russian_trolls

February 22, 2018

0.0.1 Russian Trolls: Proposal #2

This dataset released by NBC on Feb. 14 contains 200K tweets from Russian Trolls:

Topic: Interesting Dataset; Beginner Tutorial

Proposal:

1. Create a beginner tutorial on topic modeling
2. Tutorial on predicting number of retweets(Caveats: Very hard to predict)

Yellowbrick:

1. Proposal #1 :Use FreqDist to show most frequent words; Use t-SNE to visualize clustering
2. Proposal #2: Use Regression visualizers

```
In [1]: import pandas as pd
```

```
In [2]: df = pd.read_csv('tweets.csv.xz', dtype={'user_id': object,
                                                 'created_at': object,
                                                 'tweet_id': object})
```

```
In [3]: df[['user_key', 'retweet_count', 'text']].head()
        df[~df.retweet_count.isna()][['user_key', 'retweet_count',
                                      'text',
                                      'expanded_urls']].head()
```

Out[3]:

	user_key	retweet_count	text	expanded_urls
1	detroitdailynew	0.0	Clinton: Trump shouldve apologized more, atta...	["http://detne.ws/2e172jF"]
8	ameliebaldwin	0.0	RT @AriaWilsonGOP: 3 Women Face Charges After ...	["http://www.Feed24hNews.com/4MzaL"]
13	pamela_moore13	138.0	Dave Chappelle: "Black Lives Matter" is the wo...	[]
16	pamela_moore13	592.0	The war is here! \nThis gentleman made more se...	[]
20	kansasdailynews	0.0	Obama on Trump winning: 'Anything's possible' ...	["http://bit.ly/2axp6SI"]

```
In [4]: df.text.loc[1]
```

```
Out[4]: 'Clinton: Trump shouldve apologized more, attacked less https://t.co/eJampkoHFZ'
```