

Election Campaign Dynamics

Data gathering, Topic modeling, Timeline

Anahita Poulad

Emre Canbazer

Amir Jamali

Melina Chioutakou

November 4, 2021

OUTLINE

1. Data Gathering
2. Keyword Extraction Approach
3. Some New Results from Topic Modeling
4. Correlation between Topics
5. Future Possible Approaches
6. Project Milestones and Schedule

Data Gathering

TWITTER

◇ From Twitter

```
# Creates a cursor objects and iterates through it to store tweet text and metadata of the tweet in lists
for i in tweepy.Cursor(api.user_timeline, id = candidate_username , tweet_mode = 'extended').items():
    if i.created_at > start_date and i.created_at < end_date:
        #If throws timezone error, try utc.localize(i.created_at)
        name.append(i.user.name)
        tweets.append(i.full_text)
        time.append(i.created_at)
        bio.append(i.user.description)
        follower_count.append(i.user.followers_count)
        fav_count.append(i.favorite_count)
        rt_count.append(i.retweet_count)
        is_quote.append(i.is_quote_status)
        in_reply_to.append(i.in_reply_to_screen_name)
        tweet_id.append(i.id)
        is_retweet.append('retweeted_status' in dir(i))
        temp_mentions = []
        for dicti in i.entities['user_mentions']:
            temp_mentions.append(dicti['screen_name'])
        mentions.append(temp_mentions)
        temp_hashtags = []
        for dicti in i.entities['hashtags']:
            temp_hashtags.append(dicti['text'])
        hashtags.append(temp_hashtags)
for tweet in tweets:
    de_emojized = emoji.demojize(tweet.replace(':', ' '))
    emojis = re.findall(r'(:[^\:]*:)', de_emojized)
    emojicons.append(emojis)
```

[updated] script to extract tweets of a candidate starting from a certain date

TWITTER

◇ From Twitter

	name	bio	follower_count	tweet_text	tweet_id	fav_count	rt_count
0	Antoine WAECHTER 2022 🇩🇪 🌐	Candidat à la #Presidentielle2022 .\nMouvement...	138	🇫🇷 La Charte 🇫🇷 du #MEI \n\nLa charte l les-ecol...	1453649613722894336	2	1
1	Antoine WAECHTER 2022 🇩🇪 🌐	Candidat à la #Presidentielle2022 .\nMouvement...	138	https://t.co/AAnlqz5GSF\n\nLa Conférence des N...	1453648613721137153	2	1
2	Antoine WAECHTER 2022 🇩🇪 🌐	Candidat à la #Presidentielle2022 .\nMouvement...	138	RT @DBarths: L'écologie est au cœur des débats...	1453624625028648960	0	1
3	Antoine WAECHTER 2022 🇩🇪 🌐	Candidat à la #Presidentielle2022 .\nMouvement...	138	RT @audreygarric: Le constat de l'@UNEP, à que...	1453254441302429702	0	73
4	Antoine WAECHTER 2022 🇩🇪 🌐	Candidat à la #Presidentielle2022 .\nMouvement...	138	https://t.co/0YpASOO9oO \n🇫🇷 TRIBUNE 🇫🇷\n\nCO2 ...	1452829334939619330	0	0

[updated] pandas dataframe created from the lists

TWITTER

◇ From Twitter

rt_count	in_reply_to	mentions	is_retweet	is_quote	hashtags	emojicons	time
1	None	None	False	False	MEI	:red_circle:, :scroll:	2021-10-28 09:07:30
1	None	COP26, cop21_paris	False	False	waechter2022	:United_Nations:, :leaf_fluttering_in_wind:, ...	2021-10-28 09:03:31
1	None	DBarths	True	False	None	None	2021-10-28 07:28:12
73	None	audreygarric, UNEP	True	False	COP26	None	2021-10-27 06:57:13
0	None	None	False	False	waechter2022	:red_circle:, :newspaper:, :chart_increasing;,...	2021-10-26 02:48:00

(cont.)[updated] pandas dataframe created from the lists

YOUTUBE

◇ From YouTube

```
youtube = build('youtube', 'v3', developerKey= api_key)
#create a YouTube Data API v.3 object
```

```
request = youtube.videos().list(part= "contentDetails,snippet", id = video_id)
response = request.execute()
#make a request to the data of the video and execute it
#the resulting response is a dictionary with lots of metadata of the video
```

```
channel_id = response['items'][0]['snippet']['channelId']
channel_title = response['items'][0]['snippet']['channelTitle']
video_id = response['items'][0]['id']
video_title = response['items'][0]['snippet']['title']
video_description = response['items'][0]['snippet']['description']
upload_date = response['items'][0]['snippet']['publishedAt'].replace('T', ' ').replace('Z', '')
duration = response['items'][0]['contentDetails']['duration'][2:].replace('M', ':').replace('S', '')
# query the response to assign data to corresponding variable
```

```
srt = YouTubeTranscriptApi.get_transcript(video_id, languages = ['fr'])
# get the captions data
```

[updated] script to extract YT captions using yt-transcript-api package and YouTube Data API v3

DATA

	candidate	content_type	channel_id	channel_title	video_id	video_title
0	melenchon	interview	Uck- _PEY3iC6DIGJKuoEe9bw	JEAN-LUC MÉLENCHON	Rd_GGFHi- x0	Présidentielle 2022 : «Je propose ma candidature»
1	melenchon	interview	Uck- _PEY3iC6DIGJKuoEe9bw	JEAN-LUC MÉLENCHON	Rd_GGFHi- x0	Présidentielle 2022 : «Je propose ma candidature»
2	melenchon	interview	Uck- _PEY3iC6DIGJKuoEe9bw	JEAN-LUC MÉLENCHON	Rd_GGFHi- x0	Présidentielle 2022 : «Je propose ma candidature»
3	melenchon	interview	Uck- _PEY3iC6DIGJKuoEe9bw	JEAN-LUC MÉLENCHON	Rd_GGFHi- x0	Présidentielle 2022 : «Je propose ma candidature»
4	melenchon	interview	Uck- _PEY3iC6DIGJKuoEe9bw	JEAN-LUC MÉLENCHON	Rd_GGFHi- x0	Présidentielle 2022 : «Je propose ma candidature»

pandas DataFrame containing the text of the interview split by speaker

DATA

video_description	upload_date	duration	interviewer_questions	candidate_answers
« Oui, je suis prêt : je propose ma candidatur...	2020-11-09 09:33:19	6:22	Bonsoir Jean-Luc Mélenchon.	Bonsoir.
« Oui, je suis prêt : je propose ma candidatur...	2020-11-09 09:33:19	6:22	Merci beaucoup d'avoir accepté notre invitatio...	Bon, ils n'ont tiré aucune leçon du précédent ...
« Oui, je suis prêt : je propose ma candidatur...	2020-11-09 09:33:19	6:22	Oui Ça fait des jours que la rumeur de votre c...	Écoutez, quand tout va mal et que ça semble ét...
« Oui, je suis prêt : je propose ma candidatur...	2020-11-09 09:33:19	6:22	Est-ce qu'il y a eu un débat sur votre nom au ...	Non, mais j'ai posé la question parce qu'une d...
« Oui, je suis prêt : je propose ma candidatur...	2020-11-09 09:33:19	6:22	François Ruffin avait dit que ça l'intéressait.	Non mais il y a beaucoup de talents. Attention...

pandas DataFrame containing the text of the interview split by speaker

EUROPRESSE

Europresse is an information database created in 1999 accessible with a subscription (University of Lorraine provides access)

It provides:

- on-site or remote access to information
- archive database (e.g. until 1944 for Le Monde)
- media monitoring and analysis
- information research in millions of documents
- press panoramas, newsletters and personalized reports
- detection of influencers on social media

EUROPRESSE

Bienvenue sur Europresse

EUROPRESSE UNE SOLUTION DE CISION **RECHERCHER** **DOSSIERS** **PUBLICATIONS PDF** [0] English Étudiant

Recherche simple | **Recherche avancée** | Recherche express | Recherche de biographies

TEXT= emmanuel macron presidentialles 2022 Depuis 30 jours Français, France [Q] [X]

Presse Télévision et radio Médias sociaux Etudes et rapports Répertoires et références

50 sur 140 [Download] [Print] [Share] [Email] Pertinence

La Tribune (France)
Sommet Afrique-France: Emmanuel Macron bousculé par des échanges musclés avec la jeunesse africaine
 2023-10-12 • 1928 mots

Marie-France Réveillard, envoyée spéciale à Montpellier - Le 8 octobre, Montpellier accueillait le nouveau Sommet Afrique-France qui a réuni près de 3 000 personnes. Sans chefs d'Etat africains, ce rendez-vous résolument tourné vers la jeunesse ...

Aussi paru dans

Le Figaro

Tableau de bord

MÉDIAS 199 Presse

TONALITÉ 126 Positif

Presses 100% - 199 documents

Positif 63% - 126 documents

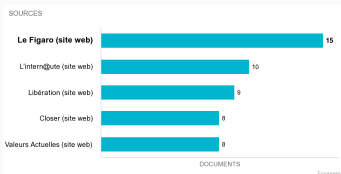
Neutre 0% - 0 documents

Négatif

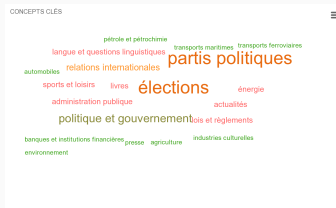
EUROPRESSE



Evolution of media coverage on the topic over the selected period



Most important sources



Terms most frequently associated with the topic

EUROPRESSE

	journal	date	title	author	content
0	Le-Monde- (site-web)	2021- 11-04	Dans-la-fabrique- opaque-des-sondages	Luc- Bronner	Ces six dernières semaines, j'ai répondu à plus de 200 sondages. Enfin, soyons précis, ce sont mes avatars qui ont été sollicités et qui ont répondu aux enquêtes proposées par les plus prestigieux instituts opérant en France : Ipsos, IFOP, Kantar, BVA, OpinionWay, Harris Interactive, GfK... Pour l'un, je m'appelle Ludvine. Pour un autre, Karim, Louis ou Géraldine. Selon les cas, j'ai 19 ans, 26 ans, 47 ans, 73 ans. J'ai déclaré habiter aussi bien la banlieue parisienne que la Bretagne, la région ...
1	La-Provence-- AUBAG	2021- 11-04	Transports--le-milliard- de-l'Etat-prend-forme	NaN	C'était le 2 septembre dernier, au Pharo. Dans son discours Marseille en grand, Emmanuel Macron annonçait pour les transports métropolitains le financement d'un milliard d'euros, dont 250 millions en subventions. Deux mois plus tard, l'examen du projet de loi de finances pour 2022 à l'Assemblée vient concrétiser les promesses présidentielles : alors que le gouvernement a déposé ce lundi un amendement visant à rehausser la subvention de l'État à l'Agence de financement des infrastructures de tr...
2	Challenges	2021- 11-04	La-vaïse-des-thèmes- de-campagne-bat-son- plein	Nicolas- Domenach	C'est la bataille du champ de bataille : la lutte pour imposer son ou ses thèmes de campagne. Percer. Les amener sans se les faire dérober. Chaque candidat(e) sait que la victoire ne dépend pas de mesurette, mais du projet qu'il ou elle incarne. Le waoah qui entre en résonance avec les Français, leurs préoccupations quotidiennes, mais aussi leur imaginaire. La meilleure stratégie de campagne, c'est poser la question dont on a la réponse, observe Gilles Finchelstein, directeur général de ...
3	L'Intern@ute- (site-web)	2021- 11-04	DIRECT-Coronavirus- en-France--ça-ne- ressemble-pas-à-une- cinquième-vague	NaN	CORONAVIRUS. Les derniers chiffres du coronavirus en France sont en forte hausse. S'ils sont à prendre avec précaution, un professeur du CHU de Rennes tempère sur la reprise épidémique, jeudi 4 novembre 2021. L'essentiel Ça ne ressemble pas du tout à un début de cinquième vague. Prudence mais pas d'alarmisme pour le Professeur Pierre Tattevin, du CHU de Rennes. Dans une interview à Ouest-France jeudi 4 novembre 2021, le spécialiste estime que cette remontée du Covid-19 est un peu préoccupant...
4	Dordogne- Libre	2021- 11-02	Éric-Ciotti-en- campagne-pour-la- primaire-de-LR	NaN	Marine PETIT m.petit@dordogne.com Après Montauban et Toulouse samedi après-midi, Éric Ciotti a terminé sa journée par une réunion publique face aux militants périgourdins des Républicains. L'objectif : défendre ses idées et convaincre pour espérer remporter la primaire de la droite dont l'élection est prévue entre le 1er et le 4 décembre prochain lors du congrès des Républicains. Une première étape dans la course à la présidentielle de 2022. Pour le candidat, actuellement député LR des ...

Keyword Extraction Approach

SOME UNSUCCESSFUL EXPERIMENTS

Motivation: Extracting **keyphrases** instead of a number of related words

- TextRank
- Multi-word keyword Scoring
- Word Attraction Rank
- Rake

Future Experiments:

- YAKE
- KeyBERT

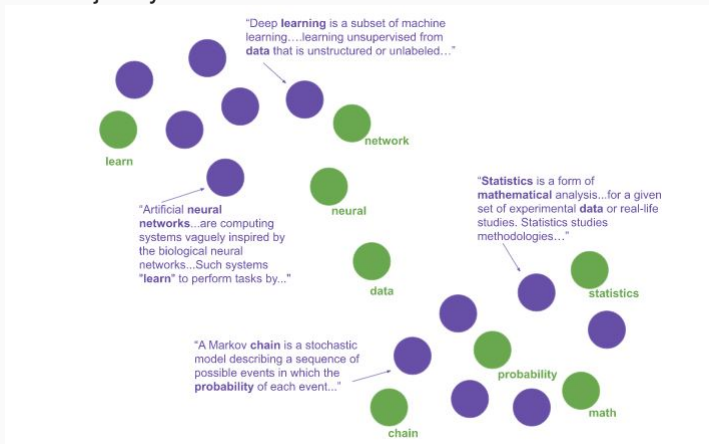
Some New Results from Topic Modeling

TOP2VEC

- It projects documents and topic word to the same space
- It uses clustering to find the most relevant topic for documents and leaves the outlier docs
- We get the sub-topic word set
- We can retrieve documents based on topics

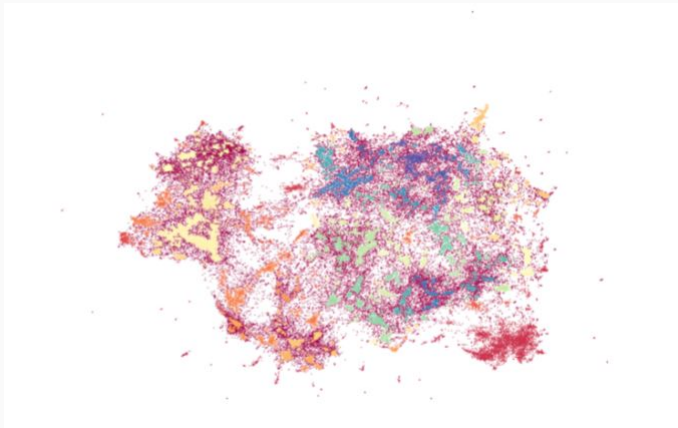
TOP2VEC

jointly embedded document and word vectors



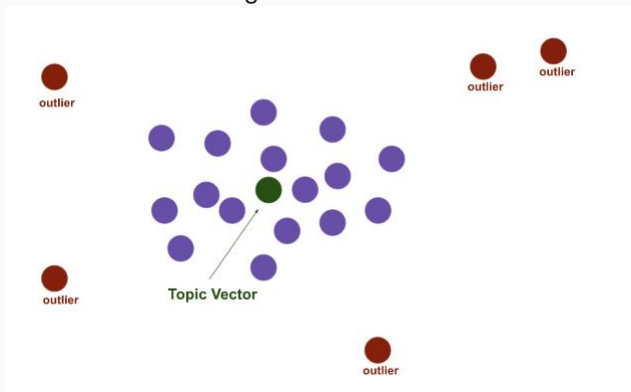
TOP2VEC

dense areas of documents using HDBSCAN



TOP2VEC

For each dense area calculate the centroid of document vectors in original dimension



Correlation between Topics

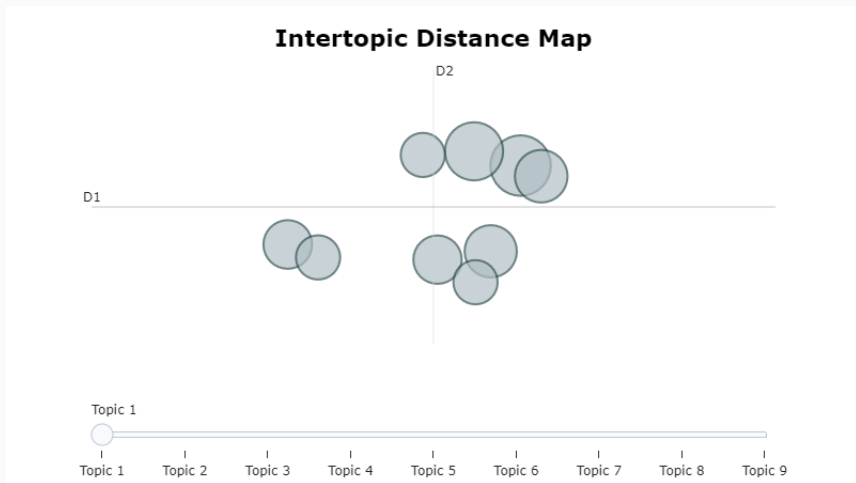
MOST FREQUENT TOPICS

More than 6000 tweets from all candidates have been analysed and more than 94 topics have been found:

Topic Word Scores

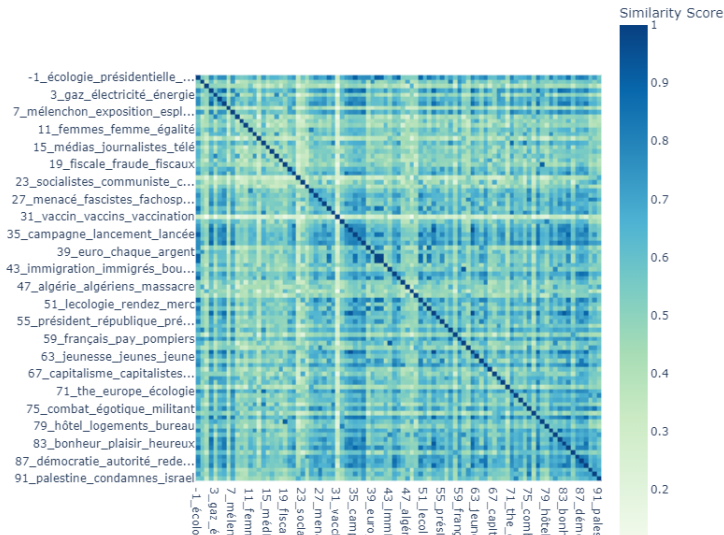


INTERTOPIC DISTANCE MAP

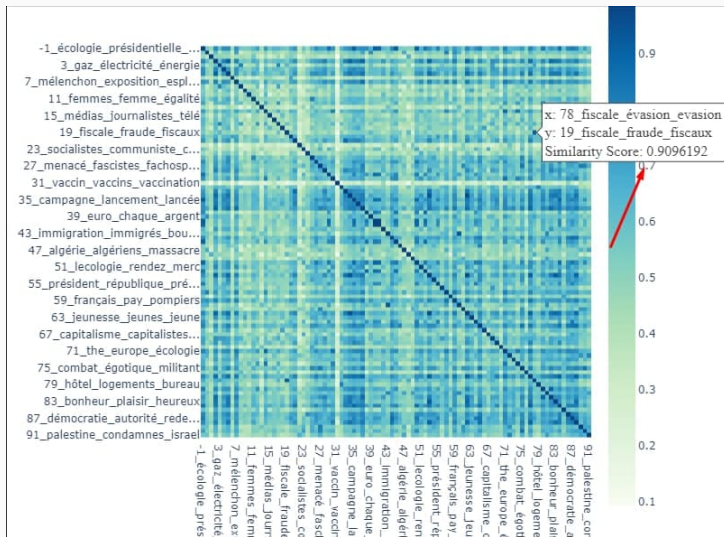


SIMILARITY MATRIX

Similarity Matrix



GETTING IDEA FOR FINDING CORRELATION



Future Possible Approaches

IRAMUTEQ- MAYBE USEFUL FOR EVALUATION

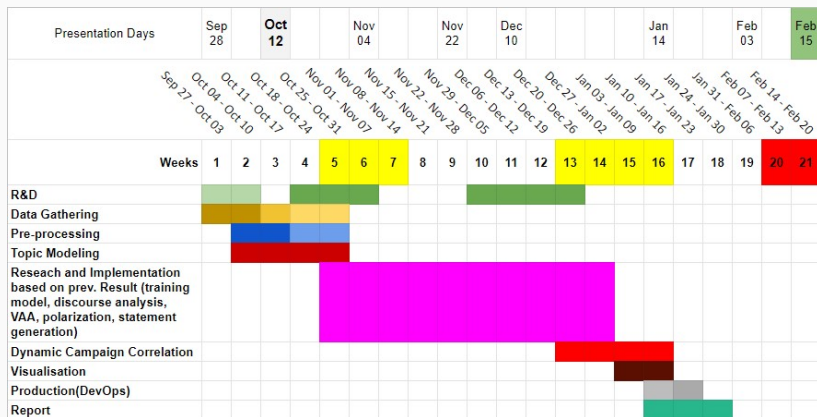
IRAMUTEQ is a licensed software based on R and Python that provides users with different text analyses such as:

- basic lexicography related to lemmatization and word frequency
- descending hierarchical classification
- post- hoc correspondence factor analysis
- similarity analysis

The vocabulary distribution is presented in a comprehensive and clear way with graphical representations

Project Milestones and Schedule

TIMELINE



GITHUB REPO

The screenshot shows the GitHub repository page for 'amirpaia / election-campaign-dynamics'. The repository is public and has 27 commits. The main branch is 'main'. The repository contains several files and folders, including 'Data Gathering', 'Topic Modeling', 'articles', 'preprocessing', 'presentations', 'LICENSE', and 'README.md'.

Navigation links: Search or jump to..., Pull requests, Issues, Marketplace, Explore.

Repository: amirpaia / election-campaign-dynamics (Public)

Navigation: Code, Issues, Pull requests, Actions, Projects, Wiki, Security, Insights, Settings.

Branches: main (selected), 1 branch, 0 tags. Buttons: Go to file, Add file, Code.

File/Folder	Description	Commit Date
amirpaia second presentation file		64adc30 21 minutes ago 27 commits
Data Gathering	Scripts for data collection.	3 days ago
Topic Modeling	Create bertopic.ipynb	3 days ago
articles	the standard structure of the project	15 hours ago
preprocessing	Update preprocessing_tweets.ipynb	3 days ago
presentations	second presentation file	21 minutes ago
LICENSE	Initial commit	14 days ago
README.md	Update README.md	14 days ago

Thank you!