lab_soln-6-4

FINISHED

```
%pyspark
tweets = sc.textFile("s3a://conor-twitter-firehose/2017/04/*/*/*")

def get_langs(tweet_str):
    '''
    INPUT: string of tweet
    OUTPUT: list of length 2 w/ the following values:
            text
            lang
    '''
    import json
    try:
        tweet = json.loads(tweet_str)
        text = tweet.get('text')
        lang = tweet.get('lang')

        if lang == 'en':
            return [text, 0]
        elif lang == 'es':
            return [text, 1]
        else:
            pass
    except ValueError:
        pass

df = tweets.map(get_langs).filter(lambda x: x is not None).toDF(['tweet', 'lang']).cache()
```

Took 33 sec. Last updated by anonymous at April 24 2017, 8:43:48 AM.

FINISHED

```
%pyspark
from pyspark.ml.feature import RegexTokenizer, HashingTF, IDF
from pyspark.ml import Pipeline
from pyspark.ml.classification import RandomForestClassifier

tokenizer = RegexTokenizer(inputCol="tweet", outputCol="words", pattern='\s+|[,.\"]')
hashingTF = HashingTF(inputCol="words", outputCol="rawFeatures", numFeatures=200)
idf = IDF(inputCol="rawFeatures", outputCol="features")
forestizer = RandomForestClassifier(labelCol="lang", featuresCol="features", numTrees=10)

pipeline = Pipeline(stages=[\
                tokenizer,
                hashingTF,
                idf,
                forestizer])
```

Took 30 sec. Last updated by anonymous at April 24 2017, 8:43:48 AM. (outdated)

FINISHED

```
%pyspark

tweets_train, tweets_test = df.randomSplit([0.7, 0.3], seed=123)

model = pipeline.fit(tweets_train)
test_model = model.transform(tweets_test)
```

Took 1 hrs 2 min 3 sec. Last updated by anonymous at April 24 2017, 9:45:51 AM.

FINISHED

```
%pyspark
import pyspark.ml.evaluation as ev
```

```
evaluator = ev.BinaryClassificationEvaluator(rawPredictionCol='probability', labelCol='lang')
print('AUC for Random Forest:', evaluator.evaluate(test_model, {evaluator.metricName: 'areaUnderROC'}))
print('APR for Random Forest:', evaluator.evaluate(test_model, {evaluator.metricName: 'areaUnderPR'}))
```

```
('AUC for Random Forest:', 0.9074535775862083)
('APR for Random Forest:', 0.8123094078113648)
```

Took 2 hrs 7 min 27 sec. Last updated by anonymous at April 24 2017, 10:51:15 AM.

```
%pyspark
pipeline.save('s3a://conor-sandbox/6-4-ML-pipeline')
model.save('s3a://conor-sandbox/6-4-ML-model')
```

READY

```
%pyspark
tokenized = tokenizer.transform(df) # Take a look at the tokenized text
tokenized.take(10)
```

FINISHED

```
[Row(tweet=u'Victor Hugo y C5N defendiendo al dictador Maduro cuando hay 2 muertos en Venezuela. Vayanse al carajo basuras, lacras.', lang=1, words=[u'victor', u'hugo', u'y', u'c5n', u'defendiendo', u'al', u'dictador', u'maduro', u'cuando', u'hay', u'2', u'muertos', u'en', u'venezuela', u'vayanse', u'al', u'carajo', u'basuras', u'lacras']), Row(tweet=u'Al C\xe9sar lo que es del C\xe9sar yyyyyyy\n\xbfAl Guaire?\nAunque usted no lo crea, SE BA\xd1ARON en el Guaire https://t.co/8IrM320fUk', lang=1, words=[u'al', u'c\xe9sar', u'lo', u'que', u'es', u'del', u'c\xe9sar', u'yyyyyyy', u'\xbfal', u'guaire?', u'aunque', u'usted', u'no', u'lo', u'crea', u'se', u'ba\xf1aron', u'en', u'el', u'guaire', u'https://t', u'co/8irm320fuk']), Row(tweet=u'A new model for lectures in the post-truth era? https://t.co/8AImi8D4pT', lang=0, words=[u'a', u'new', u'model', u'for', u'lectures', u'in', u'the', u'post-truth', u'era?', u'https://t', u'co/8aimi8d4pt']), Row(tweet=u'Those "get ___ rt\'s and I\'ll get you _____" tweets...\n\nThey\'re both annoying and sweet all at once.\n\nLow key wish that could happen to me', lang=0, words=[u'those', u'get', u'___', u"rt's", u'and', u"i'll", u'get', u'you', u'_____', u'tweets', u"they're", u'both', u'annoying', u'and', u'sweet', u'all', u'at', u'once', u'low', u'key', u'wish', u'that', u'could', u'happen', u'to', u'me']), Row(tweet=u'What do you think of the summer fashion trend. https://t.co/COPufAE6LV', lang=0, words=[u'what', u'do', u'you', u'think', u'of', u'the', u'summer', u'fashion', u'trend', u'https://t', u'co/copufae6lv']), Row(tweet=u'La Dana me pide las fotos de ese chico JAJJAJA La amo\U0001f498', lang=1, words=[u'la', u'dana', u'me', u'pide', u'las', u'fotos', u'de', u'ese', u'chico', u'jajjaja', u'la', u'amo\U0001f498']), Row(tweet=u'Something about the word "babygirl" \U0001f60d\U0001f60d\U0001f629\U0001f629', lang=0, words=[u'something', u'about', u'the', u'word', u'babygirl', u'\U0001f60d\U0001f60d\U0001f629\U0001f629']), Row(tweet=u'@joeljjs_perez @ceila_garcia @Wwjjle_Esas_mamis \U0001f60d\u2764\ufe0f\ufe0f', lang=1, words=[u'@joeljjs_perez'
```

Took 0 sec. Last updated by anonymous at April 22 2017, 4:20:37 PM. (outdated)

```
%pyspark
df.count() # Total tweets used (both training and test sets)
```

FINISHED

```
13207941
```

Took 32 sec. Last updated by anonymous at April 24 2017, 11:06:38 AM. (outdated)

```
%pyspark
```

READY