DATA SCIENCE - NLP

REDDIT API SCRAPING AND PREDICTION MODELING

JAMES BABYAK

ATX DSI-5

DID YOU KNOW?


EVERYTHING IS BIGGER IN TEXAS


INCLUDING OUR POLITICS


BETO O'ROURKE


RAFAEL "TED" CRUZ

BATTLE ROYALE!
TEXAS SHOWDOWN!

MAYBE...

PROBLEM

EXPLORE STATE OF THE RACE IN THE REDDIT REALM

WHAT CHARACTERISTICS OF A POST ON REDDIT CONTRIBUTE MOST TO WHICH SUBREDDIT IT BELONGS TO?


r/Beto_for_Senate
Search r/Beto_for_Senate
BETO
FOR SENATE
SORT
the discussion
GET STARTED
Posted by u/TrynnaFindaBalance 1 month ago
SA: Reminder about Voter Registration resources and wh
hem in this sub
r/TedCruz
Search r/TedCruz
Cruz
SORT HOT
the dis on
GET ST ED
sted by u/D ungR ubli s ys ag
eto O'Ro rke ie To ee S ene t D nk-Driving
ash Aft ing ome ne oli Rep t Says
dailywire.com/news/3...
2 Comments
Share
Save
Posted by u/trahan94 12 days ago
Poll: Cruz leads O'Rourke by 1 point in Texas Senate race
thehill.com/homene...
GATHER DATA

USE REDDIT API

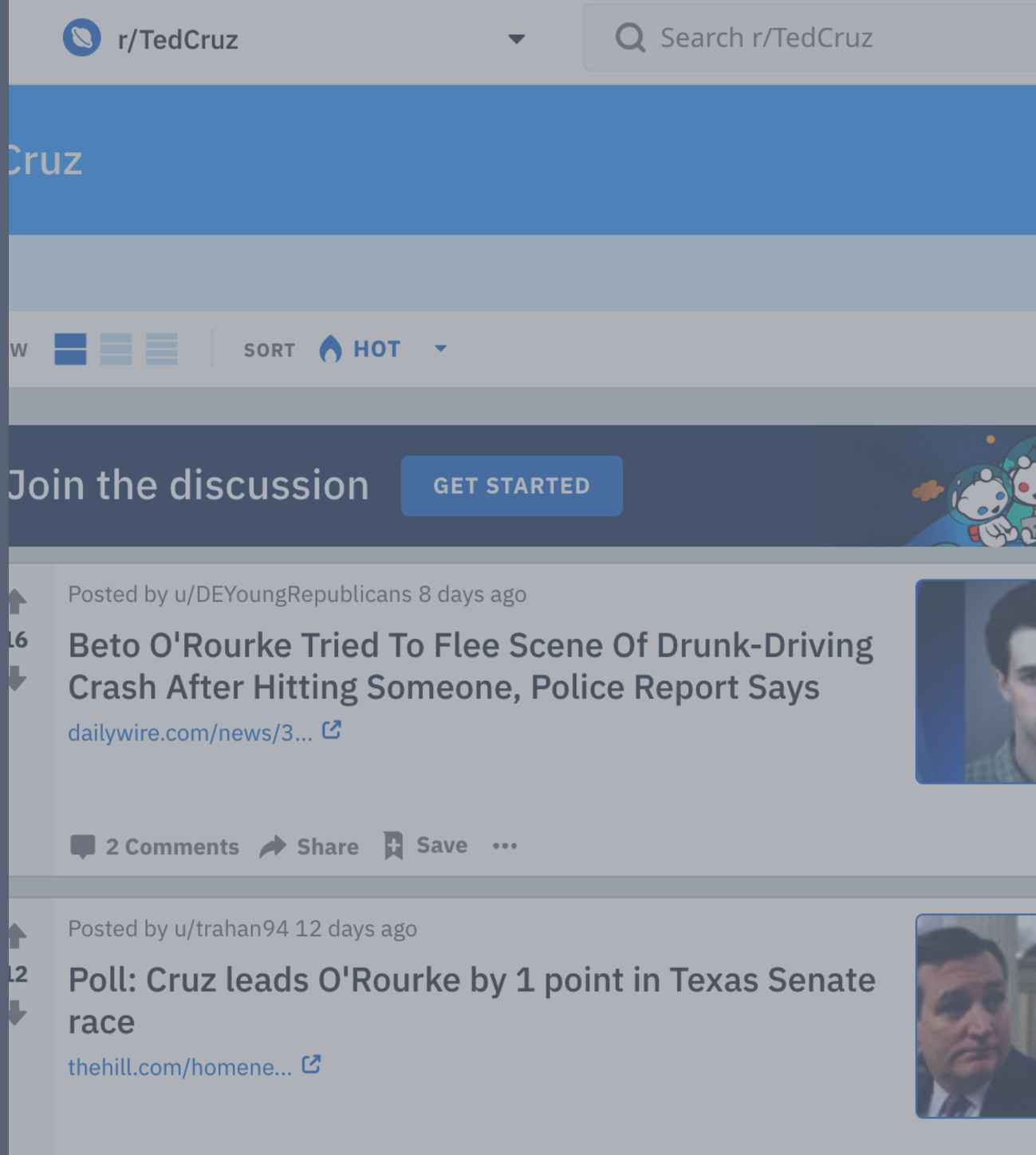
1. BUILD SCRAPE FUNCTION
2. FLEXIBLE TO DIFFERENT SUBREDDIT
3. FLEXIBLE TO AMOUNT OF DATA TO REQUEST

REDDIT SCRAPE FUNCTION

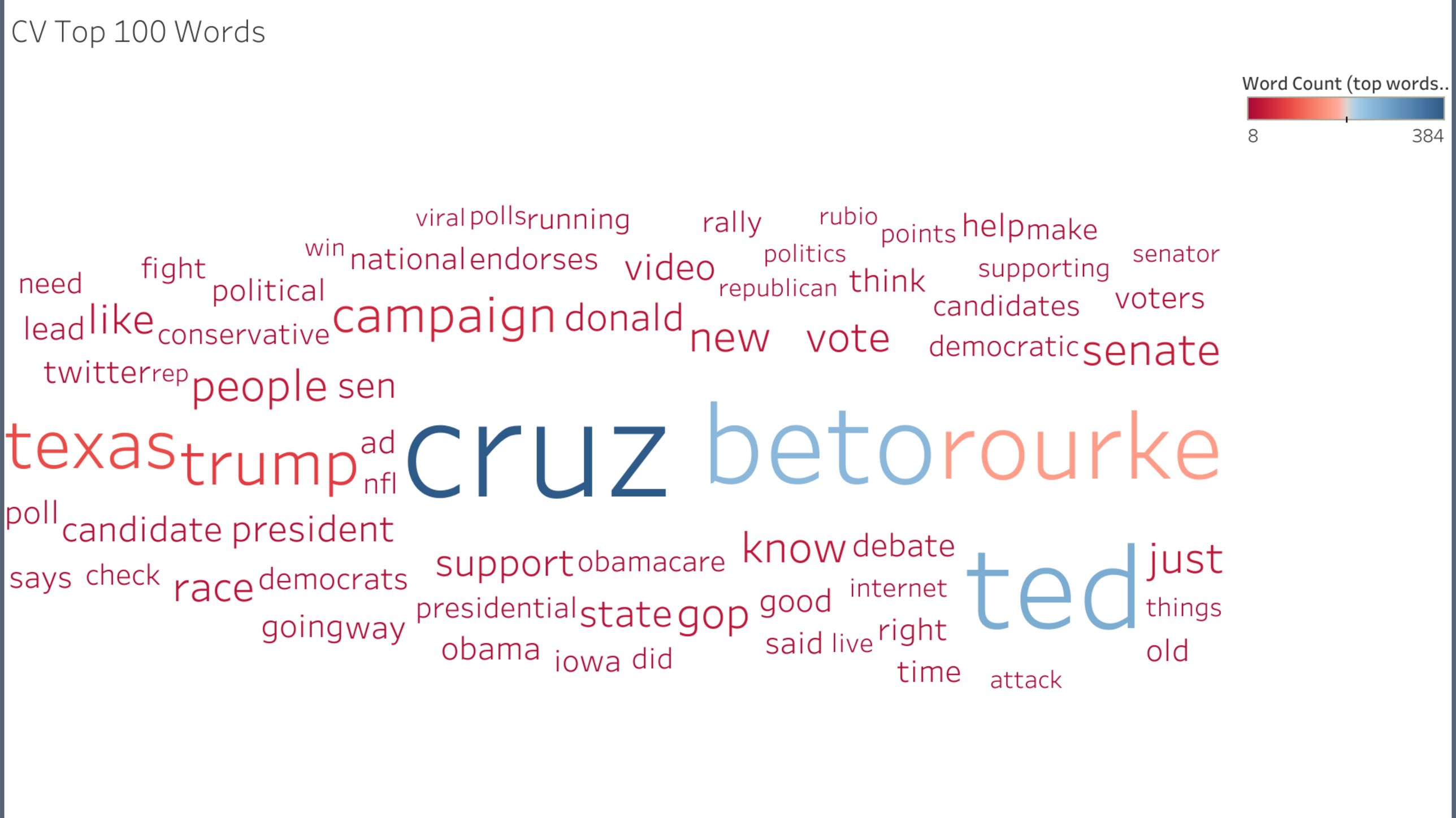
```
def reddit_scraper(subreddit, times):

url = "http://www.reddit.com/r/"+ subreddit +".json"

after = None

for i in range(times):
    posts = []
    if after == None:
        current_url = url
    else:
        current_url = url + '?after=' + after
    print(current_url)
    res = requests.get(current_url, headers={'User-agent': 'GitUpstreamMaster'})
    #print(res.status_code)
    if res.status_code != 200:
        print('Status error: ', res.status_code)
        break
    current_dict = res.json()
    current_posts = [p['data'] for p in current_dict['data']['children']]
    posts.extend(current_posts)
    after = current_dict['data']['after']
    time.sleep(4)
    if i > 0:
        current_df = pd.DataFrame(posts)
        prev_posts = pd.read_csv(subreddit + '.csv')
        all_posts = pd.concat([prev_posts,current_df])
        all_posts.to_csv(subreddit + '.csv', index=False)
    else:
        current_df = pd.DataFrame(posts)
        current_df.to_csv(subreddit + '.csv', index=False)
return all_posts.shape
```


```
reddit_scraper('Beto_for_Senate', 15)
```

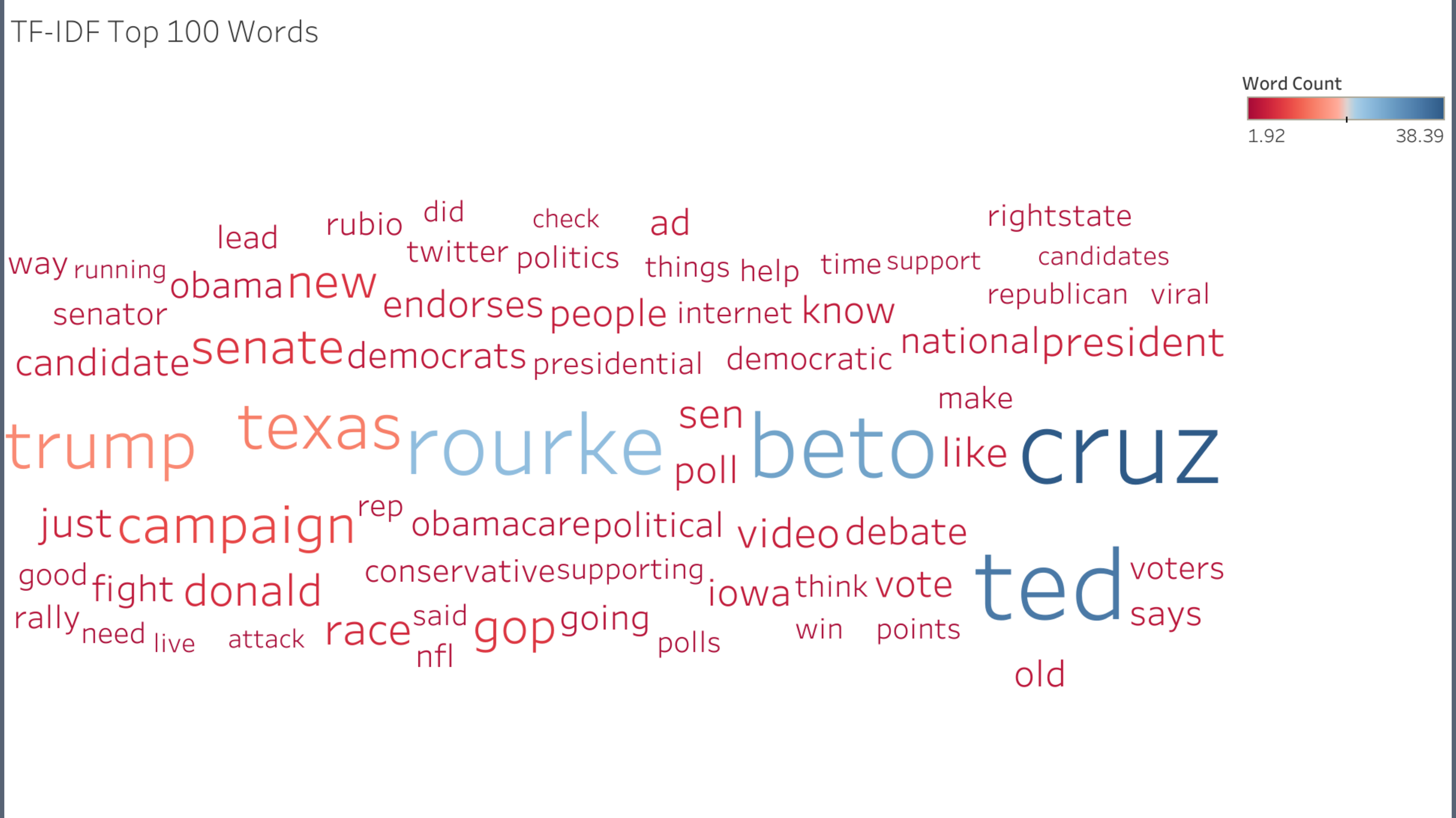
```
reddit_scraper('TedCruz', 15)
```



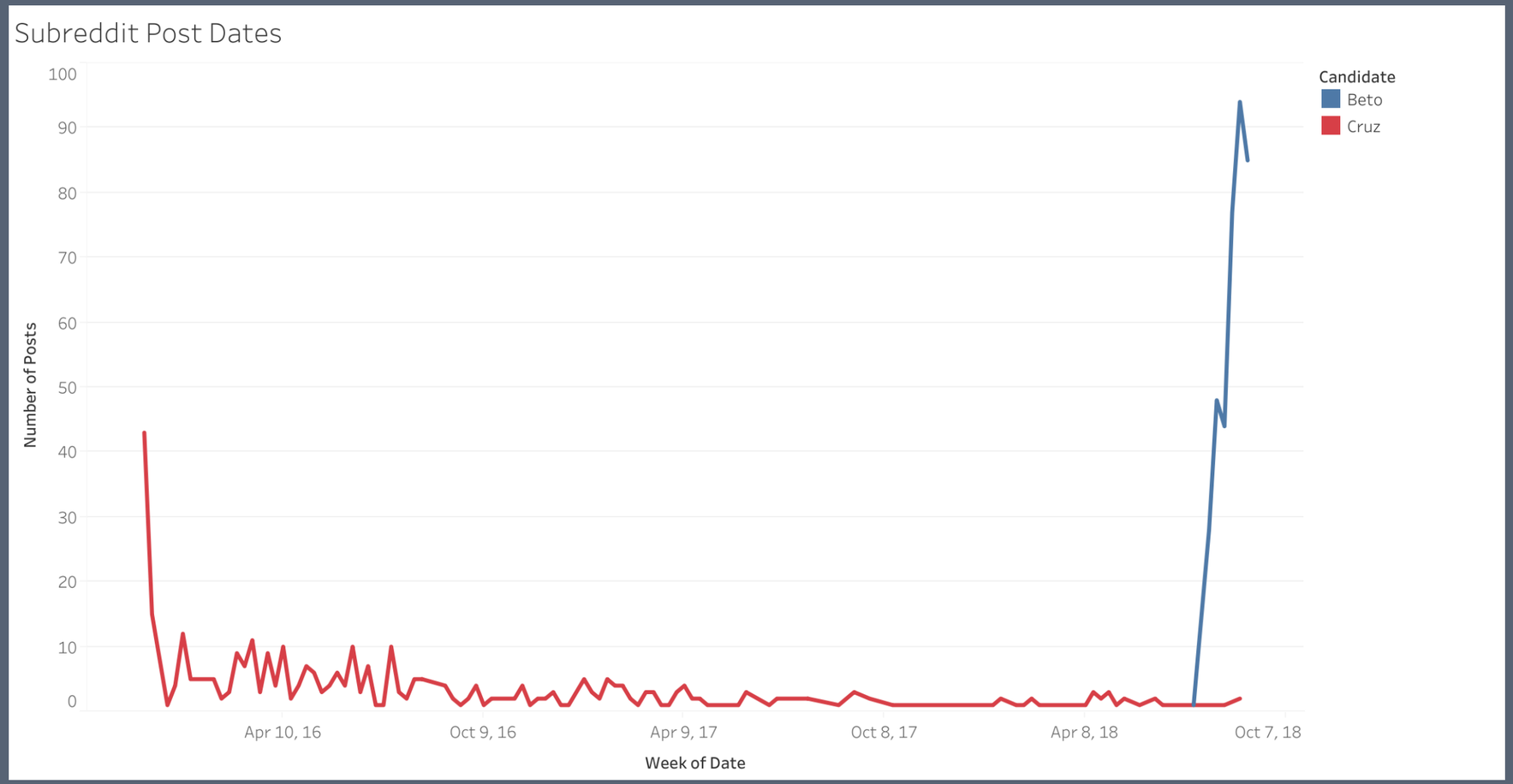
CV Top 100 Words

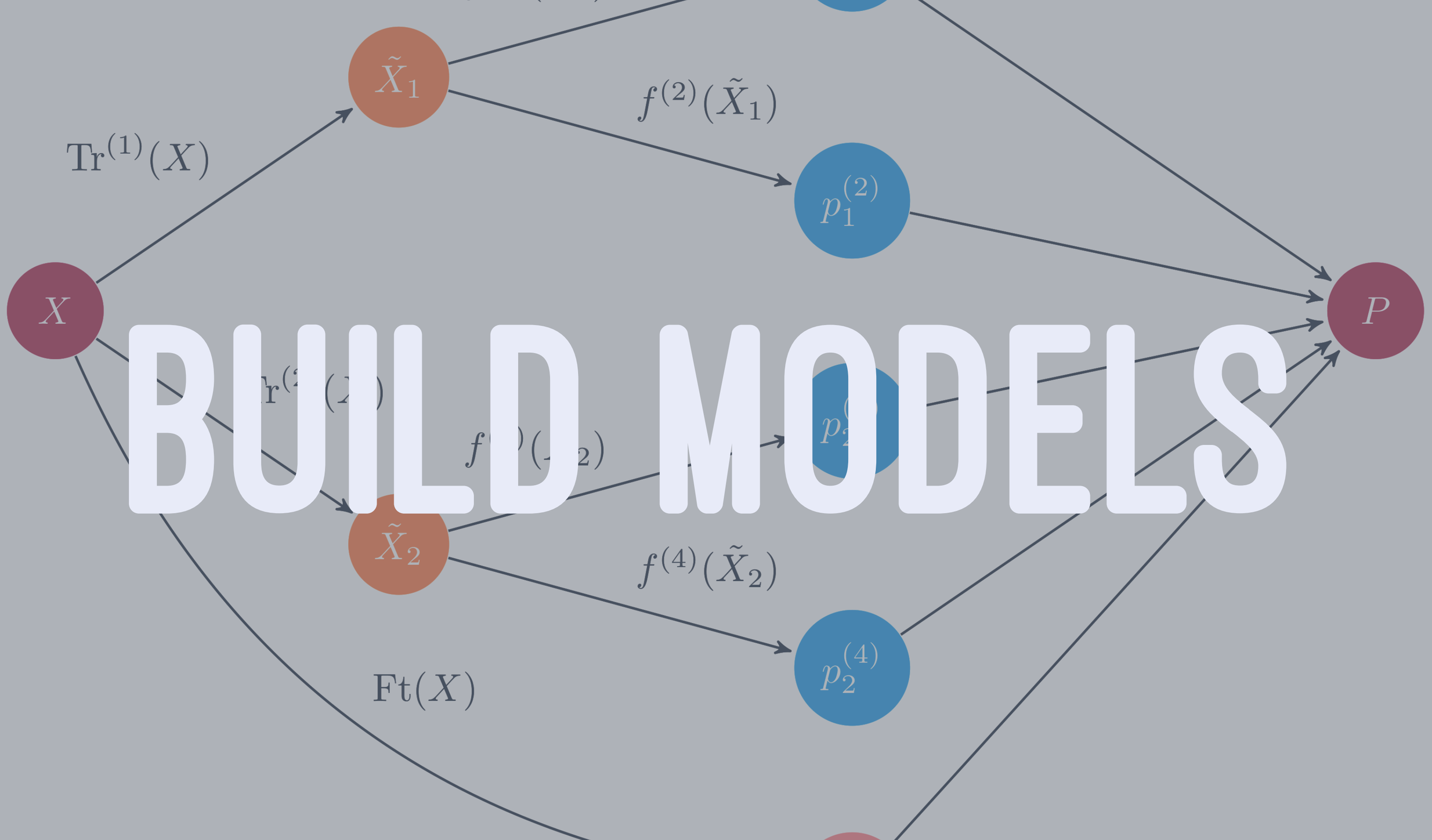


TF-IDF Top 100 Words



POST TIMELINE





BUILD MODELS

BUILD MODELS

1. COMPARE CLASSIFIERS
2. COMPARE VECTORIZERS
3. APPLY GRIDSEARCH
4. ADD TO PIPELINE

OPTIONS

MODELS

- LOGISTIC
- RANDOM FOREST
- EXTRA TREES
- GRADIENT BOOST
- MULTINOMIAL NAIVE BAYES

VECTORIZER

- COUNT VECTORIZER
- TF-IDF VECTORIZER

GRID SEARCH

SCORED BY ACCURACY

MODELS	SCORE	BEST SCORE
LOGISTIC	0.941	0.941
RANDOM FOREST	0.894	0.931
EXTRA TREES	0.904	0.910
GRADIENT BOOST	0.947	0.947
MULTINOMIAL NB	0.943	0.943

GRID SEARCH

SCORED BY ACCURACY

MODELS	SCORE	BEST SCORE
LOGISTIC	0.941	0.941
RANDOM FOREST	0.894	0.931
EXTRA TREES	0.904	0.910
GRADIENT BOOST	**0.947**	**0.947**
MULTINOMIAL NB	0.943	0.943

PIPELINE

MODELS

- LOGISTIC
- RANDOM FOREST
- EXTRA TREES
- GRADIENT BOOST
- MULTINOMIAL NB

VECTORIZER

- COUNT VECTORIZER
- TF-IDF VECTORIZER

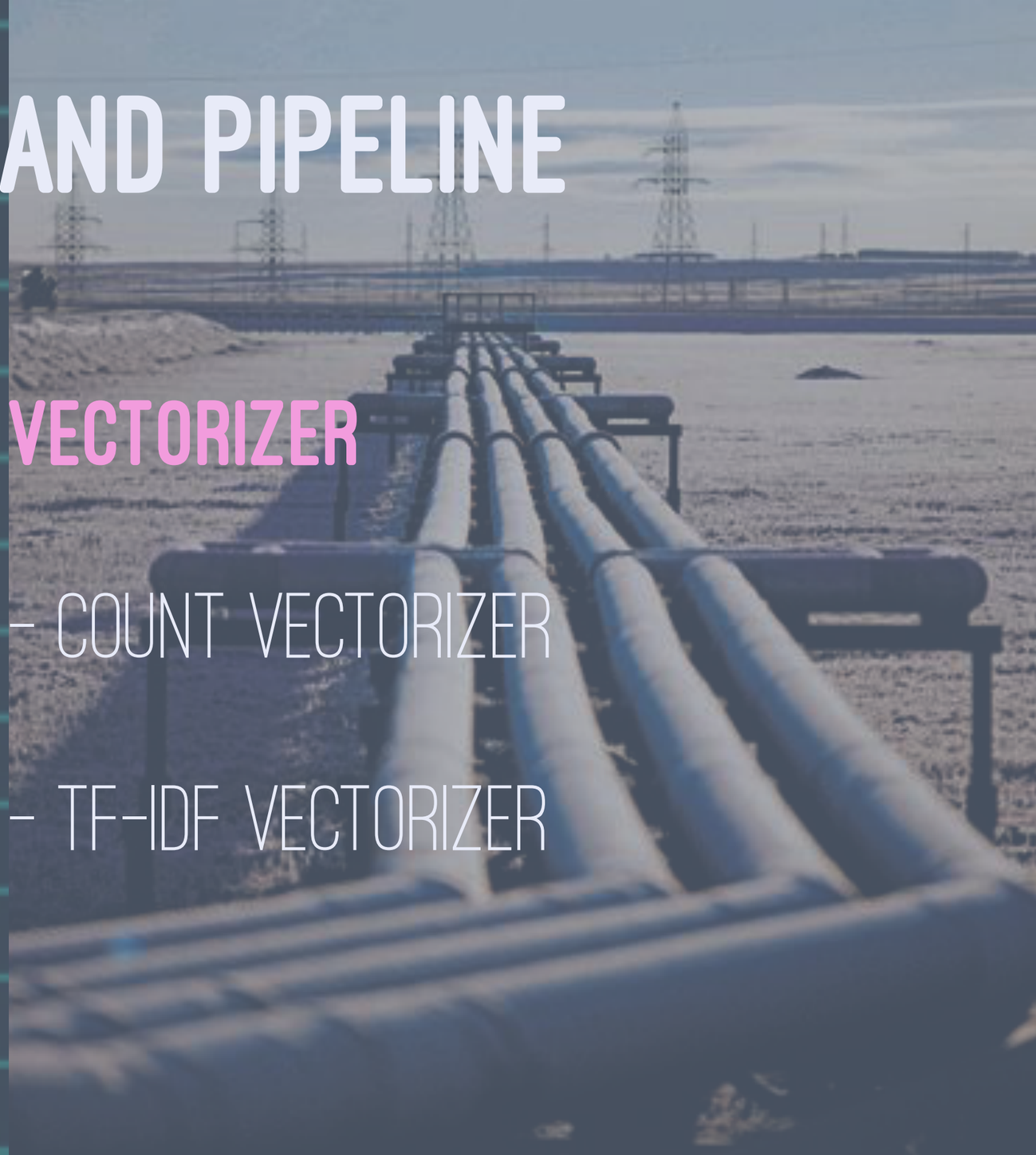
GRIDSEARCH AND PIPELINE

MODELS

- LOGISTIC
- GRADIENT BOOST

VECTORIZER

- COUNT VECTORIZER
- TF-IDF VECTORIZER



GRIDSEARCH AND PIPELINE

MODELS

- **LOGISTIC**
- GRADIENT BOOST

VECTORIZER

- COUNT VECTORIZER
- **TF-IDF VECTORIZER**

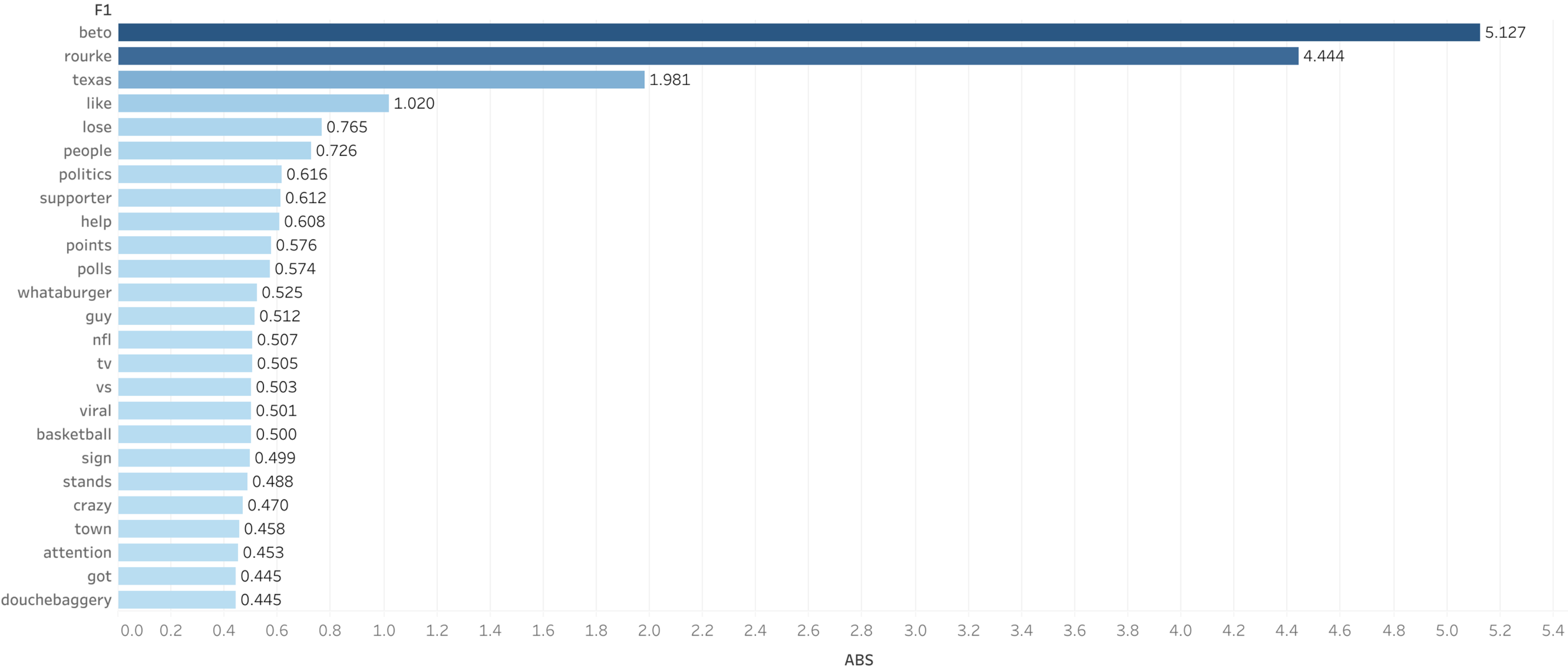
PERFORMANCE

LOGISTIC AND TFIDF

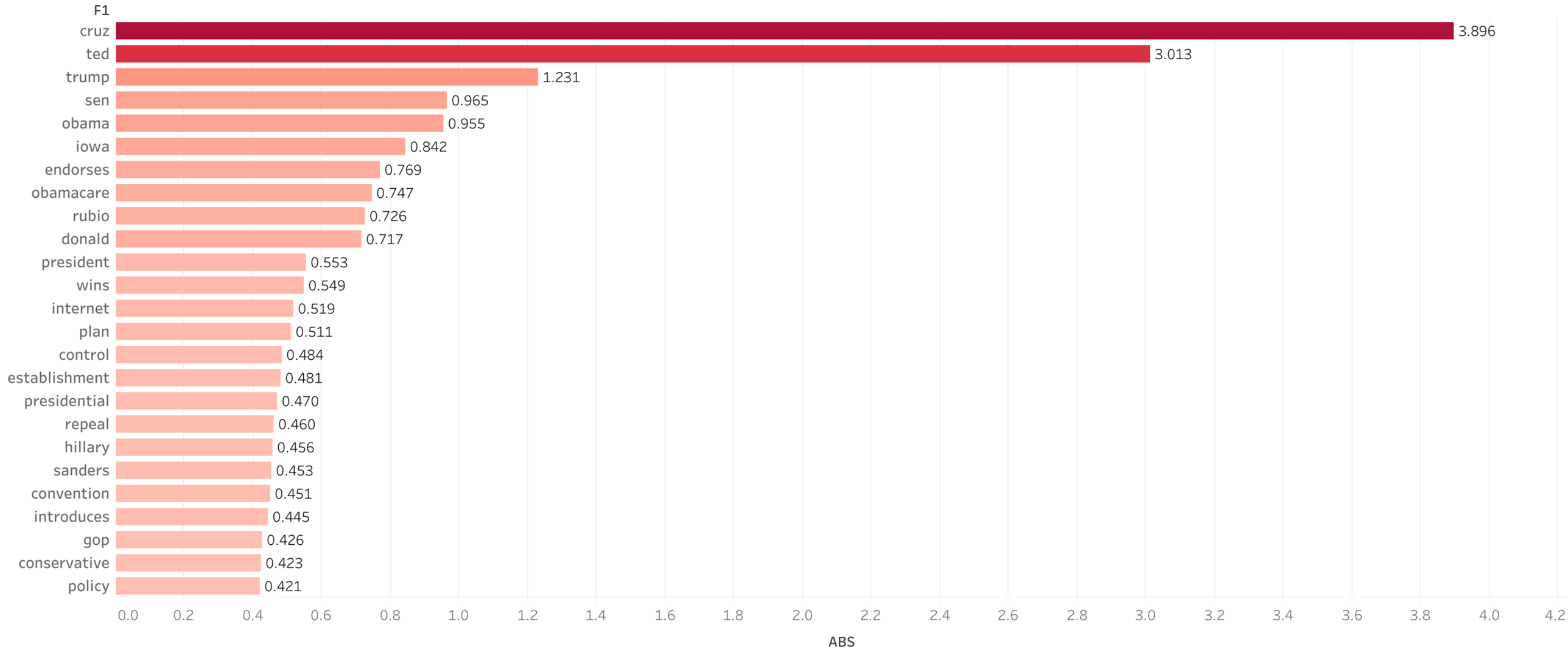
ACCURACY: 0.943

GOOD RIGHT?

Coefficients - Beto



Coefficients - Cruz



MODEL RELIES HEAVILY ON NAMES OF CANDIDATES EVEN WITH TF-IDF VECTORIZER

CHANGES AFFECT ACCURACY

1. REMOVE CANDIDATES NAMES

> 0.729

2. REMOVE 'BETO' AND 'ROURKE'

> 0.835

3. REMOVE 'TED' AND 'CRUZ'

> 0.910

ACCURACY PARADOX

COLLECT MORE DATA: COULD HELP BALANCE THE DATASET

CHANGE YOU METRIC: PRECISION, RECALL OR F1 SCORE

OVERSAMPLE THE DATA: RANDOMLY SAMPLE THE MINORITY CLASS TO CREATE MORE 'FAKE' DATA.

PENALIZED MODEL: BIAS THE MODEL TOWARDS THE MINORITY CLASS.

APPLICATION

TEXAS SUBREDDIT

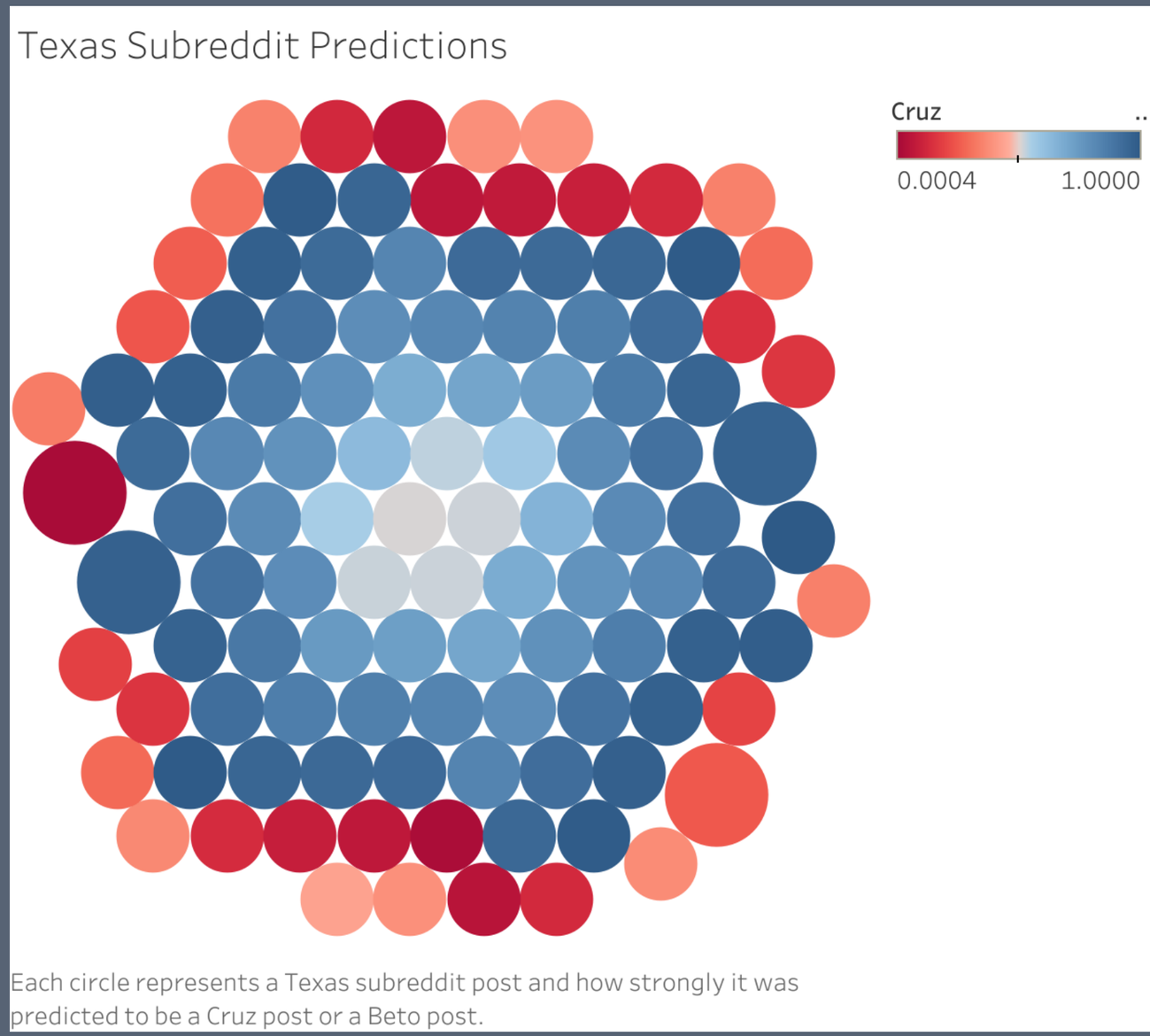
- GO TO A "NEUTRAL" SITE
- EXPLORE THE TEXAS SUBREDDIT
  1. WITH FLAIR TAG "POLITICS"
  1. BY CANDIDATE NAME
- SEE WHAT Y'ALL ARE TALKING ABOUT

968 TOTAL POSTS -> 171 BY NAME

RESULTS

BETO: 69%

CRUZ: 31%



	prediction	P_Beto	P_Cruz	title_text
75	TedCruz	0.000415	0.999585	Ted Cruz: Brainless, Heartless Spineless
89	TedCruz	0.000415	0.999585	Ted Cruz: Servile Puppy
47	TedCruz	0.007553	0.992447	Ted Cruz: Donald Trump Is a 'Pathological Liar'
31	TedCruz	0.039838	0.960162	Ted Cruz: Americans Have No Right To Masturbate
162	TedCruz	0.046096	0.953904	Ted Cruz says it was President Obama who was '...

TAKEAWAYS

TAKEAWAYS

MODELS AND METRICS

- SEEKING BEST SCORE - ACCURACY
- BROAD OPTIONS OF MODEL FINE TUNING
- WHAT DO YOU NEED?
- TASK: FIND ME A CAT


NAILED IT!
100%

TAKEAWAYS

DATA

- LIMITATION OF QUALITY DATA
- POSTS ARE SHORT. NOT TEXT HEAVY
- CRUZ CAMPAIGN NOT HEAVY REDDIT USERS

Tweets 19.4K | Following 7,740 | Followers 3.01M | Likes 1,248

Ted Cruz

@tedcruz

Father of two, @heidiscruz's husband, fighter for liberty. Representing the great state of Texas in the U.S. Senate.

Houston, Texas

https://t.co/dJ4Q3979gw

Joined March 2009

Born on December 22

Tweet to Ted Cruz

Likes

Sexuall Posts 😈 @SexuallPosts · Sep 10


Following
Likes
1,248
Ted Cruz
@tedcruz
Houston, Texas
https://t.co/dJ4Q3979gw
Joined March 2009
Born on December 22
Tweet to Ted Cruz
REAL
SCREENSHOT


THANK YOU
QUESTIONS?