



Dynamic Context Analysis in Twitter:

Brown Biggers



Issues with Regular Expressions

- Parsing errors created problems with word count
- Defaults caused incorrect tokenization



Examples:

```
re_hash_test = '# #34 4#3 A#36 3 A# #hashtag'
print(re.sub(r'\b#\B', '!', re_hash_test))
print(re.sub(r'\b#\b', '!', re_hash_test))
print(re.sub(r'\B#\B', '!', re_hash_test))
```

```
# #34 4#3 A#36 3 A! #hashtag
# #34 4!3 A!36 3 A# #hashtag
! #34 4#3 A#36 3 A# #hashtag
```

```
text="this is a tweet # #### #1 #hashtag #123 #12345 apm yooooo"
tokens_no_stopwords(text)
```

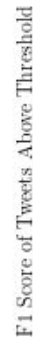
```
['tweet', '#1', '#hashtag', '#123', '#12345', 'yooo']
```

```
TAGGED 2017 #1 @ US 1-Biscayne Blvd. https://t.co/1foe57DVyC
['tagged', '2017', '#', '1', '@', 'us', '1-biscayne', 'blvd', '.', 'https', ':', '///t.co/1foe57dvyc']
['tagged', '2017', '#1', '@', 'us', '1-biscayne', 'blvd.', 'https://t.co/1foe57dvyc']
S: tagged 2017 #1 @ us 1-biscayne blvd. https://t.co/1foe57dvyc
1: tagged 2017 #1 @ us 1-biscayne blvd.
2: tagged 2017 #1 @ us 1-biscayne blvd.
3: tagged 2017 #1 us 1biscayne blvd
4: tagged 2017 #1 us 1biscayne blvd
5: tagged 2017 #1 us 1biscayne blvd
6: tagged 2017 #1 1biscayne blvd
7: tagged #1 1biscayne blvd
['tagged', '#', '1', '1biscayne', 'blvd']
['tagged', '#1', '1biscayne', 'blvd']
```



Topics for next week:

- Neural network configuration
 - Negative Sampling v. Hierarchical Softmax
- Supporting documentation for Word2Vec in Twitter analysis





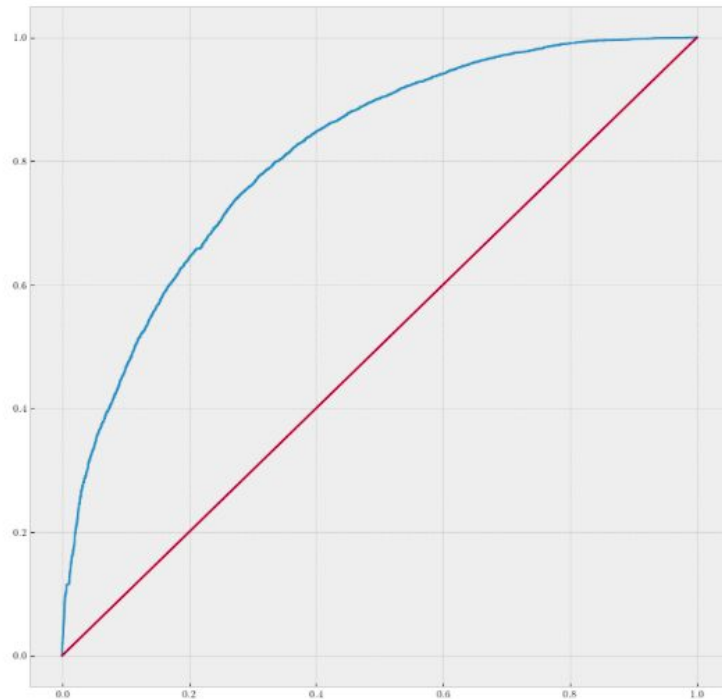
Negative Sampling Testing

	MM_Score	F1_Score_Max
0	0	0.000000
1	23	0.716133
2	23	0.713791
3	23	0.709994
4	23	0.707330
5	23	0.706207
6	24	0.703772
7	24	0.701751
8	24	0.701094
9	23	0.699942
10	24	0.697631

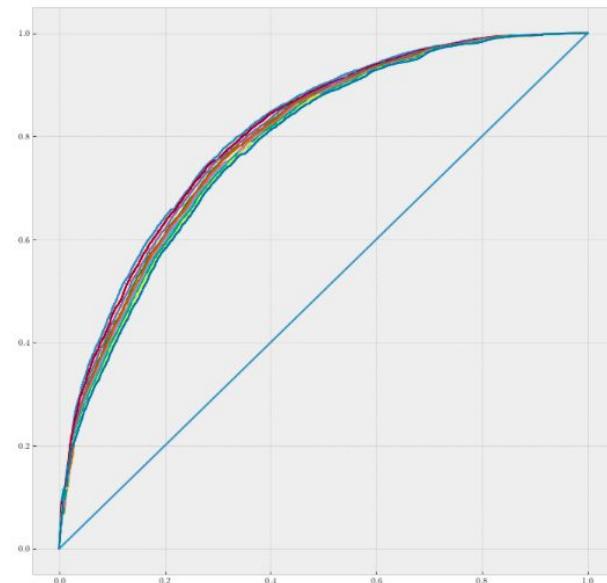
AU-ROC Curves

```
fpr, tpr, thresholds = roc_curve(tweet_encoded['irma_rel'], tweet_encoded['MM_score_window_1'])
```

```
fig_roc_1 = plt.figure(figsize=(12,12),facecolor='w')  
plt.plot(fpr,tpr)  
plt.plot([0,1],[0,1])  
plt.show()
```



```
fig_roc_1 = plt.figure(figsize=(12,12),facecolor='w')  
for i in range(1,11):  
    fpr, tpr, thresholds = roc_curve(tweet_encoded['irma_rel'], tweet_encoded[f'MM_score_window_{i}'])  
    plt.plot(fpr,tpr)  
plt.plot([0,1],[0,1])  
plt.show()
```





Topics for next week:

- Supporting documentation for Word2Vec in Twitter analysis
- Clean up related code for results.