Performance Preprocessing elements

EmoticonReplace

Dict size: 224

Performance on train_pos_full:
replaced: 117661
not recognized: 21297

Performance on train_neg_full:
replaced: 70818
not recognized: 42064

- Not recognized is the count of word containing a special character that are longer than 1 (excluding hashtags). Most of these misses are not emoticons, they're times, dates and names spelled with underscore.
- Negative tweets contain less emojis. Also the distribution over the emojis is different for negative and positive tweets (see appendix).
- Runtime about 17 seconds per data set on my laptop.

ApposRemove

Dict size: 86

Performance on train_pos_full after ER
replaced: 394360
not recognized: 72790 (distinct: 9929)

Performance on train_neg_full after ER:
replaced: 362152
not recognized: 106732 (distinct: 13941)

- Not recognized is the count of words that contain a ' . Most of the remaining ones are either something like Justin's, or something f'k. Both of these are hard to tackle.
- Runtime about 14 seconds per data set on my laptop.

Appendix

EmoticonReplace

Emoticon count:

Train_pos_full:

```
{'<3': 41821, ":')": 2137, ':p': 8499, ':d': 17163, ';d': 1363, '(': 36054, ';]': 44, 'xd': 2465, 'xp': 126, ':/':
3467, ':\\': 144, ':)': 7, '=)': 1199, '=d': 241, '=]': 105, ':|': 498, 'oo': 426, '=p': 87, 'o_o': 362, 'd;': 69,
'd:': 373, '8-)': 35, '80': 295, ':}': 68, ':]': 278, '8)': 42, '=/': 66, ':@': 27, ':o)': 27, ':[': 22, "d':": 5, '=\\':
23, '>:p': 5, ':{': 27, ':-}': 5, ':-]': 6, 'dx': 27, '8d': 14, '>:/': 12, ':(': 3, '>:[': 4, 'd8': 11, 'd:<': 4, 'd=': 2,
'owo': 2, ';)': 1}
```

Train_neg_full:

```
{'xd': 1322, '3': 27359, '<3': 15394, 'd:': 1161, ':/': 6188, ':d': 7280, ":')": 1845, ':p': 4834, ';d': 562,
'dx': 281, ':\\': 266, ':|': 1006, '80': 975, ':]': 108, 'xp': 229, 'd8': 244, 'oo': 407, 'o_o': 373, 'd;': 158, ':(':
23, '8-)': 35, ':}': 43, '8)': 135, ':@': 232, ':[': 46, "d':": 102, ':{': 42, ';]': 25, '>:/': 32, ':o)': 10, 'owo': 3,
';)': 1, '>:[': 8, 'uwu': 2, 'd:<': 13, 'd=': 14, '8d': 44, ':)': 1, ':-]': 3, '>:p': 8, ':-}': 2, '>:\\': 2}
```