Classification of Arabic Documents

Project Final Presentation – Dec. 6, 2012

CS 5604 : Information Storage and Retrieval

Instructor: Prof. Edward Fox

GTA : Tarek Kanan

ProjArabic Team

Ahmed Elbery

Outline

- Arabic documents classification: Motivation
- Arabic documents classification: Challenges
- Model
- Model Details
- Results and Evaluation

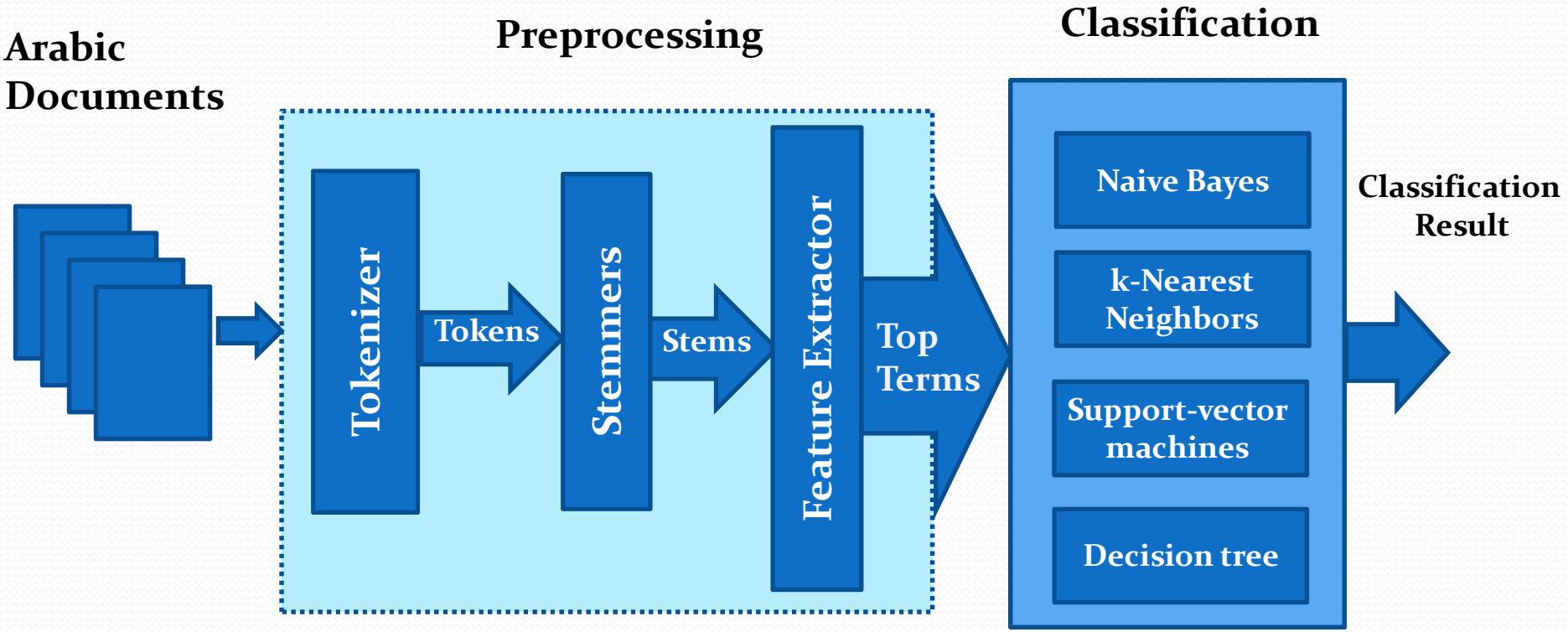
Arabic documents classification: Motivation

- Rich set of Arabic documents
- Now > 65M Internet users of Arabic
- Arabic NLP needed for increasing Arabic internet content

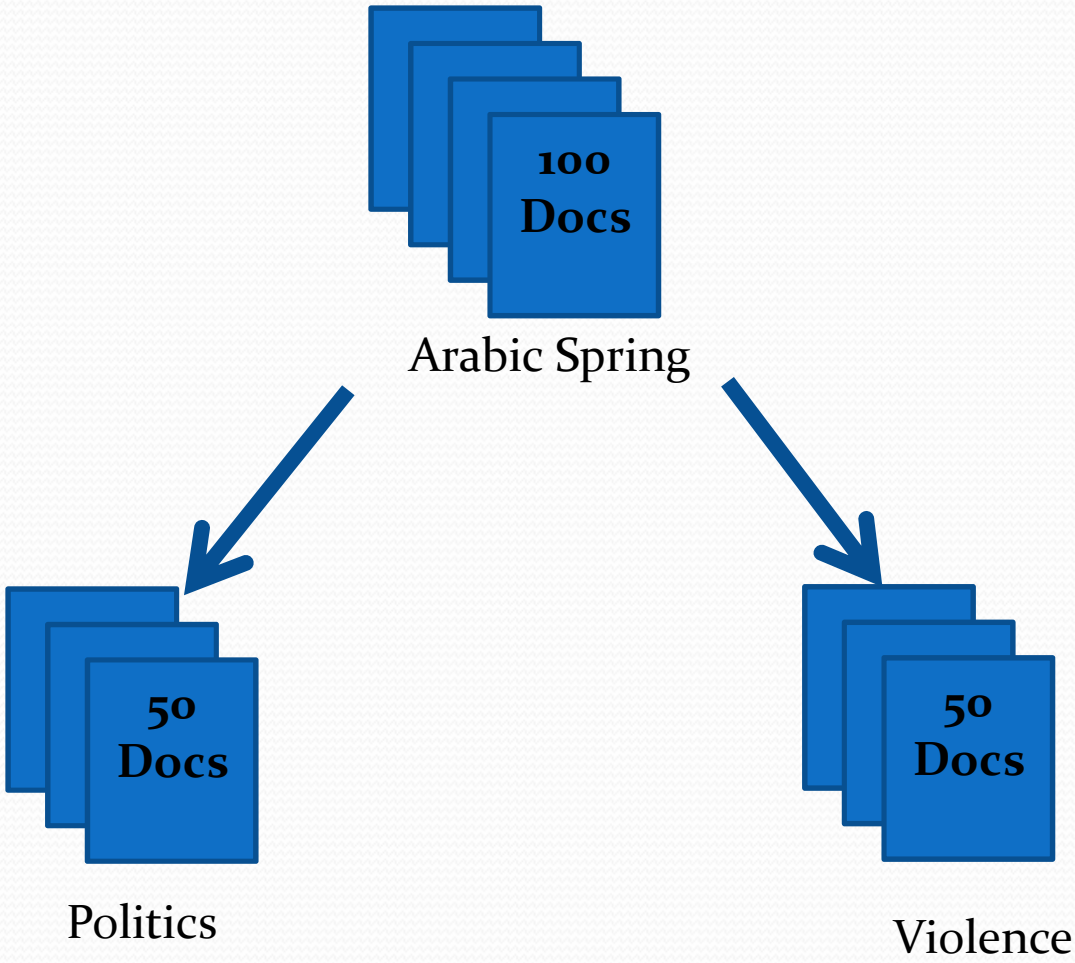
Arabic documents classification: Challenges

- ***Techniques built for English language processing may not apply to Arabic because:-***
  - *Arabic is very rich with complex morphology*
  - *Arabic has a very different and difficult syntax and grammar*

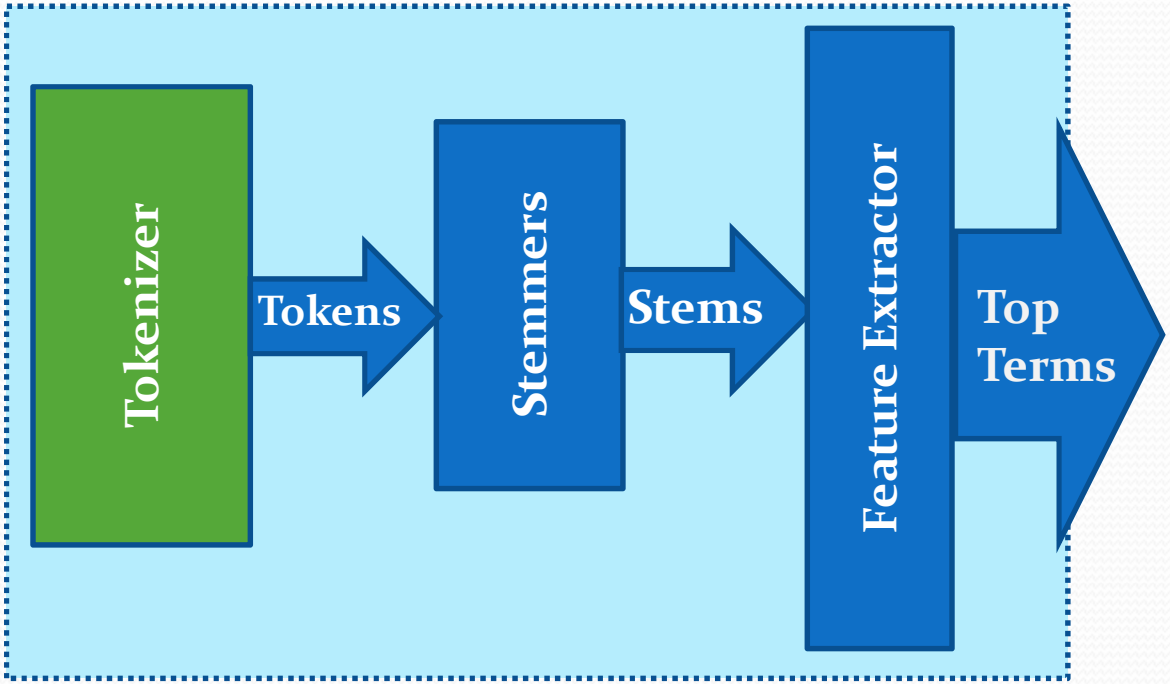
Project model



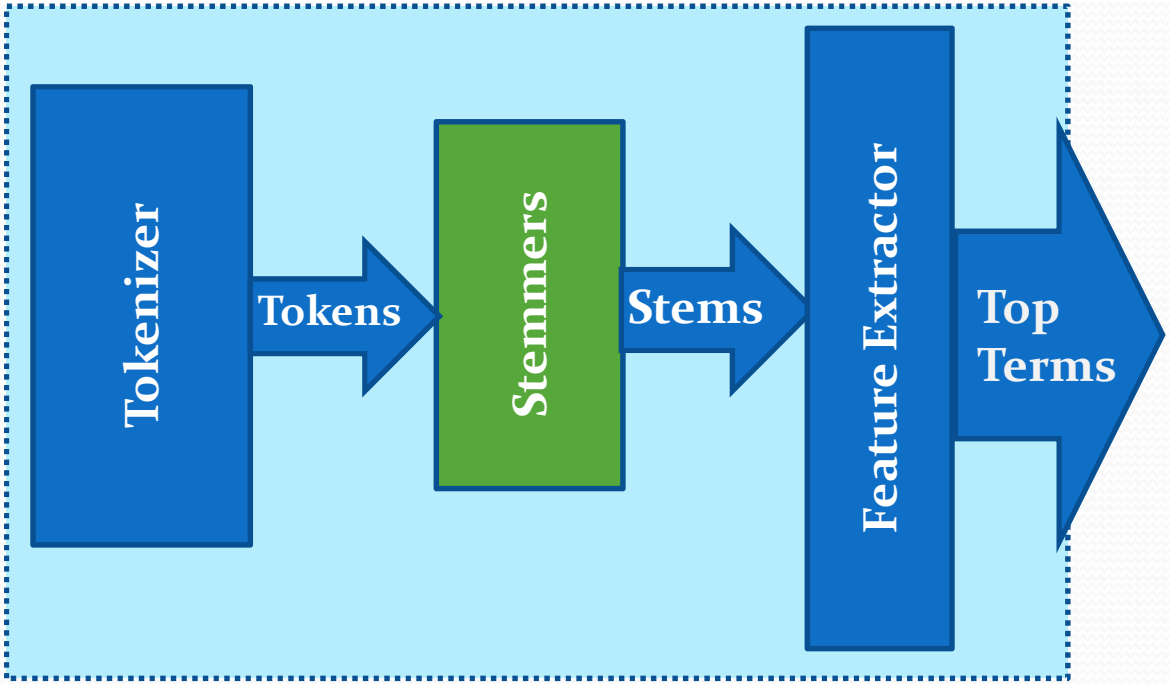
Data Set



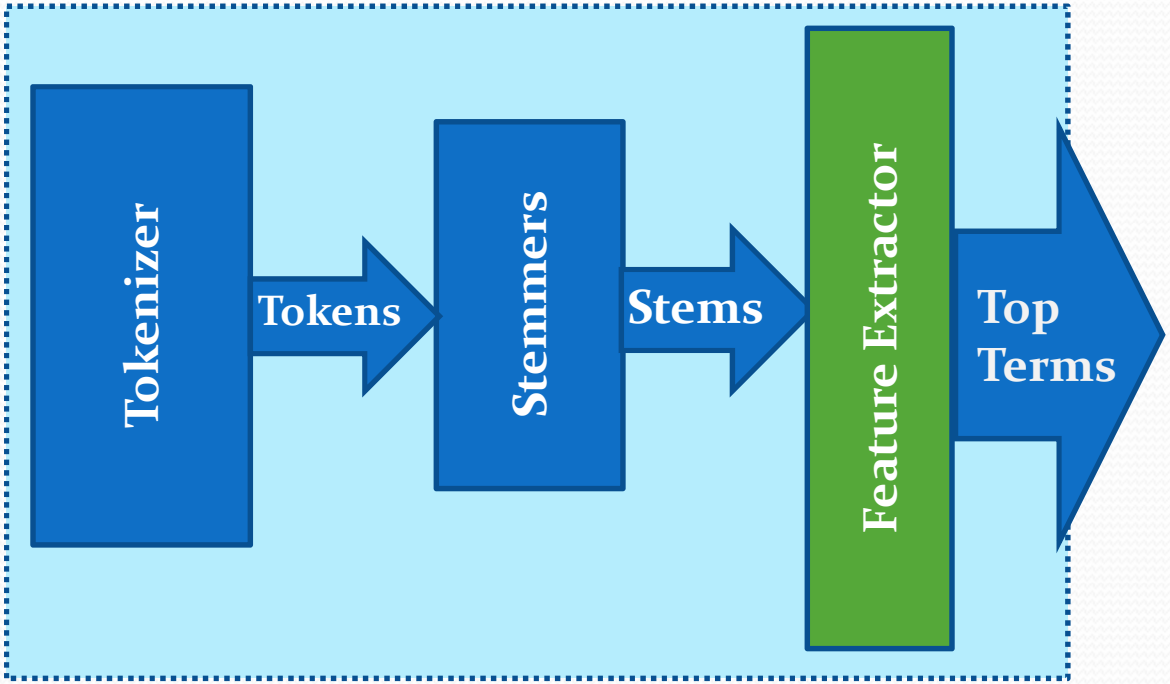
Preprocessing



Preprocessing



Preprocessing



Example

Doc P-1 :

Systems Politics nation area Liberty International Politics Government

Doc P-2

Politics Systems nation area Kill nation Politics Government

Doc V-1

Violence Systems Weapon Weapon Militias Violence Kill Government Burn

Doc V-1

Burn Systems Weapon Militias Violence Kill Kill Government

Example- Cont.

		'Burn'	'Government'	'International'	'Kill'	'Liberty'	'Militias'	'Politics'	'Systems'	'Violence'	'Weapon'	'area'	'nation'
Term count	Doc P-1	0	1	1	0	1	0	2	1	0	0	1	1
	Doc V-1	1	1	0	1	0	1	0	1	2	2	0	0
	Doc V-2	1	1	0	2	0	1	0	1	1	1	0	0
tf	Doc P-1	0	0.125	0.125	0	0.125	0	0.25	0.125	0	0	0.125	0.125
	Doc V-1	0.1111	0.1111	0	0.1111	0	0.1111	0	0.1111	0.2222	0.2222	0	0
	Doc V-2	0.125	0.125	0	0.25	0	0.125	0	0.125	0.125	0.125	0	0
df		2	3	1	2	1	2	1	3	2	2	1	1
idf		0.4055	0	1.0986	0.4055	1.0986	0.4055	1.0986	0	0.4055	0.4055	1.0986	1.0986
tf.idf	Doc P-1	0	0	0.1373	0	0.1373	0	0.2747	0	0	0	0.1373	0.1373
	Doc V-1	0.0451	0	0	0.0451	0	0.0451	0	0	0.0901	0.0901	0	0
	Doc V-2	0.0507	0	0	0.1014	0	0.0507	0	0	0.0507	0.0507	0	0
sum(tf.idf)		0.0957	0	0.1373	0.1464	0.1373	0.0957	0.2747	0	0.1408	0.1408	0.1373	0.1373

Example- Cont.

		'Burn'	'Government'	'International'	'Kill'	'Liberty'	'Militias'	'Politics'	'Systems'	'Violence'	'Weapon'	'area'	'nation'
Term count	Doc P-1	0	1	1	0	1	0	2	1	0	0	1	1
	Doc V-1	1	1	0	1	0	1	0	1	2	2	0	0
	Doc V-2	1	1	0	2	0	1	0	1	1	1	0	0
tf	Doc P-1	0	0.125	0.125	0	0.125	0	0.25	0.125	0	0	0.125	0.125
	Doc V-1	0.1111	0.1111	0	0.1111	0	0.1111	0	0.1111	0.2222	0.2222	0	0
	Doc V-2	0.125	0.125	0	0.25	0	0.125	0	0.125	0.125	0.125	0	0
df		2	3	1	2	1	2	1	3	2	2	1	1
idf		0.4055	0	1.0986	0.4055	1.0986	0.4055	1.0986	0	0.4055	0.4055	1.0986	1.0986
tf.idf	Doc P-1	0	0	0.1373	0	0.1373	0	0.2747	0	0	0	0.1373	0.1373
	Doc V-1	0.0451	0	0	0.0451	0	0.0451	0	0	0.0901	0.0901	0	0
	Doc V-2	0.0507	0	0	0.1014	0	0.0507	0	0	0.0507	0.0507	0	0
sum(tf.idf)		0.0957	0	0.1373	0.1464	0.1373	0.0957	0.2747	0	0.1408	0.1408	0.1373	0.1373

Example- Cont.

	'International'	'Kill'	'Liberty'	'Politics'	'Violence'	'Weapon'
Doc P-1	0.137327	0	0.137327	0.274653	0	0
Doc V-1	0	0.045052	0	0	0.090103	0.090103
Doc V-2	0	0.101366	0	0	0.050683	0.050683

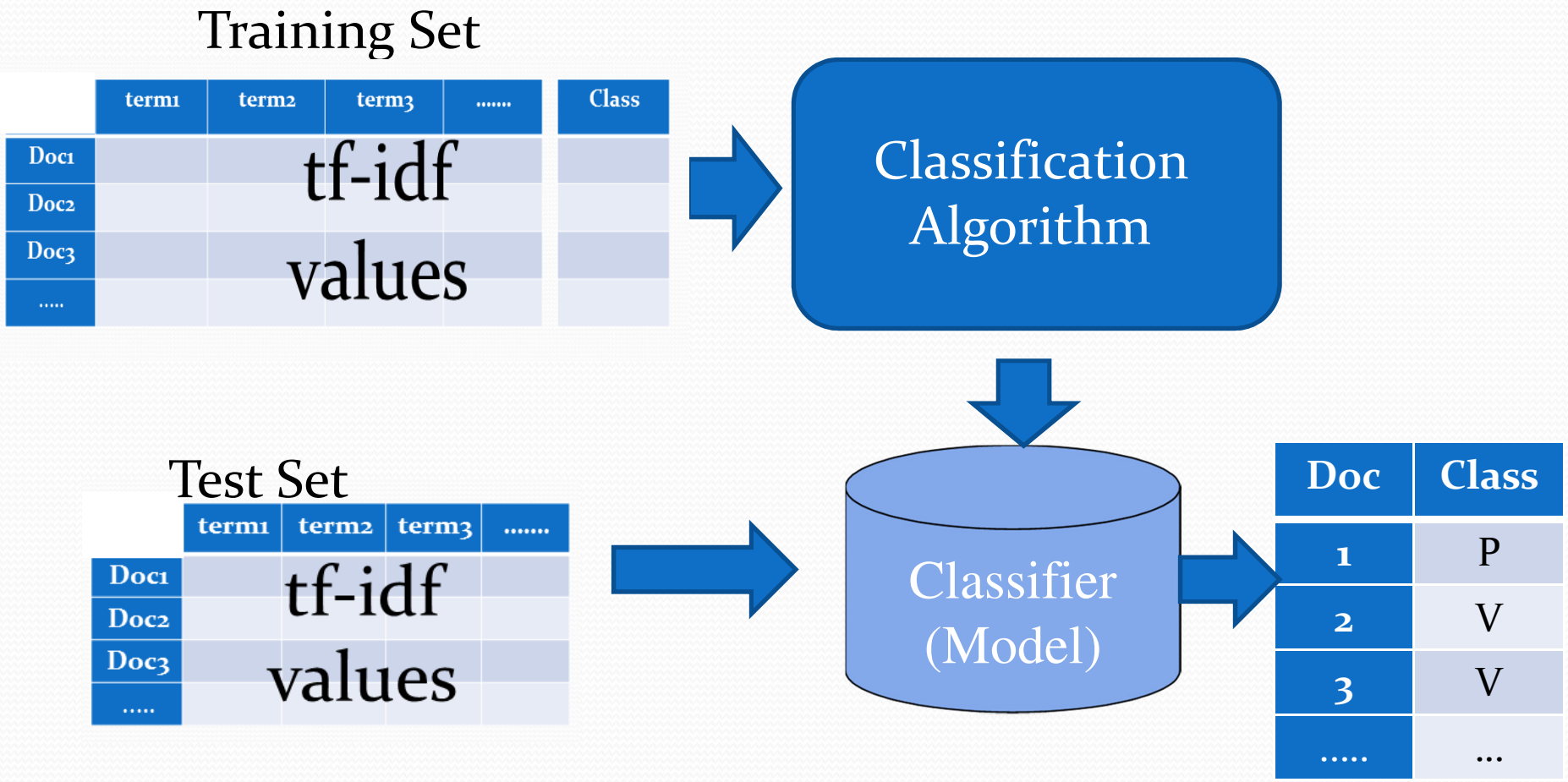
Preprocessing

The output matrix

	term1	term2	term3	…….	Class
Doc1					
Doc2					
Doc3					
…..					

tf-idf values

Classifier


Training Set
term1
term2
term3
.....
Class
Doc1
Doc2
Doc3
.....
tf-idf
values
Classification Algorithm
Test Set
term1
term2
term3
.....
Doc1
Doc2
Doc3
.....
tf-idf
values
Classifier (Model)
Doc
Class
1
P
2
V
3
V
.....
...

Results and Evaluation

Accuracy

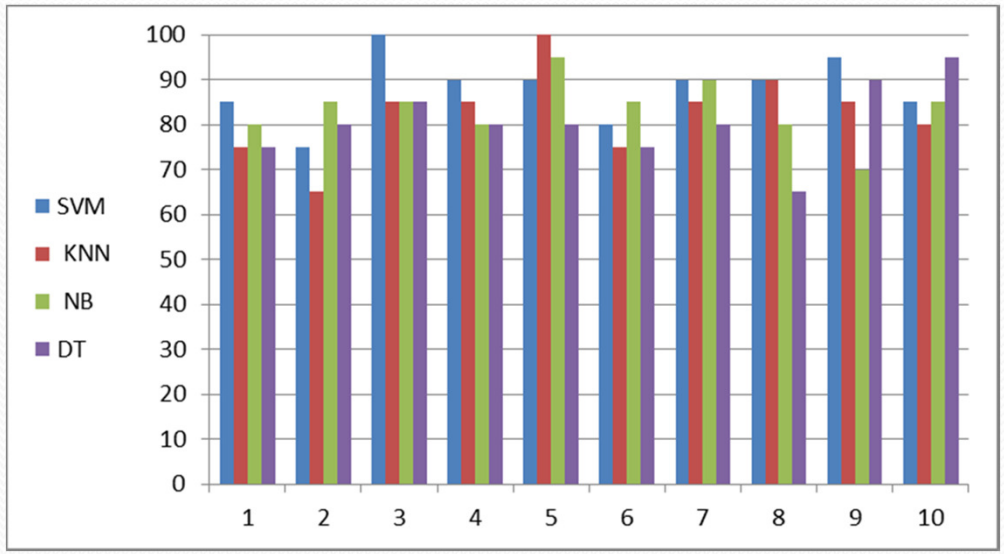
- 100 Docs (50+50)
- 10 times
- 80% training
- 20% test

Accuracy

	SVM	KNN	NB	DT
1	85	75	80	75
2	75	65	85	80
3	100	85	85	85
4	90	85	80	80
5	90	100	95	80
6	80	75	85	75
7	90	85	90	80
8	90	90	80	65
9	95	85	70	90
10	85	80	85	95
Av.	89.5	89	86.5	83

Results and Evaluation

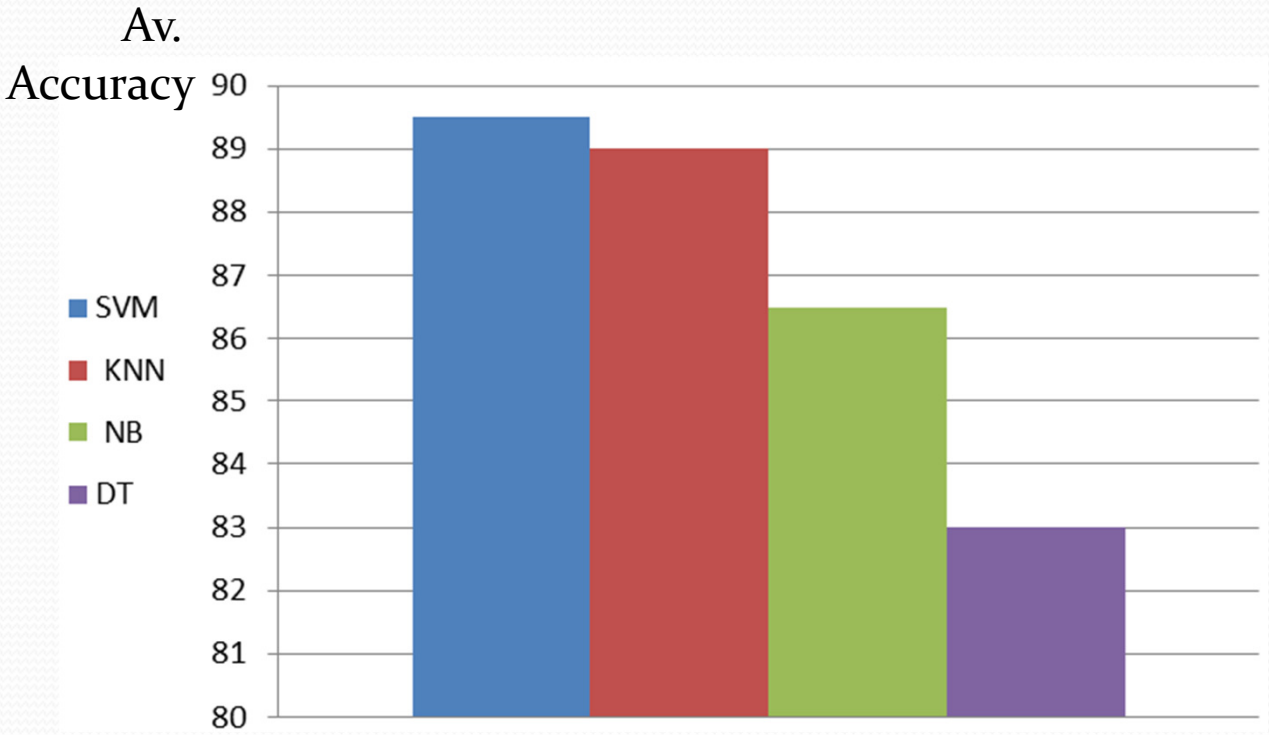
Accuracy



Correlation coefficient

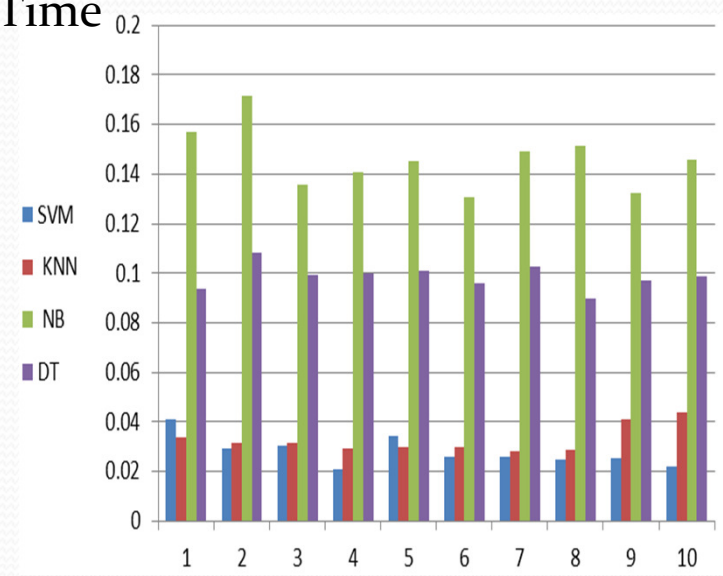
	SVM	*KNN*	*NB*	*DT*
SVM	1			
KNN	0.695182	1		
NB	-0.18592	0.240441	1	
DT	0.205563	-0.05273	-0.08491	1

Results and Evaluation

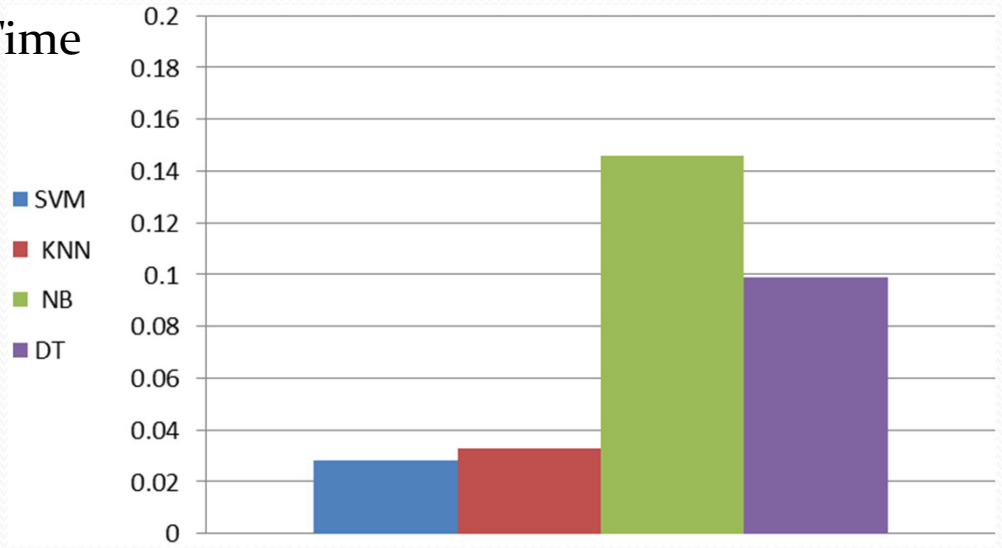


Results and Evaluation

Time



Av. Time



	SVM	*KNN*	*NB*	*DT*
SVM	1			
KNN	-0.13369	1		
NB	0.335572	-0.12174	1	
DT	-0.08792	-0.05281	0.370353	1

Future work

- Test the different parameters of the classifier
  - Feature ratio
  - Feature selection parameters

- Classifier parameters.
- Statistically analysis the results.

Ahmed Elbery