Bi-weekly report

Pan Yiqiao 2015 4.13

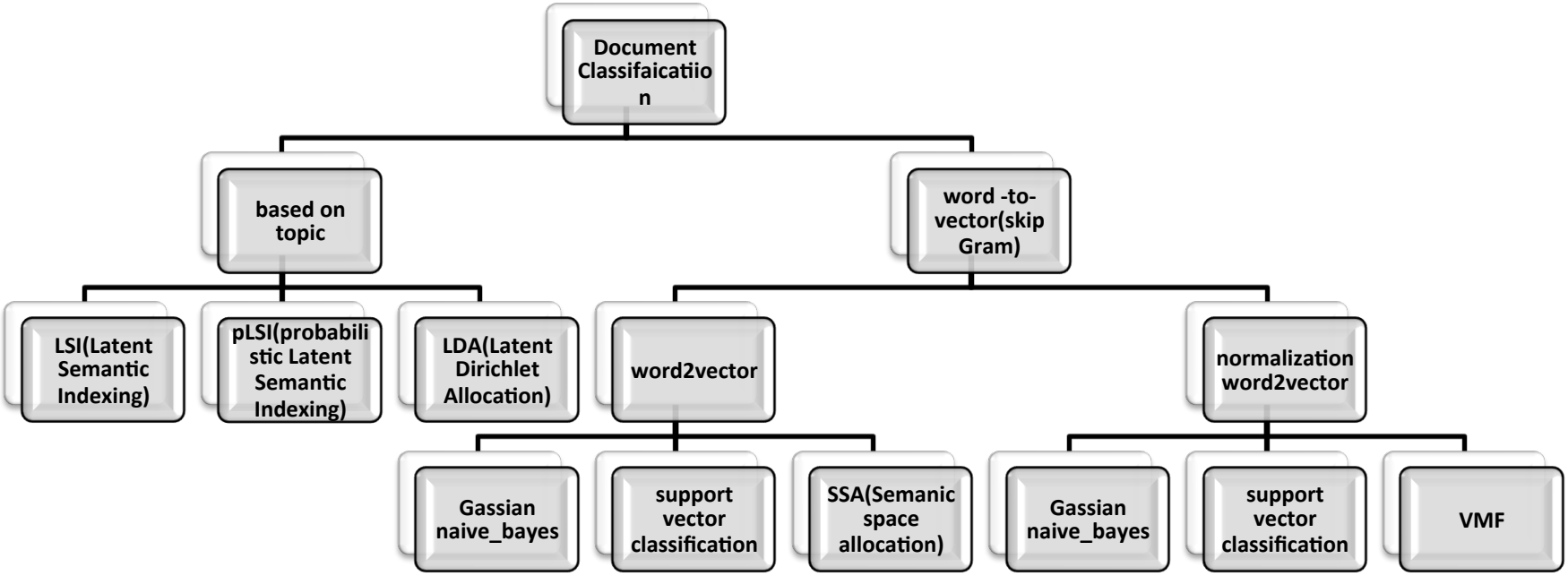
Part1 Learn tools

1） Python

2） java

3） LINUX

Part2 Learn document classification process



Part3 Test pre_procession

1) The article classified（test.py）（Corpus database：Reuters）
2) Match the test and train set（task0.py）
3) Tokenizer（task1.py using tokennizer.perl）
4) Remove duplicate sentence ， Remove non-English word ， Remove the empty text （task2.py）
5) average pooling (average.py)

A total of 3409 articles, 64 topics

	Word2vector	Norm-word2vector	Norm-word2vector（norm）
NB-100	53.0%	49.4%	49.4%
NB-200	52.7%	48.3%	48.3%
NB-300	51.9%	46.1%	46.1%
NB-400	51.8%	48.6%	48.6%
NB-500	52.6%	49.4%	49.4%
SVM-100	74.1%	49.5%	49.5%
SVM-200	71.7%	49.3%	49.3%
SVM-300	71.0%	37.5%	37.5%
SVM-400	70.0%	31.2%	31.2%
SVM-500	69.1%	31.2%	31.2%

Part 4 Learning materials

1）基于词向量的文本分类方法

2）Document Classification with Distributions of Word Vectors

3）Neural Networks

http://ufldl.stanford.edu/wiki/index.php/Neural_Networks

4）Network open class for CS229 Machine Learning

http://cs229.stanford.edu