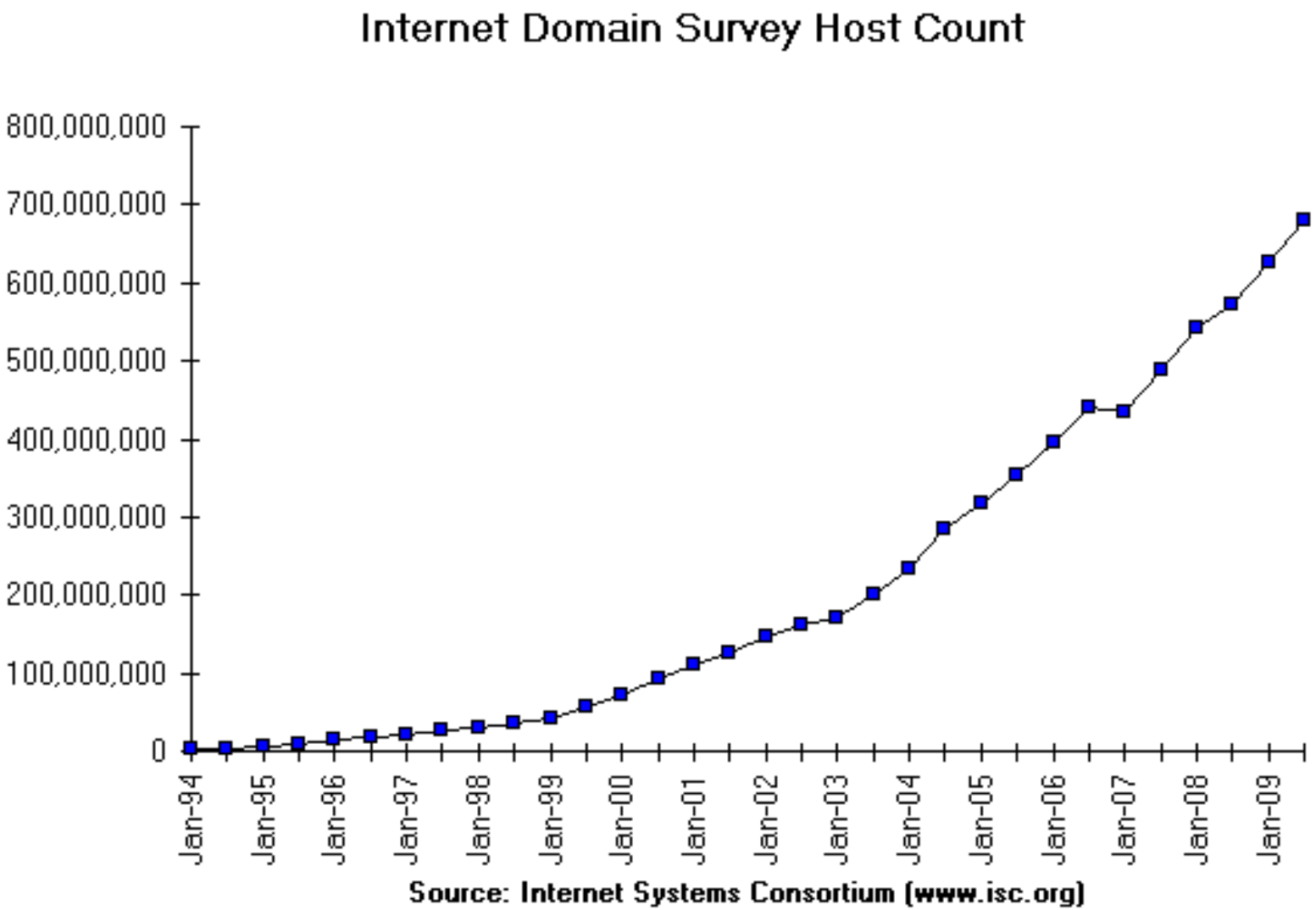
Automated Text Classification in the DMOZ Hierarchy

Lachlan Henderson

November 6, 2009


Internet Domain Survey Host Count
800,000,000
700,000,000
600,000,000
500,000,000
400,000,000
300,000,000
200,000,000
100,000,000
0
Jan-94
Jan-95
Jan-96
Jan-97
Jan-98
Jan-99
Jan-00
Jan-01
Jan-02
Jan-03
Jan-04
Jan-05
Jan-06
Jan-07
Jan-08
Jan-09
Source: Internet Systems Consortium (www.isc.org)

How can can large amounts of data be organised?

Hierarchy Level	Topic Count	Document Count	Topic (non-empty) Count
0	1	0	0
1	17	88	3
2	656	6421	335
3	7764	128888	5974
4	39946	472830	33364
5	89934	778173	77495
6	109847	785585	89268
7	167528	737462	128985
8	165460	663991	125173
9	107407	520592	86859
10	56265	332234	50205
11	15903	147529	15097
12	3906	34436	3750
13	648	5674	537
Total	765282	4613903	617045

Table: Frequency of ODP Topics and Documents at Hierarchy Level

Level 2 Label	Document Count
Top/Regional/North_America	694831
Top/World/Deutsch	501512
Top/Regional/Europe	285859
Top/World/Français	233029
Top/World/Italiano	203114
Top/World/Japanese	184078
Top/World/Español	163367
Top/Society/Religion_and_Spirituality	103862
Top/World/Nederlands	97338
Top/Arts/Music	80618
Total	2547608

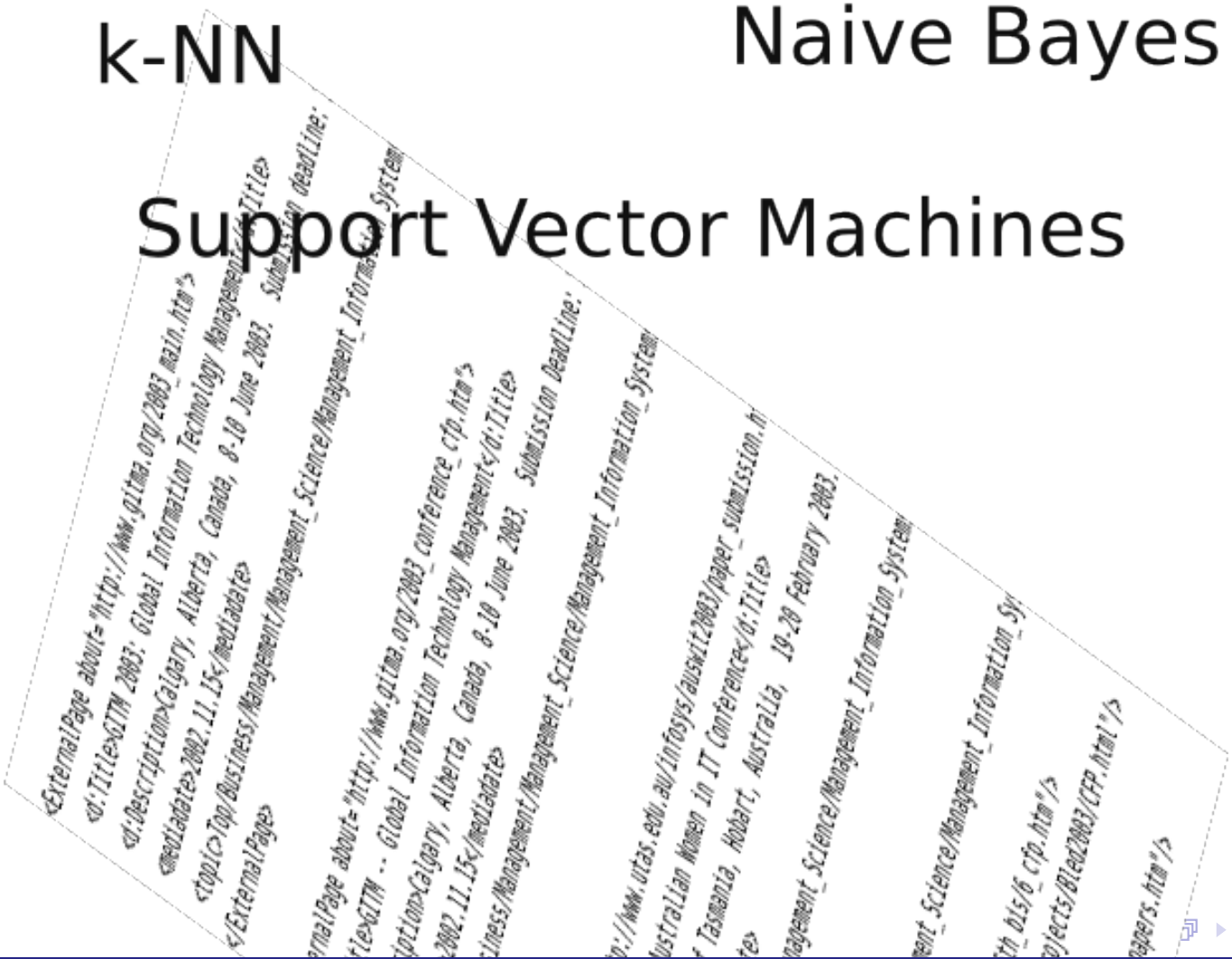
Table: Top 10 ODP Level-2 Labels And Document Counts

Centroid Classifier

k-NN

Naive Bayes

Support Vector Machines



Supervised learning requires labeled examples.

Use the description data as a proxy for a webpage!

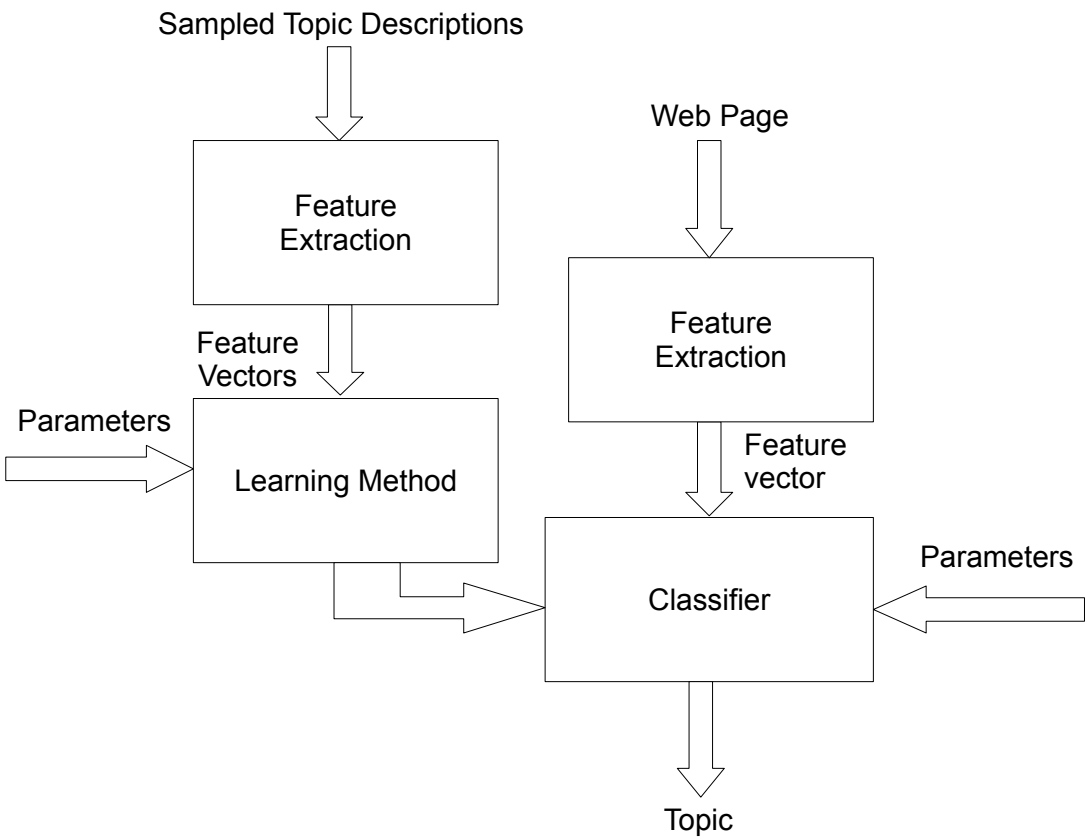
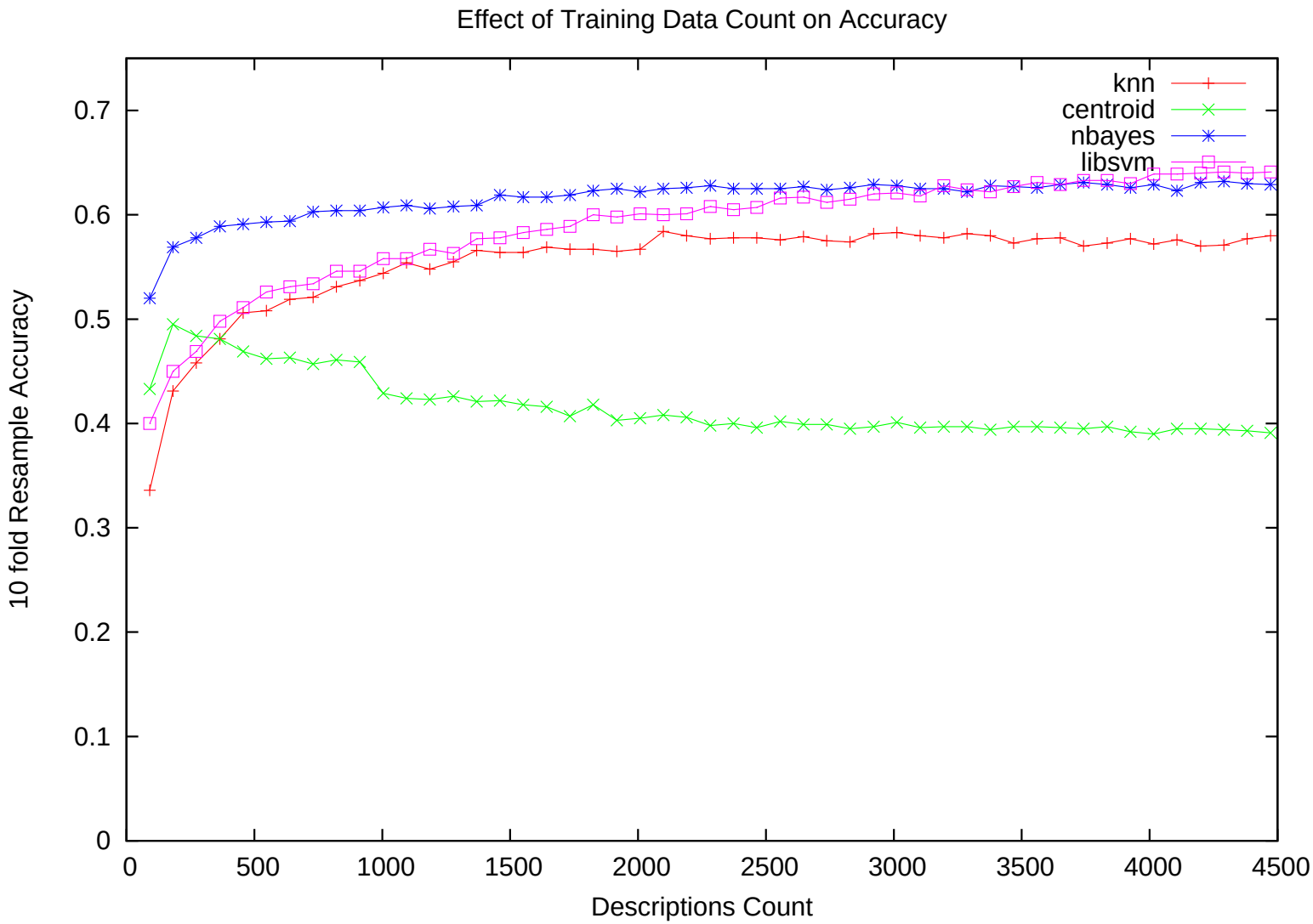
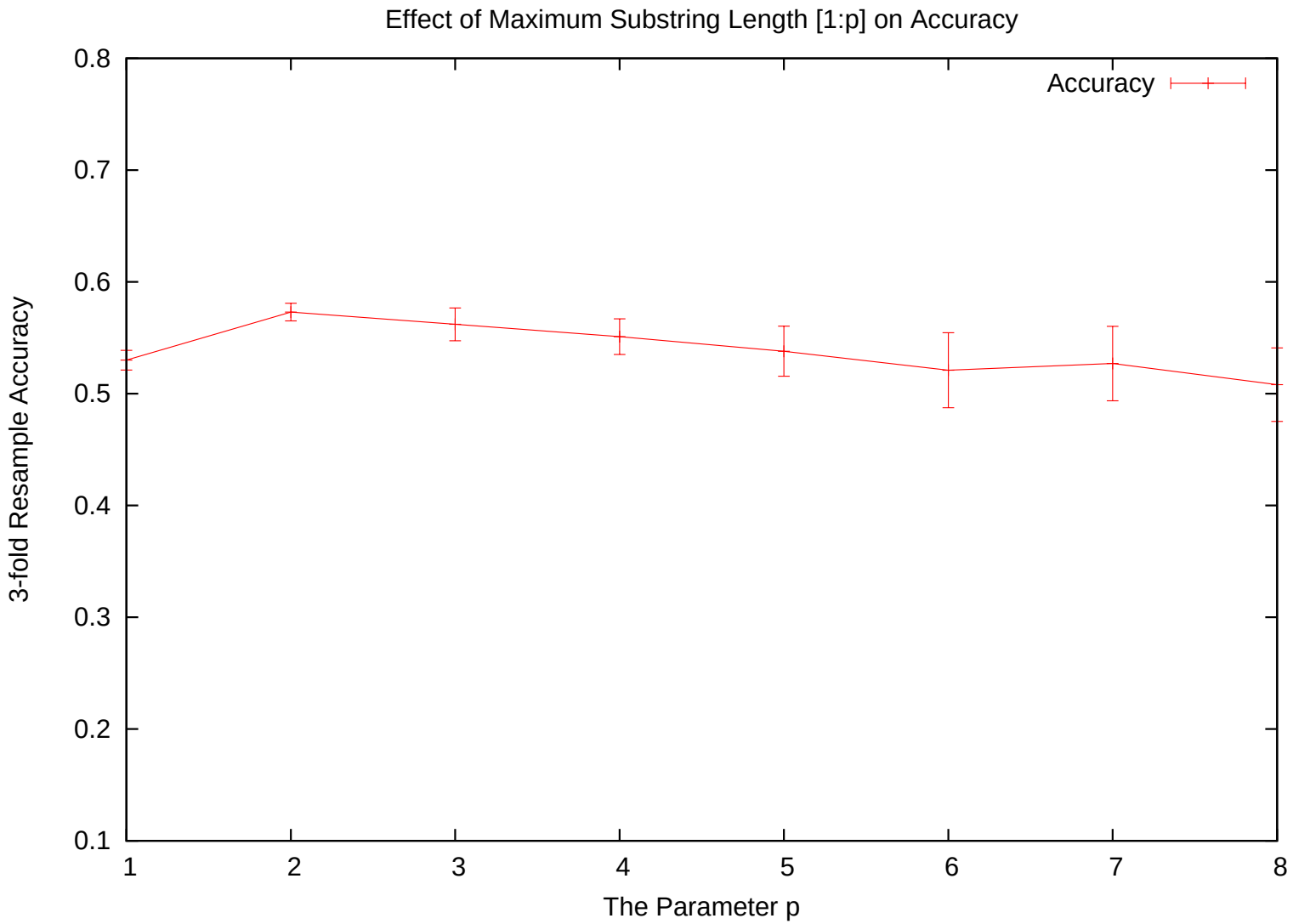


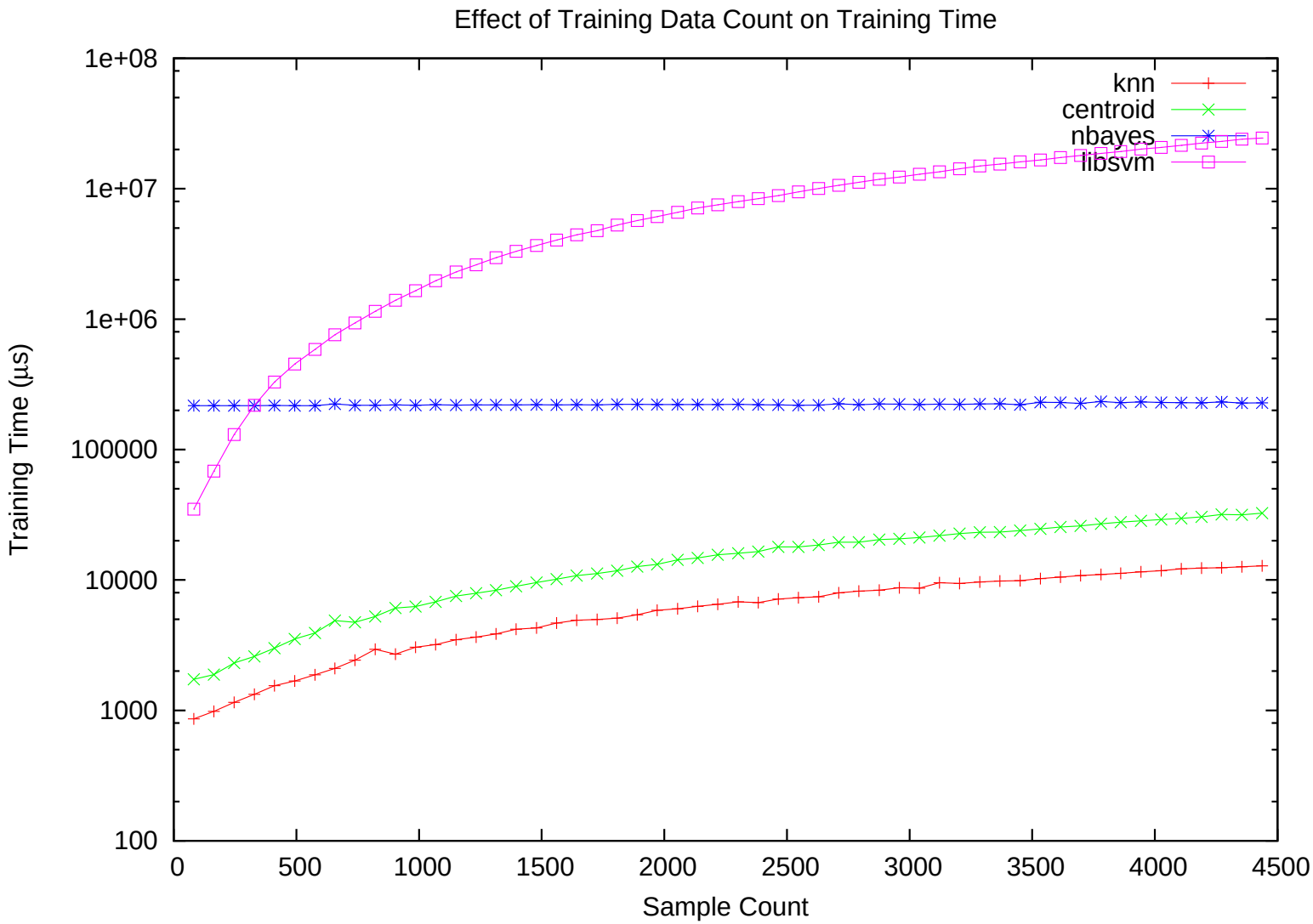
Figure: Web Page Classification Process

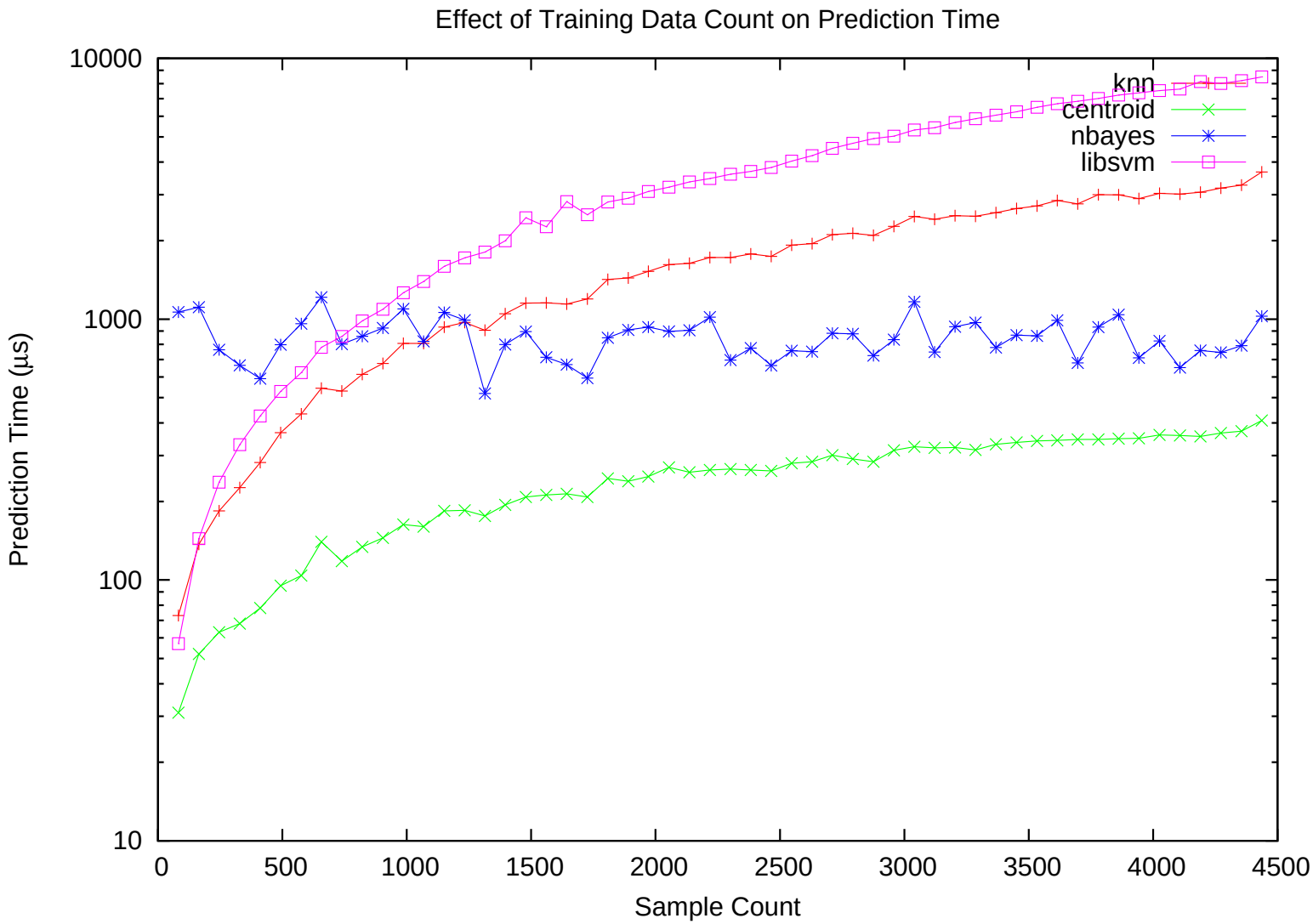
File	knn		centroid		nbayes		libsvm	
	Det.	Ref.	Det.	Ref.	Det.	Ref.	Det.	Ref.
oh0.wc.arff	87.1±0.2	84.4	90.7±0.2	89.3	89.2±0.2	89.1	89.7±0.1	-
oh5.wc.arff	84.1±0.1	85.6	87.0±0.2	88.2	84.5±0.2	87.1	89.9±0.3	-
oh10.wc.arff	76.8±0.3	77.5	81.0±0.3	85.3	82.0±0.2	81.2	81.5±0.2	-
oh15.wc.arff	79.3±0.3	81.7	82.8±0.3	87.4	81.6±0.3	84.0	83.9±0.1	-

Table: Classifier 10-fold Cross Validation % Accuracy on the OHSUMED Collection Compared With Experimental Results of Karypis et. el. (95 % Confidence Limits)


Effect of Training Data Count on Accuracy
10 fold Resample Accuracy
Descriptions Count
knn
centroid
nbayes
libsvm
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
500
1000
1500
2000
2500
3000
3500
4000
4500


Effect of Maximum Substring Length [1:p] on Accuracy
Accuracy
3-fold Resample Accuracy
The Parameter p
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
1
2
3
4
5
6
7
8


Effect of Training Data Count on Training Time
Training Time (µs)
Sample Count
knn
centroid
nbayes
libsvm
1e+08
1e+07
1e+06
100000
10000
1000
100
0
500
1000
1500
2000
2500
3000
3500
4000
4500


Effect of Training Data Count on Prediction Time
Prediction Time (µs)
Sample Count
10000
1000
100
10
0
500
1000
1500
2000
2500
3000
3500
4000
4500
knn
centroid
nbayes
libsvm

Feature extraction/selection and algorithm are closely tied.

Large scale text classification is difficult.

Language independent methods are very important.

Thank You