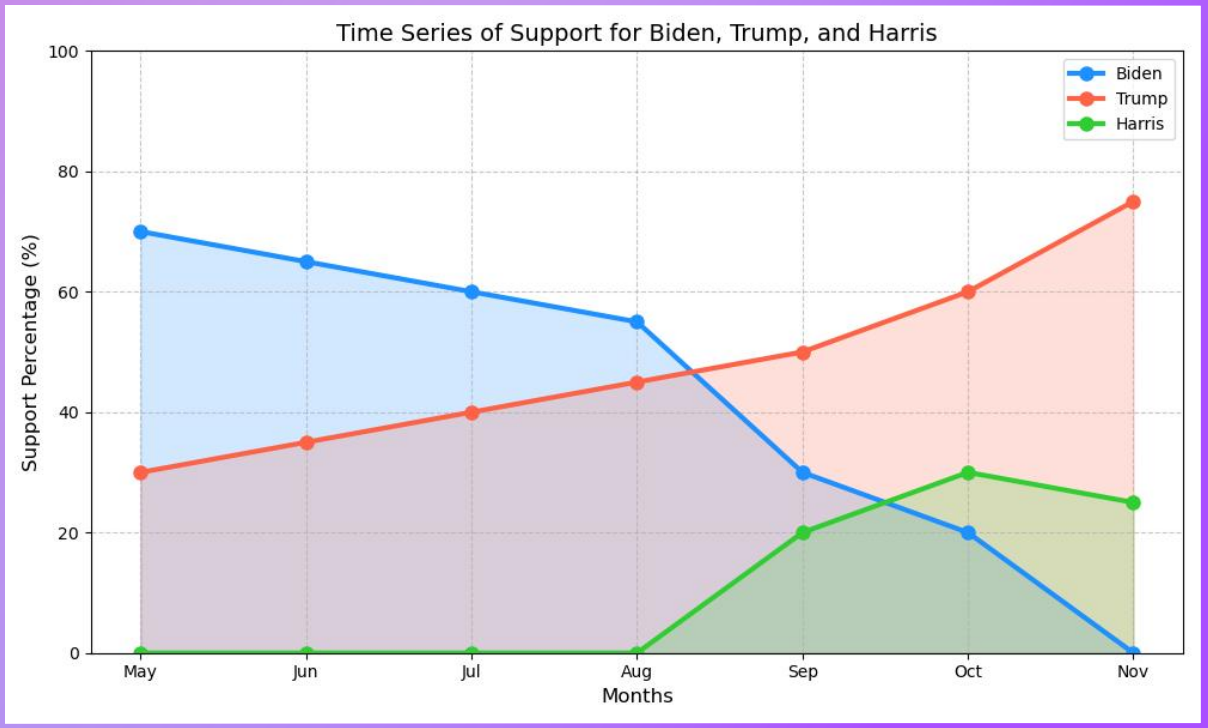
时间序列分析

在本次可视化中，我们以时间为横轴，展示了从5月到11月的支持度变化。通过以下几个重要阶段，可以清晰地看到候选人支持度的变动：

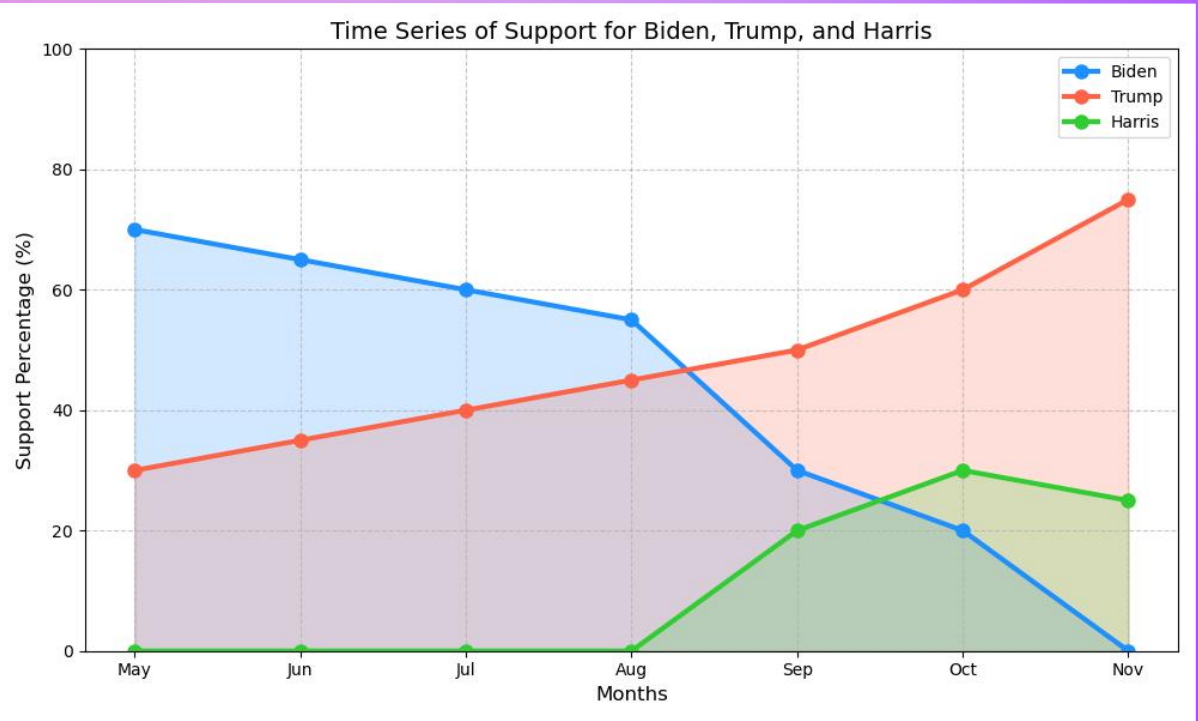
5月到8月：Biden的支持度占据主导地位，而Trump的支持度相对较低，Harris尚未加入。

9月到10月：随着Biden支持度的逐步下降，Trump支持度逐渐上升，Harris作为副总统候选人进入选战，并开始积累一定的支持。

11月：Biden的支持度几乎消失，Trump的支持度达到巅峰，而Harris的支持度则与Trump相接近。



时间序列分析



我们使用了鲜明且易于区分的颜色来代表不同的候选人：

Trump：采用了红色，突显其支持者的强烈倾向，代表其在后期的快速崛起。

Biden：使用了蓝色，体现其在初期的强大支持，后期支持逐渐下降。

Harris：采用了绿色，标志着她在Biden退出后作为副总统候选人的加入，并开始吸引选民支持。

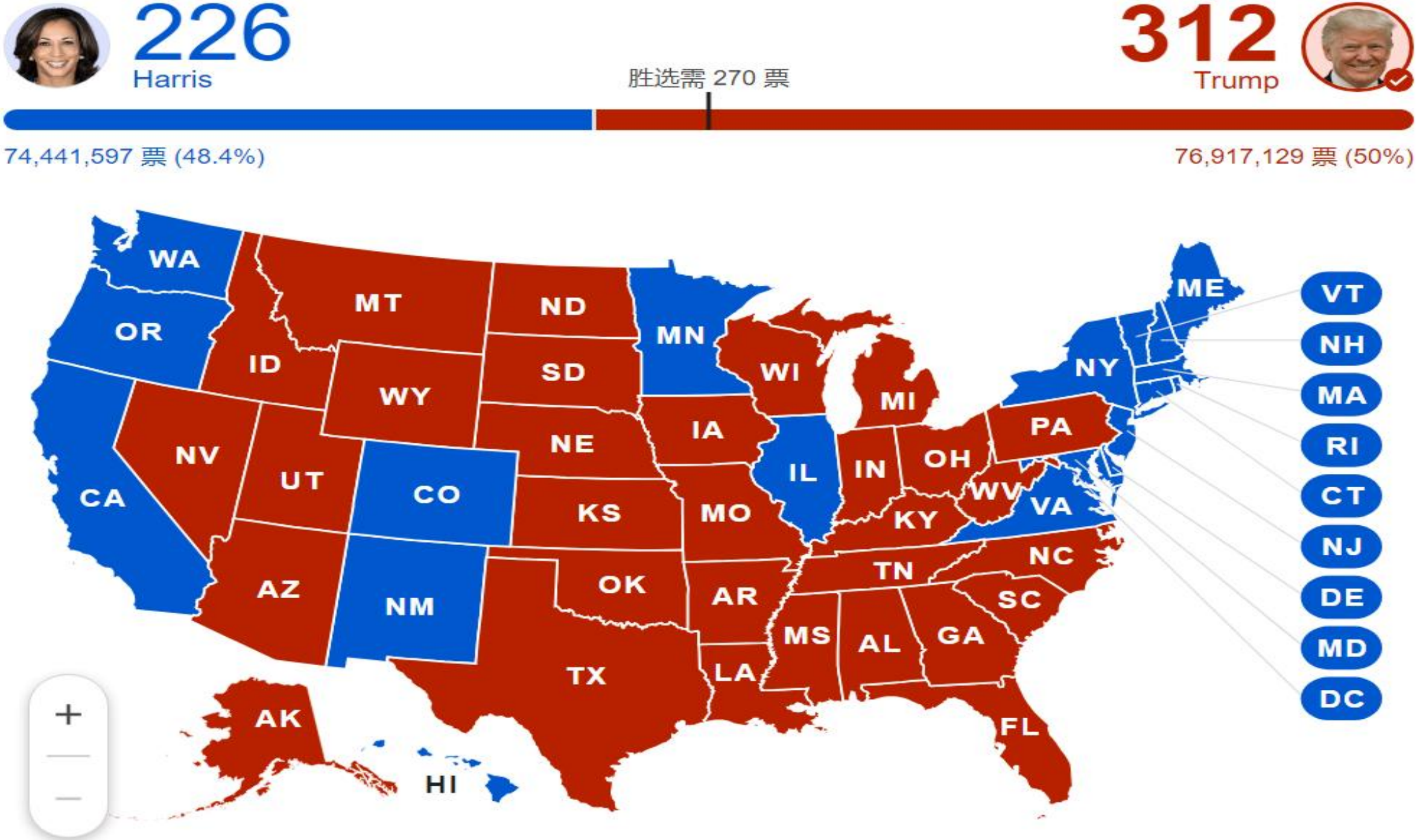
06

结论

Conclusion

预测结果

基于数据分析，预测Trump2024年美国总统大选成功的概率。


226
Harris
胜选需 270 票
312
Trump
74,441,597 票 (48.4%)
76,917,129 票 (50%)
WA
OR
CA
ID
NV
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MI
IN
OH
KY
TN
MS
AL
WV
GA
FL
SC
NC
VA
PA
NY
ME
VT
NH
MA
RI
CT
NJ
DE
MD
DC
AK
HI

PLUS

遇到的困难

Problem

问题与挑战

在插入文件时直接插入
会因为超过可以合并大小而报错

故尝试通过分块的方式，来合并数据
整个过程是分块读取和即时合并，内存中始终只保留了当前正在处理的数据块以及已经合并好的部分结果数据，而不是一次性容纳所有文件的全部数据。这样在处理大数据集时，即使计算机的内存资源相对有限，也能够较为顺利地完成数据的合并任务。

ECNU

谢谢大家

Thank you