In [42]:

```
from nltk.tokenize import TweetTokenizer
from nltk.corpus import stopwords
from textblob import TextBlob
from googletrans import Translator
from PIL import Image
from wordcloud import WordCloud, ImageColorGenerator
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import nltk
import string
import re
```

In [2]:

```
df = pd.read_csv('log.csv')
df1 = pd.read_csv('log1.csv')
df = df.append(df1, ignore_index=True)
```

In [81]:

```
df.head()
```

	index	date	text	text_token	text_en	words
0	0	2019-09-24 14:26:16+00:00	@Miltonneves "A ONU nao e a organizacao do int...	[onu, nao, organizacao, interesse, global, org...	Miltonneves The UN is not the organization of ...	onu nao organizacao interesse global organizac...
1	1	2019-09-24 14:26:16+00:00	RT @conexaopolitica: Apresentador da Band diz ...	[apresentador, band, diz, vivo, nao, compra, d...	RT conexaopolitica Band host says live that he...	apresentador band diz vivo nao compra discurso...
2	2	2019-09-24 14:26:16+00:00	RT @Bolsoneas: Alguem anotou a placa? Bolsonar...	[alguem, anotou, placa, bolsonaro, atropelou, ...	RT Bolsoneas Did anyone write down the sign Bo...	alguem anotou placa bolsonaro atropelou esquer...
3	3	2019-09-24 14:26:16+00:00	RT @loenxxii: E TA ERRADO PORRA?????? https://...	[ta, errado, porra]	RT loenxxii Is it fucking wrong https t.co/yS8...	ta errado porra
4	4	2019-09-24 14:26:16+00:00	RT @IvanValente: Show do horror e pouco para d...	[show, horror, pouco, descrever, bolsonaro, on...	RT IvanValente Show of horror and little to de...	show horror pouco descrever bolsonaro onu aluc...

In [3]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13272 entries, 0 to 13271
Data columns (total 2 columns):
date    13272 non-null object
text    13272 non-null object
dtypes: object(2)
memory usage: 207.5+ KB
```

In [38]:

```
sw_br = stopwords.words('portuguese')

stemmer = nltk.stem.RSLPStemmer()

# Happy Emoticons
emoticons_happy = set([
    ':-)', ':)', ';)', ':o)', ':]', ':3', ':c)', ':>', '=]', '8)', '=)', ':}',
    ':^)', ':-D', ':D', '8-D', '8D', 'x-D', 'xD', 'X-D', 'XD', '=-D', '=D',
    '=-3', '=3', ':-))', ":'-)", ":')", ':*', ':^*', '>:P', ':-P', ':P', 'X-P',
    'x-p', 'xp', 'XP', ':-p', ':p', '=p', ':-b', ':b', '>:)', '>;)', '>:-)',
    '<3'
])

# Sad Emoticons
emoticons_sad = set([
    ':L', ':-/', '>:/', ':S', '>:[', ':@', ':-(', ':[', ':-||', '=L', ':<',
    ':-[', ':-<', '=\\', '=/', '>:(', ':(', '>.<', ":'-(", ":'(", ':\\', ':-c',
    ':c', ':{', '>:\\', ';('
])

# all emoticons (happy + sad)
emoticons = emoticons_happy.union(emoticons_sad)


def clean_tweets(tweet,to_stem=True):
    # remove stock market tickers like $GE
    tweet = re.sub(r'\$\w*', '', tweet)

    # remove old style retweet text "RT"
    tweet = re.sub(r'^RT[\s]+', '', tweet)

    # remove hyperlinks
    tweet = re.sub(r'https?:\/\/.*[\r\n]*', '', tweet)

    # remove hashtags
    # only removing the hash # sign from the word
    tweet = re.sub(r'#', '', tweet)
    # tokenize tweets
    tokenizer = TweetTokenizer(
        preserve_case=False, strip_handles=True, reduce_len=True)
    tweet_tokens = tokenizer.tokenize(tweet)
    tweets_clean = []
    for word in tweet_tokens:
        if (word not in sw_br and  # remove stopwords
            word not in emoticons and  # remove emoticons
                word not in string.punctuation):  # remove punctuation
            # tweets_clean.append(word)
            if to_stem:
                word = stemmer.stem(word)  # stemming word
            tweets_clean.append(word)

    return tweets_clean
```

Removendo Duplicados

In [5]:

```
df = df.loc[~df.duplicated(subset='text', keep='first')]
df.reset_index(inplace=True)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5573 entries, 0 to 5572
Data columns (total 3 columns):
index    5573 non-null int64
date     5573 non-null object
text     5573 non-null object
dtypes: int64(1), object(2)
memory usage: 130.7+ KB
```

In [6]:

```
df['date'] = pd.to_datetime(df['date'])
```

In [7]:

```
df['text_token'] = df.text.apply(lambda x: clean_tweets(x,False))
df['words'] = df.text_token.apply(lambda x: " ".join(x))
```

In [35]:

```
df.to_csv('twitter_en.csv',index=False)
```

In [36]:

```
final = pd.read_csv('twitter_en.csv')
#removendo indices
final = final.iloc[:,1:]
final.head(1)
```

	date	text	text_token	text_en	words
0	2019-09-24 14:26:16+00:00	@Miltonneves "A ONU nao e a organizacao do int...	['onu', 'nao', 'organizacao', 'interesse', 'gl...	Miltonneves The UN is not the organization of ...	onu nao organizacao interesse global organizac...

WordClouds sobre nossos twitters

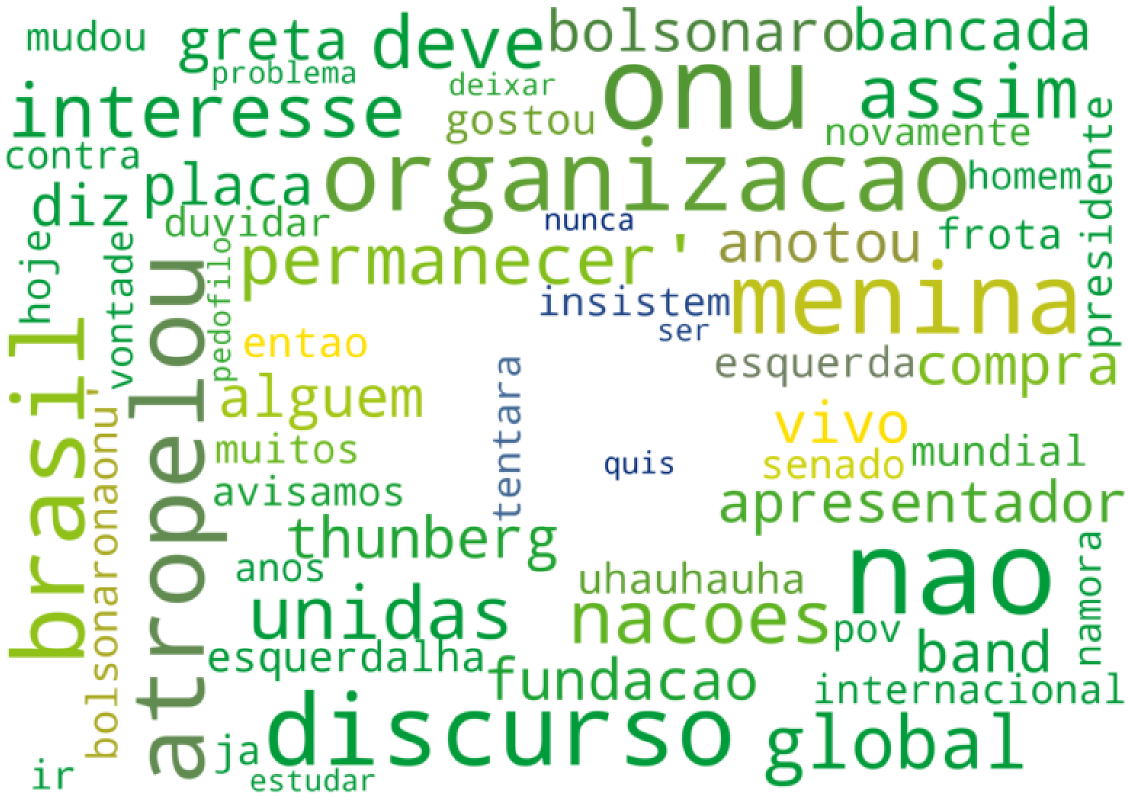
In [41]:

```
mask = np.array(Image.open("brasil.png"))
wordcloud_br = WordCloud(background_color="white",
                          mode="RGBA",
                          max_words=1000, mask=mask).generate(str(final['words'].values))

# create coloring from image
image_colors = ImageColorGenerator(mask)
plt.figure(figsize=[14,13])
plt.imshow(wordcloud_br.recolor(color_func=image_colors), interpolation="bilinear")
plt.axis("off")

# store to file
#plt.savefig("img/us_wine.png", format="png")

plt.show()
```

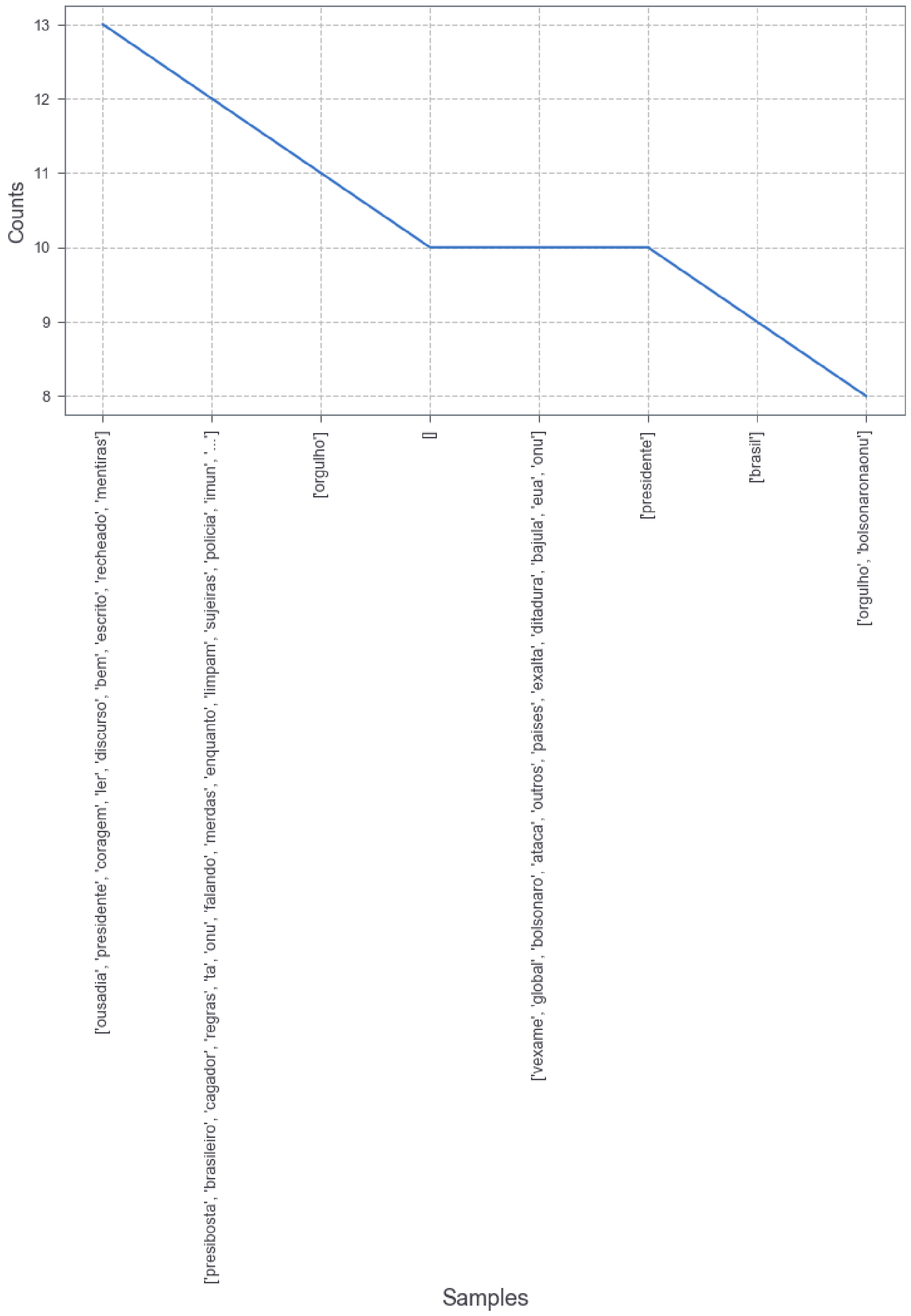


Observando a Frequencia

In [56]:

```
from nltk.probability import FreqDist
fdist = FreqDist(final.text_token)
print(fdist)
plt.figure(figsize=(12,6))
plt.xlabel('Termos', fontsize=18)
plt.ylabel('Counts', fontsize=16)
import matplotlib.pyplot as plt
fdist.plot(8,cumulative=False, )
plt.show()
```

```
<FreqDist with 5273 samples and 5573 outcomes>
```



10 sequencias mais ocorridas

In [80]:

```
for x in fdist.most_common(10):
    print(x[0])
```

```
['ousadia', 'presidente', 'coragem', 'ler', 'discurso', 'bem', 'escrito', 'recheado', 'mentiras']
['presibosta', 'brasileiro', 'cagador', 'regras', 'ta', 'onu', 'falando', 'merdas', 'enquanto', 'limpam', 'sujeiras', 'polici
a', 'imun', '...']
['orgulho']
[]
['vexame', 'global', 'bolsonaro', 'ataca', 'outros', 'paises', 'exalta', 'ditadura', 'bajula', 'eua', 'onu']
['presidente']
['brasil']
['orgulho', 'bolsonaronaonu']
['bolsonaro', 'ataca', 'cuba', 'franca', 'venezuela', 'nega', 'devastacao', 'amazonia']
['parabens', 'presidente', 'arrasou', 'bolsonaronaonu']
```