# 语种识别解决方案

## 语种识别方案

### Langdetect模型优势

Langdetect为基于字符的n-gram+朴素贝叶斯模型（n为1，2，3）。

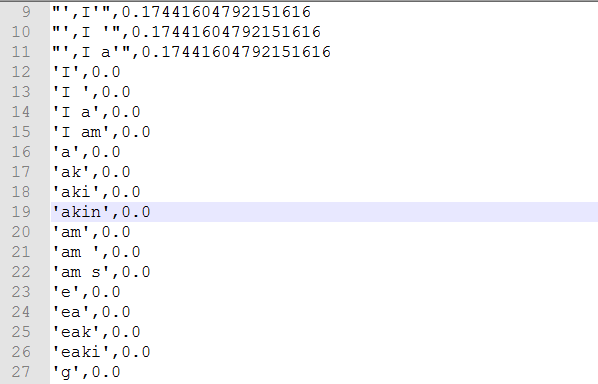
Langdetect模型对zh-ch中文简体和zh-tw中文繁体支持比较好，而langid、ldig、C2V2L模型对对zh-ch中文简体和zh-tw中文繁体支持不够好，原因为只有langdetect对zh-ch中文简体和zh-tw中文繁体做了汉字的正则化处理，而langid、ldig、C2V2L都没有。

### Langid模型优势

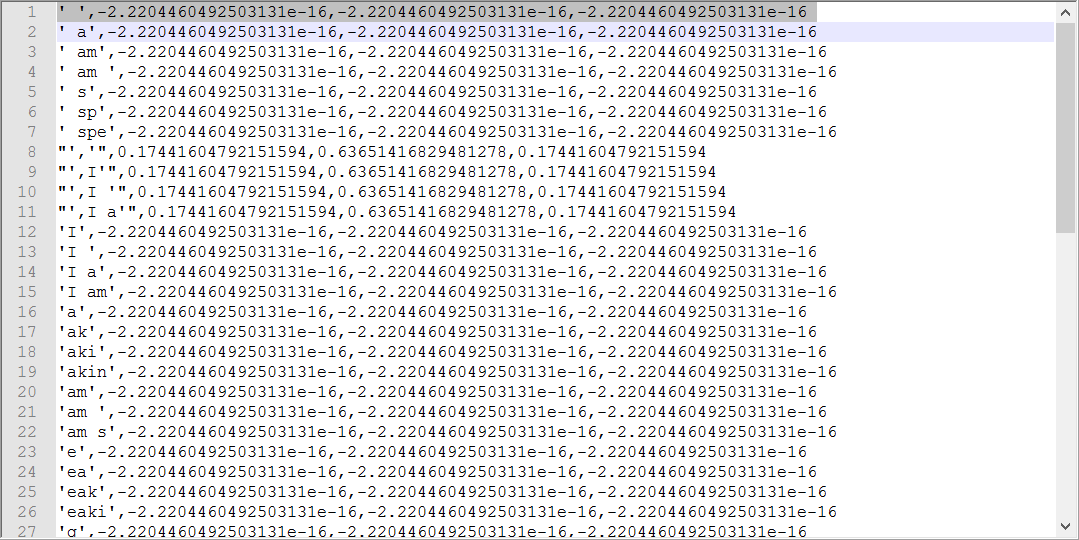
Langid为基于字节的n-gram+朴素贝叶斯模型（n为1，2，3，4），对于中文只能理解一个字符，因为1个中文占用3个字节。基于字节的n-gram可以直接提取特征不需做编码检测。

Langid模型的优势在于领域适应性。如果希望降低上述N-gram频率表的词表大小，可以选择具有较高平均互信息的特征N-gram项；且这些项对于同一语言，在不同场景下的分布应该相似，即对场景不敏感，即 (language-domain) score较低。

域的IG信息熵：

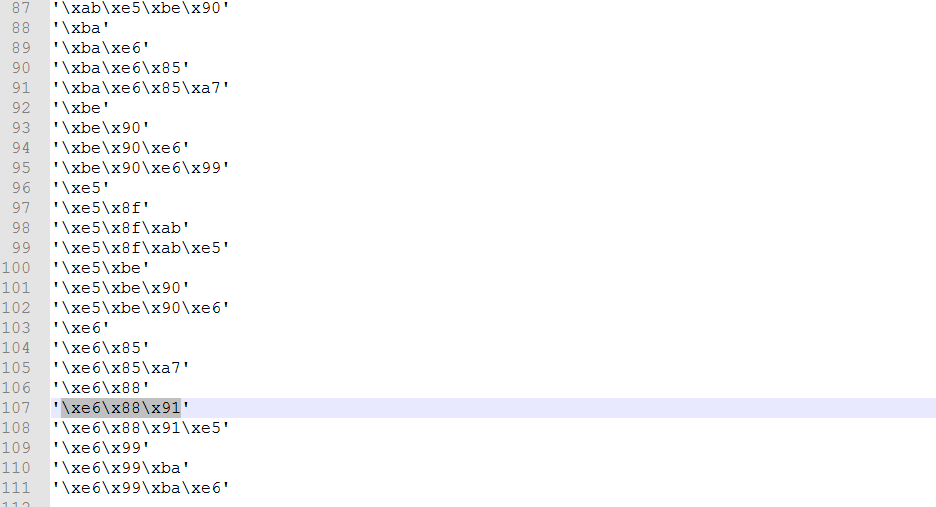


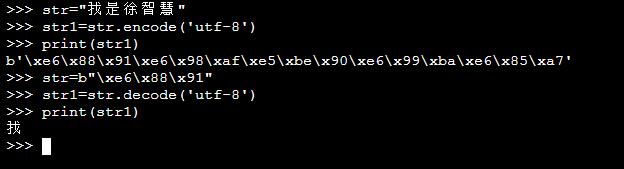
语种的IG信息熵：



下图为“我是徐智慧”使用langid生成的DFfeats文件。下图可以看出langid模型中1个中文占用3个字节。如果langid模型不区分zh-ch中文简体和zh-tw中文繁体的话，只识别zh中文的话效果没问题。

Langid无论训练集多少，最后生成的模型大小是一定的，这样会可能会导致随着训练集增大，测试集准确率下降。





### Ldig模型优势

Ldig对推特短文本支持比较好，使用最长字串+逻辑回归算法。目前只支持带有空格的语种。

### C2V2L模型优势

C2V2L模型使用CNN+BiLSTM模型，可以对文本中每个单词识别出相应的语种。测试C2V2L没有发现它比其他模型有很大的优点，并且C2V2L模型对zh-ch中文简体和zh-tw中文繁体支持不够好。

### 语种识别方案

对于语种识别，应该分应用情况使用不同的模型，如对篇章长文本语种识别、推特短文本语种识别、文本中带有多种语言的文本识别，分别使用相应的模型可以达到最佳性能。

篇章长文本语种识别使用langdetect模型比较好。

推特短文本语种识别使用ldig模型比较好。

如果不知道文本的领域使用通用语种识别模型langid比较好。

如果文本中存在多种语言则使用C2V2L模型比较好。

并且langdetect、ldig、langid、C2V2L这四个模型本身也有需要改进的地方。

Langdetect需要对一些相近的语种根据语种的个性增加相应的正则化。

Ldig、langid、C2V2L则需要增加对zh-ch中文简体和zh-tw中文繁体的处理，如增加与langdetect一样的正则化处理。

Ldig还需要对不带空格的语种进一步调试，达到良好的性能。

语种类别挑选方面也需要做一定的处理，如去掉iw希伯来语，因为谷歌翻译iw希伯来语和he希伯来文翻译出来的文本一样。