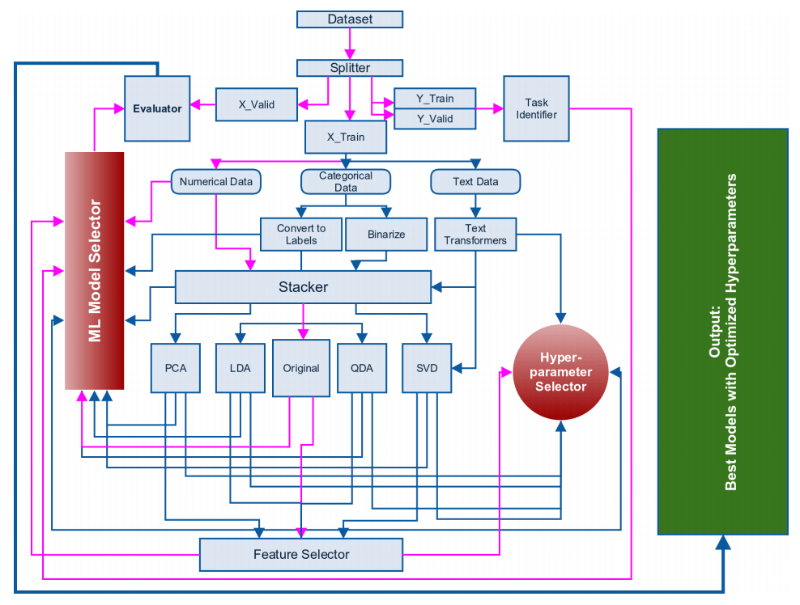
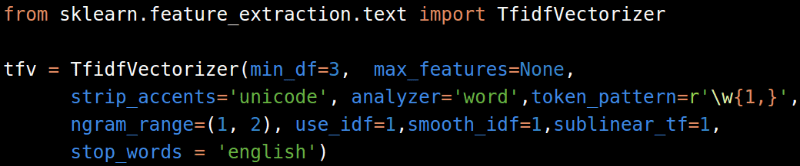
一个中等水平的数据科学家每天都要处理大量的数据。一些人说超过60%到70%的时间都用于数据清理、数据处理及格式转化，以便于在之后应用机器学习模型。这篇文章的重点便在后者—— 应用机器学习模型（包括预处理的阶段）。此文讨论到的内容来源于我参加的过的数百次的机器学习竞赛。请大家注意这里讨论的方法是大体上适用的，当然还有很多被专业人士使用的非常复杂的方法。

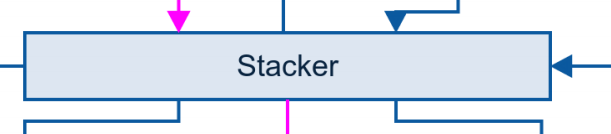
[](https://i1.wp.com/blog.kaggle.com/wp-content/uploads/2016/07/abhishek_2.png)

The TfidfVectorizer performs better than the counts most of the time and I have seen that the following parameters for TfidfVectorizer work almost all the time.

[](https://i0.wp.com/blog.kaggle.com/wp-content/uploads/2016/07/abhishek_12.png)

If you are applying these vectorizers only on the training set, make sure to dump it to hard drive so that you can use it later on the validation set.

Next, we come to the stacker module. Stacker module is not a model stacker but a feature stacker. The different features after the processing steps described above can be combined using the stacker module.

[](https://i2.wp.com/blog.kaggle.com/wp-content/uploads/2016/07/abhishek_14.png)

**FeatureUnion: composite(组合)feature spaces**

FeatureUnion把若干个transformer object组合成一个新的estimators。这个新的transformer组合了他们的输出，一个FeatureUnion对象接受一个transformer对象列表。