

Brute force : multinomial naive bayes on basic features

- Default parameters for multinomial naive bayes
- No cross validation, no hyper parameters
- train is done on unbalanced data (distribution of is_duplicate is not the same on train(37%) and challenge data (17%))

But a bunch of features are tested

- Compute common_words between question1 & question2
- Compute Nb common_words between question1 & question2
- Compute Nb common words/nb words in question1
- Compute Nb common words/nb words in question2
- Compute Nb words in question1/question2 not in common words
- Compute (Nb common words)/(nb words in question1+nb word in question2) These features are combined to generate 255 models (*2 as bot models with/without tests are trained)

Sequence

- All features are generated
 - on training data
 - on challenge data (overkill as some features may be rejected later but easier)
- AUC graphs are computed for each single feature, so we can visually see if it is potentially useful or obviously useless
- 2 Models are trained on AUL combination of features
 - on a partition 80/20
 - on full training data
- Performances of all models are saved in an excel file
 - logloss on proba (test+full)
 - logloss on decision (test+full)
 - accuracy (test+full)
 - score (weighted combination of f1-score) (test+full)
- Various requests on this database of performances are done
 - Apply is done on best 3 models (test & full)
 - csv is generated in a csv
 - proba is zipped
 - command line to publish in kaggle is displayed (no automatic submission to keep control)
- Kaggle performances are downloaded and saved
- Various requests on kaggle performances are done

```
In [4]: # Ugly incantation to make our framework working
import sys
sys.path.insert(0, r'./SAPDeveloP/QuoraPairs/BruteForce/Tools')

# Import all our small tools (paths, cache, print, zip, excel, pandas, progress...)
from Tools.all import *

# Setup the name of our experiment
# It will be used to store every result in a unique place
```

```
In [5]: def add_column_from_columns(dataframe, output_column_name, function):
    dataframe[output_column_name] = dataframe.progress_apply(function, axis=1)
    return dataframe[output_column_name]

def add_column_from_column(dataframe, output_column_name, input_column_name, function):
    dataframe[output_column_name] = dataframe[input_column_name].progress_apply(function)
    return dataframe[output_column_name]

def build_all_simple_features(dataframe):
    print_small()
    print_warning('Compute common words between question1 & question2')
    add_column_from_columns(dataframe, 'nb_words_question1', 'question1', lambda x: len(x.split()))
    add_column_from_columns(dataframe, 'nb_words_question2', 'question2', lambda x: len(x.split()))
    print_warning('Compute Nb common words between question1 & question2')
    add_column_from_columns(dataframe, 'common_words', lambda r: list(set(r.question1.split()) & set(r.question2.split())))
    add_column_from_column(dataframe, 'nb_common_words', 'common_words', len)

    print_warning('Compute Nb common words/nb words in question1')
    add_column_from_columns(dataframe, 'nb_words_question1-common_words', lambda r: r.nb_common_words / r.nb_words_question1)
    print_warning('Compute Nb common words/nb words in question2')
    add_column_from_columns(dataframe, 'nb_words_question2-common_words', lambda r: r.nb_common_words / r.nb_words_question2)
    print_warning('Compute (nb common words)/(nb words in question1+nb word in question2)')
    add_column_from_columns(dataframe, 'nb_common_words/(nb_words_question1+nb_words_question2)', lambda r: r.nb_common_words / (r.nb_words_question1 + r.nb_words_question2))
    end_small()
    dataframe = dataframe.drop(columns='common_words')
    return dataframe
```

```
In [6]: EXPERIMENT='multinomial_basic_features_unbalanced'

prepare_environment(EXPERIMENT)
train_dataframe=load_dataframe(CLEAN_TRAINING_DATA)
challenge_dataframe=load_dataframe(CLEAN_CHALLENGE_DATA)
print_section('Unouched input data has been loaded. Training: %d lines Challenge: %d lines' % (len(train_dataframe), len(challenge_dataframe)))

train_dataframe=load_or_build_dataframe('Training data + basic features', 'training_basic_features', build_all_simple_features(train_dataframe))
challenge_dataframe=load_or_build_dataframe('Challenge data + basic features', 'challenge_basic_features', build_all_simple_features(challenge_dataframe))

Prepare multinomial_basic_features_unbalanced environment in ./multinomial_basic_features_unbalanced
```

```
Create ./multinomial_basic_features_unbalanced
Make local copy of ./PandasStore/clean_training.pkl
Make local copy of ./PandasStore/clean_challenge.pkl
Done

Unouched input data has been loaded. Training: 404290 lines Challenge: 2345796 lines
```

Training data + basic features: Load or rebuild training_basic_features

```
!!! ./multinomial_basic_features_unbalanced/training_basic_features.pkl does not exist!!!

Rebuild and save it

Compute common_words between question1 & question2

Compute Nb common_words between question1 & question2

Compute Nb common words/nb words in question1

Compute Nb common words/nb words in question2

Compute Nb words in question1/question2 not in common words

Compute (nb common words)/(nb words in question1+nb word in question2)

Save training_basic_features
Done training_basic_features contains 404290 lines in 113.7 s
```

Challenge data + basic features: Load or rebuild challenge_basic_features

```
!!! ./multinomial_basic_features_unbalanced/challenge_basic_features.pkl does not exist!!!

Rebuild and save it

Compute common_words between question1 & question2

Compute Nb common_words between question1 & question2

Compute Nb common words/nb words in question1

Compute Nb common words/nb words in question2

Compute Nb words in question1/question2 not in common words

Compute (nb common words)/(nb words in question1+nb word in question2)

Save challenge_basic_features
Done challenge_basic_features contains 2345796 lines in 675.1 s
```

```
In [5]: train_dataframe.head()[1:1]

Out[5]:
```

id	qid1	qid2	question1	question2	is_duplicate	nb_words_question1	nb_words_question2	nb_common_words	nb_common_words/nb_words
0	0	1	2	What is the step by step guide to invest in market in india?	0	12	10		

```
In [6]: challenge_dataframe.head()[1:1]

Out[6]:
```

test_id	question1	question2	nb_words_question1	nb_words_question2	nb_common_words	nb_common_words/nb_words_question1	nb_common_words/nb_words_question2
0	How does the Surface Pro himself 4 compare with iPad Pro?	Why did Microsoft choose core m3 and not core i3 home Surface Pro 4?	11	14	2		0.181818

Compute AUC of these basic features and try to figure if there is a bit of information inside each one

ie it is helping to separate 1 from 0?

```
In [7]: from sklearn.metrics import roc_auc_score

def simple_auc(dataframe, column_name):
    return roc_auc_score(dataframe['is_duplicate'], dataframe[column_name])

def show_auc(dataframe, column_name):
    print_bullet('AUC: %s' % 'f' % (column_name, simple_auc(dataframe, column_name)))

def display_simple_auc(dataframe, column_name):
    palette = sns.color_palette()
    # Let's make a figure of the size
    plot.figure(figsize=(10, 7))
    # Plot histogram of the size
    plot.hist(dataframe[column_name][dataframe['is_duplicate']==1], bins=50, color=palette[3], label='Same')
    plot.hist(dataframe[column_name][dataframe['is_duplicate']==0], bins=50, color=palette[2], label='Different', alpha=0.75, histtype='step')
    plot.title('AUC: %s' % 'f' % (column_name, simple_auc(dataframe, column_name)), fontsize=10)
    plot.xlabel(column_name)
    plot.ylabel('Nb')
    plot.legend()
```

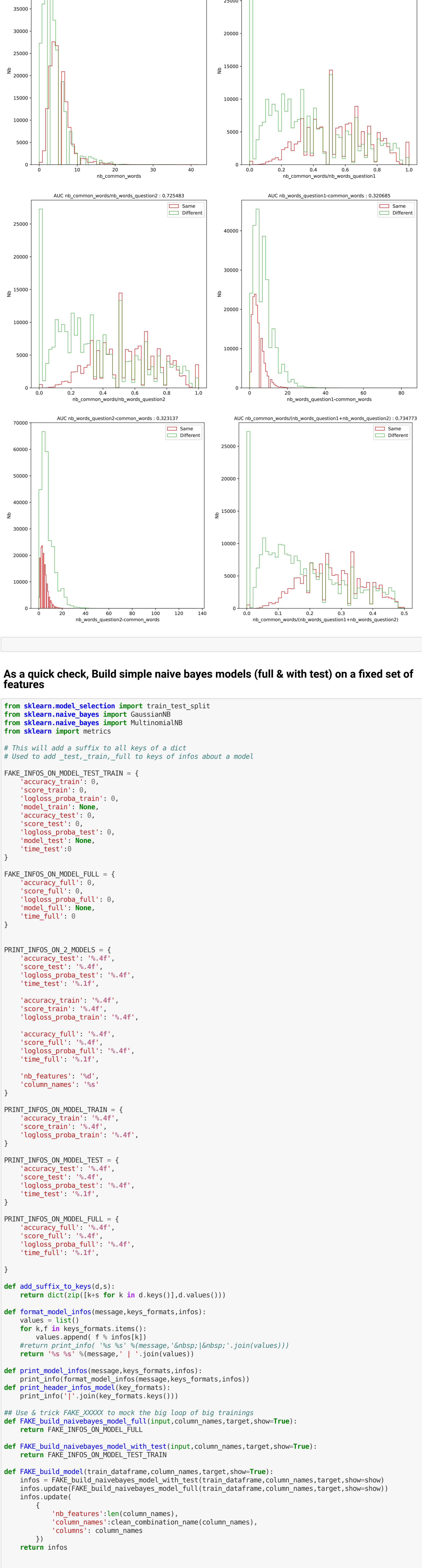
```
def show_all_simple_auc(dataframe):
    print_section('Show AUC on %d unique features' % len(all))
    for name in all:
        show_auc(dataframe, name)
    yield
    display_simple_auc(dataframe, name)
    print_done('Done')

def show_all_simple_auc_in_grid(dataframe, nb_columns=2):
    multiplot_from_generator(show_all_simple_auc(dataframe, nb_columns), nb_columns)

show_all_simple_auc_in_grid(train_dataframe, nb_columns=2)

Show AUC on 8 unique features
```

- AUC nb_words_question1: 0.416034
- AUC nb_words_question2: 0.420794
- AUC nb_common_words: 0.680781
- AUC nb_common_words/nb_words_question1: 0.723389
- AUC nb_common_words/nb_words_question2: 0.725483
- AUC nb_words_question1-common_words: 0.320685
- AUC nb_words_question2-common_words: 0.320337
- AUC nb_common_words/(nb_words_question1+nb_words_question2): 0.734773



As a quick check, Build simple naive bayes models (full & with test) on a fixed set of features

```
In [21]: from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, support_score

# This will add a suffix to all keys of a dict
# Used to add model_train, full to keys of infos about a model

FAKE_INFOS_ON_MODEL_TEST_TRAIN = {
    'accuracy_train': 0,
    'score_train': 0,
    'logloss_proba_train': 0,
    'model_train': None,
    'accuracy_test': 0,
    'score_test': 0,
    'logloss_proba_test': 0,
    'model_test': None,
    'time_test': 0
}

FAKE_INFOS_ON_MODEL_FULL = {
    'accuracy_full': 0,
    'score_full': 0,
    'logloss_proba_full': 0,
    'model_full': None,
    'time_full': 0
}

PRINT_INFOS_ON_2_MODELS = {
    'accuracy_test': '%.4f',
    'score_test': '%.4f',
    'logloss_proba_test': '%.4f',
    'time_test': '%.1f',
    'accuracy_train': '%.4f',
    'score_train': '%.4f',
    'logloss_proba_train': '%.4f',
    'accuracy_full': '%.4f',
    'score_full': '%.4f',
    'logloss_proba_full': '%.4f',
    'time_full': '%.1f',
    'nb_features': 'nd',
    'column_names': 'ns'
}

PRINT_INFOS_ON_MODEL_TRAIN = {
    'accuracy_train': '%.4f',
    'score_train': '%.4f',
    'logloss_proba_train': '%.4f',
}

PRINT_INFOS_ON_MODEL_TEST = {
    'accuracy_test': '%.4f',
    'score_test': '%.4f',
    'logloss_proba_test': '%.4f',
    'time_test': '%.1f',
}

PRINT_INFOS_ON_MODEL_FULL = {
    'accuracy_full': '%.4f',
    'score_full': '%.4f',
    'logloss_proba_full': '%.4f',
    'time_full': '%.1f',
}

def add_suffix_to_keys(d, s):
    return dict(zip([k+s for k in d.keys()], d.values()))

def build_model_infos(message, keys_formats, infos):
    values = list()
    for k, f in keys_formats.items():
        values.append(f % infos[k])
    return '\n%s\n' % message, '\n%s\n' % message, '\n%s\n' % message, '\n%s\n' % message, '\n%s\n' % message

def print_model_infos(message, keys_formats, infos):
    print_info(format_model_infos(message, keys_formats, infos))

def print_header_infos_model(key_formats):
    print_info(' '.join(key_formats))

# Use & trick FAKE_XXXXX to mock the big loop of big trainings
def FAKE_build_naivebayes_model_full(input, column_names, target, show=True):
    return FAKE_INFOS_ON_MODEL_FULL

def FAKE_build_naivebayes_model_with_test(input, column_names, target, show=True):
    return FAKE_INFOS_ON_MODEL_TEST_TRAIN

def FAKE_build_naivebayes_model_full_with_test(input, column_names, target, show=True):
    return FAKE_INFOS_ON_MODEL_FULL

def FAKE_build_model(train_dataframe, column_names, target, show=True):
    infos = FAKE_build_naivebayes_model_full(input, column_names, target, show=show)
    infos.update(FAKE_build_naivebayes_model_full(train_dataframe, column_names, target, show=show))
    infos.update({
        'nb_features': len(column_names),
        'column_names': clean_combination_name(column_names),
        'columns': column_names
    })
    return infos

def compute_metrics_model(model, input_df, target_df, suffix, show=True):
    prediction_df = model.predict(input_df)
    prediction_proba = model.predict_proba(input_df)
    res = metrics.classification_report(target_df, prediction_df, output_dict=True)
    accuracy = res['accuracy']
    score = res['f1-score'] * (1.0 - 17) + res['f1-score'] * 17
    logloss_proba = metrics.log_loss(target_df, prediction_proba, df)
    if show:
        print('Classification report on %s' % suffix)
        print(metrics.classification_report(target_df, prediction_df))
    return add_suffix_to_keys({
        'accuracy': accuracy,
        'score': score,
        'logloss_proba': logloss_proba,
        'model': model,
    }, suffix)

def build_naivebayes_model_with_test(input, column_names, target, show=True):
    # Print bullet 'Multinomial Naive Bayes with test on %s' % clean_combination_name(column_names)
    input_train = pandas.DataFrame()
    for column_name in column_names:
        input_train[column_name] = input[column_name]
    target_train = target
    input_train, input_test, target_train, target_test = train_test_split(input_train, target_train, random_state=42, test_size=0.2)
    if show:
        print_info('Training %d Lines Test %d Nb Features: %d' % (len(input_train), len(input_test), len(input_train.columns)))
    model = MultinomialNB()
    # naive_bayes_classifier_with_test=ComplementNB()
    start = time.time()
    model.fit(input_train, target_train)
    duration = time.time() - start
    infos = compute_metrics_model(model, input_test, target_test, 'test', show=show)
    infos.update(compute_metrics_model(model, input_train, target_train, 'train', show=show))
    if show:
        print_done('time_test: %s' % duration)
    if show:
        print_model_infos('Test', PRINT_INFOS_ON_MODEL_TEST, infos)
        print_model_infos('Train', PRINT_INFOS_ON_MODEL_TRAIN, infos)
    return infos

def build_naivebayes_model_full(input, column_names, target, show=True):
    input_full = pandas.DataFrame()
    for column_name in column_names:
        input_full[column_name] = input[column_name]
    if show:
        print_info('Training on %dx%d' % (len(input_full), len(input_full.columns)))
    target_full = target
    model = MultinomialNB()
    start = time.time()
    model.fit(input_full, target_full)
    duration = time.time() - start
    infos = compute_metrics_model(model, input_full, target_full, 'full', show=show)
    infos.update('time_full: %s' % duration)
    if show:
        print_model_infos('Full', PRINT_INFOS_ON_MODEL_FULL, infos)
    return infos

def build_model(train_dataframe, column_names, target, show=True):
    infos = build_naivebayes_model_with_test(train_dataframe, column_names, target, show=show)
    infos.update(build_naivebayes_model_full(train_dataframe, column_names, target, show=show))
    infos.update({
        'nb_features': len(column_names),
        'column_names': clean_combination_name(column_names),
        'columns': column_names
    })
    return infos

print_section('Quick check: Basic model using feature nb_words_question1')
build_model(train_dataframe, ['nb_words_question1'], train_dataframe['is_duplicate'])

Training 323432 lines Test 80858 Nb Features: 1

Classification report on_test
precision recall f1-score support
0 0.63 1.00 0.77 50803
1 0.00 0.00 0.00 30955
accuracy 0.31 0.50 0.63 80858
macro avg 0.39 0.50 0.63 80858
weighted avg 0.39 0.63 0.48 80858

Classification report on_train
precision recall f1-score support
0 0.63 1.00 0.77 264224
1 0.00 0.00 0.00 119268
accuracy 0.32 0.50 0.63 323432
macro avg 0.32 0.50 0.39 323432
weighted avg 0.40 0.63 0.49 323432

Test 0.6283(0.6405) 0.6599(0.0)
Train 0.6514(0.6425) 0.6582

Training 404290 lines Test 80858 Nb Features: 1

Classification report on_full
precision recall f1-score support
0 0.63 1.00 0.77 255027
1 0.00 0.00 0.00 149263
accuracy 0.32 0.50 0.63 404290
macro avg 0.32 0.50 0.39 404290
weighted avg 0.40 0.63 0.49 404290

Full 0.6508(0.6421) 0.6585(0.1)
```

```
Out[21]: {'accuracy_test': 0.62829893268099,
'score_test': 0.640531212735738,
'logloss_proba_test': 0.658075918235348,
'model_test': MultinomialNB(),
'accuracy_train': 0.631427935392911,
'score_train': 0.6424864684567217,
'logloss_proba_train': 0.6581913334694939,
'model_train': MultinomialNB(),
'time_test': 0.041047811590317,
'accuracy_full': 0.638802146973707,
'score_full': 0.6428960175454294,
'logloss_proba_full': 0.6585273839844356,
'model_full': MultinomialNB(),
'time_full': 0.05285835266113281,
'nb_features': 1,
'column_names': ['nb_words_question1'],
'columns': ['nb_words_question1']}
```

Build models (test & full) on all combinations of features

```
In [46]: # Had design choice : a DataFrame can be more convenient than a dict
# But then, it is convenient to suppress all non numeric/string columns

def models_dict_to_df(models_dict):
    # Remove non numeric/string columns in a way I use more convenient
    return pandas.DataFrame.from_dict(models_dict, orient='index').reindex(columns=['logloss_proba_test', 'logloss_proba_train', 'logloss_proba_full', 'nb_features', 'column_names', 'accuracy_test', 'accuracy_train', 'accuracy_full', 'score_test', 'score_train', 'score_full', 'model_test', 'model_train', 'model_full', 'columns', 'time_test', 'time_train', 'time_full'])

def build_model_on_all_subset_of_simple_features(dataframe, target):
    print_section('Build all models (with test+full) on every combination of simple features - %d lines' % (len(dataframe)))
    models_dict = dict()
    print_header_infos_model(PRINT_INFOS_ON_2_MODELS)
    all_combinations = list(all_subsets(all_numeric_columns(dataframe)))
    progress = tqdm(all_combinations)
    for c in progress:
        if (len(c)) == 1:
            if (len(c)) >= 0:
                infos = FAKE_build_model(dataframe, c, target, show=False)
                models_dict[clean_combination_name(c)] = infos
                progress.refresh()
                print_done('Done - top-start')
            # Design mistake : need to convert dict to dataframe : (
            return models_dict_to_df(models_dict)

def FAKE_build_model_on_all_subset_of_simple_features(dataframe, target):
    start = time.time()
    print_section('Build all models (with test+full) on every combination of simple features - %d rows' % (EXPERIMENT, len(dataframe)))
    models_dict = dict()
    print_header_infos_model(PRINT_INFOS_ON_2_MODELS)
    all_combinations = list(all_subsets(all_numeric_columns(dataframe)))
    progress = tqdm(all_combinations)
    for c in progress:
        if (len(c)) >= 0:
            infos = FAKE_build_model(dataframe, c, target, show=False)
            models_dict[clean_combination_name(c)] = infos
            print_info(format_model_infos(' ', PRINT_INFOS_ON_2_MODELS, infos))
            progress.refresh()
            print_done('Done - top-start')
            # Design mistake : need to convert dict to dataframe : (
            return models_dict_to_df(models_dict)

def graph_all_metrics_all_models(results, y_label, metric_sort, metrics):
    palette = sns.color_palette()
    data = results.sort_values(metric_sort+'test', ascending=(metric_sort=='logloss_proba'))
    # x = plot
    # fig, ax = plot.subplots()
    x = numpy.arange(len(data))
    num_col = 1
    if len(metrics) > 1:
        kinds = ['test']
    else:
        kinds = ['test', 'train', 'full']
    for k in kinds:
        for m in metrics:
            plot.plot(x, data[m+k], color = palette[num_col], label = m+k)
            num_col += 1
    plot.ylabel(y_label)
    # Plot plot(x, data[m+k], color = palette[num_col])
    plot.plot(x, data['time_full'], color = palette[num_col+1], label = 'time_full')
    plot.title(EXPERIMENT+'-'+y_label, fontsize=8)
    # ax.set_xticklabels(labels)
    plot.legend()
```

If you want to quickly validate some code, activate this massive subsampling

Results will be useless but code coverage will be ok. This is also the good place to rebalance training data (cf next experiments)


```

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040

```


common_words+nb_words_question2-common_words+nb_common_words/(nb_words_question1+nb_words_question2)
0.6767 | 0.7120 | 0.9546 | 0.0 | 0.6752 | 0.7097 | 0.9469 | 0.6756 | 0.7101 | 0.9481 | 0.0 | 7 |
nb_words_question1+nb_words_question2+nb_common_words+nb_common_words/nb_words_question1+nb_common_words/nb_words_question2
common_words+nb_words_question2-common_words
0.6710 | 0.7065 | 0.8101 | 0.0 | 0.6693 | 0.7041 | 0.8078 | 0.6697 | 0.7047 | 0.8069 | 0.0 | 7 |
nb_words_question1+nb_words_question2+nb_common_words+nb_common_words/nb_words_question1+nb_common_words/nb_words_question2
common_words+nb_common_words/(nb_words_question1+nb_words_question2)
0.6767 | 0.7109 | 0.8420 | 0.0 | 0.6739 | 0.7074 | 0.8317 | 0.6746 | 0.7082 | 0.8348 | 0.0 | 7 |
nb_words_question1+nb_words_question2+nb_common_words+nb_common_words/nb_words_question1+nb_common_words/nb_words_question2
common_words+nb_common_words/(nb_words_question1+nb_words_question2)
0.6775 | 0.7128 | 0.9379 | 0.0 | 0.6746 | 0.7092 | 0.9298 | 0.6752 | 0.7099 | 0.9310 | 0.0 | 7 |
nb_words_question1+nb_words_question2+nb_common_words+nb_common_words/nb_words_question1+nb_words_question1+nb_words_question2
common_words+nb_common_words/(nb_words_question1+nb_words_question2)
0.6777 | 0.7131 | 0.9396 | 0.0 | 0.6753 | 0.7099 | 0.9316 | 0.6759 | 0.7106 | 0.9328 | 0.0 | 7 |
nb_words_question1+nb_words_question2+nb_common_words+nb_common_words/nb_words_question1+nb_words_question1+nb_words_question2
common_words+nb_common_words/(nb_words_question1+nb_words_question2)
0.6730 | 0.7113 | 0.6648 | 0.0 | 0.6676 | 0.7057 | 0.6631 | 0.6690 | 0.7070 | 0.6632 | 0.0 | 7 |
nb_words_question1+nb_words_question2+nb_common_words+nb_common_words/nb_words_question1+nb_common_words/nb_words_question2
common_words+nb_common_words/(nb_words_question1+nb_words_question2)
0.6773 | 0.7123 | 0.9703 | 0.0 | 0.6748 | 0.7091 | 0.9625 | 0.6752 | 0.7097 | 0.9639 | 0.0 | 7 |
nb_words_question1+nb_words_question2+nb_common_words+nb_common_words/nb_words_question1+nb_common_words/nb_words_question2+nb_words_question1
common_words+nb_common_words/(nb_words_question1+nb_words_question2)
0.6767 | 0.7124 | 0.9645 | 0.0 | 0.6745 | 0.7093 | 0.9575 | 0.6749 | 0.7099 | 0.9582 | 0.0 | 7 |
nb_words_question1+nb_words_question2+nb_common_words+nb_common_words/nb_words_question1+nb_words_question2+nb_words_question1
common_words+nb_common_words/(nb_words_question1+nb_words_question2)
0.6773 | 0.7123 | 0.9722 | 0.0 | 0.6749 | 0.7092 | 0.9646 | 0.6755 | 0.7100 | 0.9657 | 0.0 | 8 |
nb_words_question1+nb_words_question2+nb_common_words+nb_common_words/nb_words_question1+nb_common_words/nb_words_question2
common_words+nb_words_question2-common_words+nb_common_words/(nb_words_question1+nb_words_question2)

Done in 30.6 s

The big loop

nb_words_question2+nb_common_words/nb_words_question1+nb_common_words/nb_words_question2+nb_words_question1-
common_words+nb_words_question2-common_words+nb_common_words/(nb_words_question1+nb_words_question2)
0.6729 | 0.7065 | 0.9553 | 0.0 | 0.6729 | 0.7068 | 0.9535 | 0.6729 | 0.7067 | 0.9537 | 0.1 | 6 |
nb_common_words+nb_common_words/nb_words_question1+nb_common_words/nb_words_question2+nb_words_question1-
nb_common_words+nb_words_question2-common_words+nb_common_words/(nb_words_question1+nb_words_question2)
0.6737 | 0.7071 | 0.9536 | 0.1 | 0.6734 | 0.7071 | 0.9515 | 0.6734 | 0.7072 | 0.9517 | 0.1 | 7 |
nb_words_question1+nb_words_question2+nb_common_words+nb_common_words/nb_words_question1+nb_common_words/nb_words_question2
common_words+nb_words_question2-common_words
0.6696 | 0.7036 | 0.8074 | 0.0 | 0.6704 | 0.7045 | 0.8073 | 0.6702 | 0.7043 | 0.8082 | 0.1 | 7 |
nb_words_question1+nb_words_question2+nb_common_words+nb_common_words/nb_words_question1+nb_common_words/nb_words_question2
common_words+nb_common_words/(nb_words_question1+nb_words_question2)
0.6676 | 0.7005 | 0.8449 | 0.0 | 0.6699 | 0.7028 | 0.8416 | 0.6695 | 0.7024 | 0.8412 | 0.1 | 7 |
nb_words_question1+nb_words_question2+nb_common_words+nb_common_words/nb_words_question1+nb_words_question1-
common_words+nb_words_question2-common_words/(nb_words_question1+nb_words_question2)
0.6723 | 0.7069 | 0.9363 | 0.1 | 0.6723 | 0.7072 | 0.9341 | 0.6723 | 0.7072 | 0.9344 | 0.1 | 7 |
nb_words_question1+nb_words_question2+nb_common_words+nb_common_words/nb_words_question1+nb_words_question1-
common_words+nb_words_question2-common_words/(nb_words_question1+nb_words_question2)
0.6734 | 0.7071 | 0.9384 | 0.0 | 0.6737 | 0.7076 | 0.9362 | 0.6735 | 0.7076 | 0.9364 | 0.1 | 7 |
nb_words_question1+nb_words_question2+nb_common_words+nb_common_words/nb_words_question2+nb_words_question1-
common_words+nb_words_question2-common_words+nb_common_words/(nb_words_question1+nb_words_question2)
0.6651 | 0.7028 | 0.6704 | 0.0 | 0.6664 | 0.7041 | 0.6683 | 0.6703 | 0.7038 | 0.6686 | 0.1 | 7 |
nb_words_question1+nb_words_question2+nb_common_words/nb_words_question1+nb_words_question2+nb_words_question1-
common_words+nb_words_question2-common_words+nb_common_words/(nb_words_question1+nb_words_question2)
0.6734 | 0.7066 | 0.9608 | 0.0 | 0.6731 | 0.7067 | 0.9677 | 0.6732 | 0.7067 | 0.9678 | 0.1 | 7 |
nb_words_question1+nb_common_words+nb_common_words/nb_words_question1+nb_common_words/nb_words_question2+nb_words_question1-
common_words+nb_words_question2-common_words+nb_common_words/(nb_words_question1+nb_words_question2)
0.6726 | 0.7065 | 0.9629 | 0.0 | 0.6727 | 0.7069 | 0.9612 | 0.6726 | 0.7068 | 0.9615 | 0.1 | 7 |
nb_words_question2+nb_words_question2-common_words+nb_common_words/nb_words_question1+nb_words_question2+nb_words_question1-
common_words+nb_words_question2-common_words+nb_common_words/(nb_words_question1+nb_words_question2)
0.6734 | 0.7066 | 0.9714 | 0.1 | 0.6732 | 0.7068 | 0.9693 | 0.6732 | 0.7068 | 0.9695 | 0.1 | 8 |
common_words+nb_words_question2-common_words+nb_common_words/(nb_words_question1+nb_words_question2)
common_words+nb_words_question2-common_words+nb_common_words/(nb_words_question1+nb_words_question2)

Done in 396.5 s

In [32]: model_results.head().transpose()

Out [32]:

	nb_words_question1	nb_words_question2	nb_common_words	nb_common_words/nb_words_question1	nb_common_w
logloss_proba_test	0.659876	0.659876	0.659876	0.659876	
logloss_proba_train	0.658191	0.658191	0.658191	0.658191	
logloss_proba_full	0.658527	0.658527	0.658527	0.658527	
nb_features	1	1	1	1	
column_names	nb_words_question1	nb_words_question2	nb_common_words	nb_common_words/nb_words_question1	nb_common_w
accuracy_test	0.628299	0.628299	0.628299	0.628299	
accuracy_train	0.631428	0.631428	0.631428	0.631428	
accuracy_full	0.630802	0.630802	0.630802	0.630802	
score_test	0.640531	0.640531	0.640531	0.640531	
score_train	0.642486	0.642486	0.642486	0.642486	
score_full	0.642096	0.642096	0.642096	0.642096	
model_test	MultinomialNB()	MultinomialNB()	MultinomialNB()	MultinomialNB()	
model_train	MultinomialNB()	MultinomialNB()	MultinomialNB()	MultinomialNB()	
model_full	MultinomialNB()	MultinomialNB()	MultinomialNB()	MultinomialNB()	
columns	(nb_words_question1)	(nb_words_question2)	(nb_common_words)	(nb_common_words/nb_words_question1)	(nb_common_w
time_test	0.0417576	0.0350451	0.0387604	0.0385814	
time_full	0.0490961	0.0480876	0.0449538	0.0470052	

Graph metrics on all models generated

