Focused Crawler

Wil Collins, Will Dickerson

Client: Mohamed Magdy and CTRnet

Vector Space Modeling

Models: tf-idf, LSI

D_1="One apple a day keeps the doctor away"

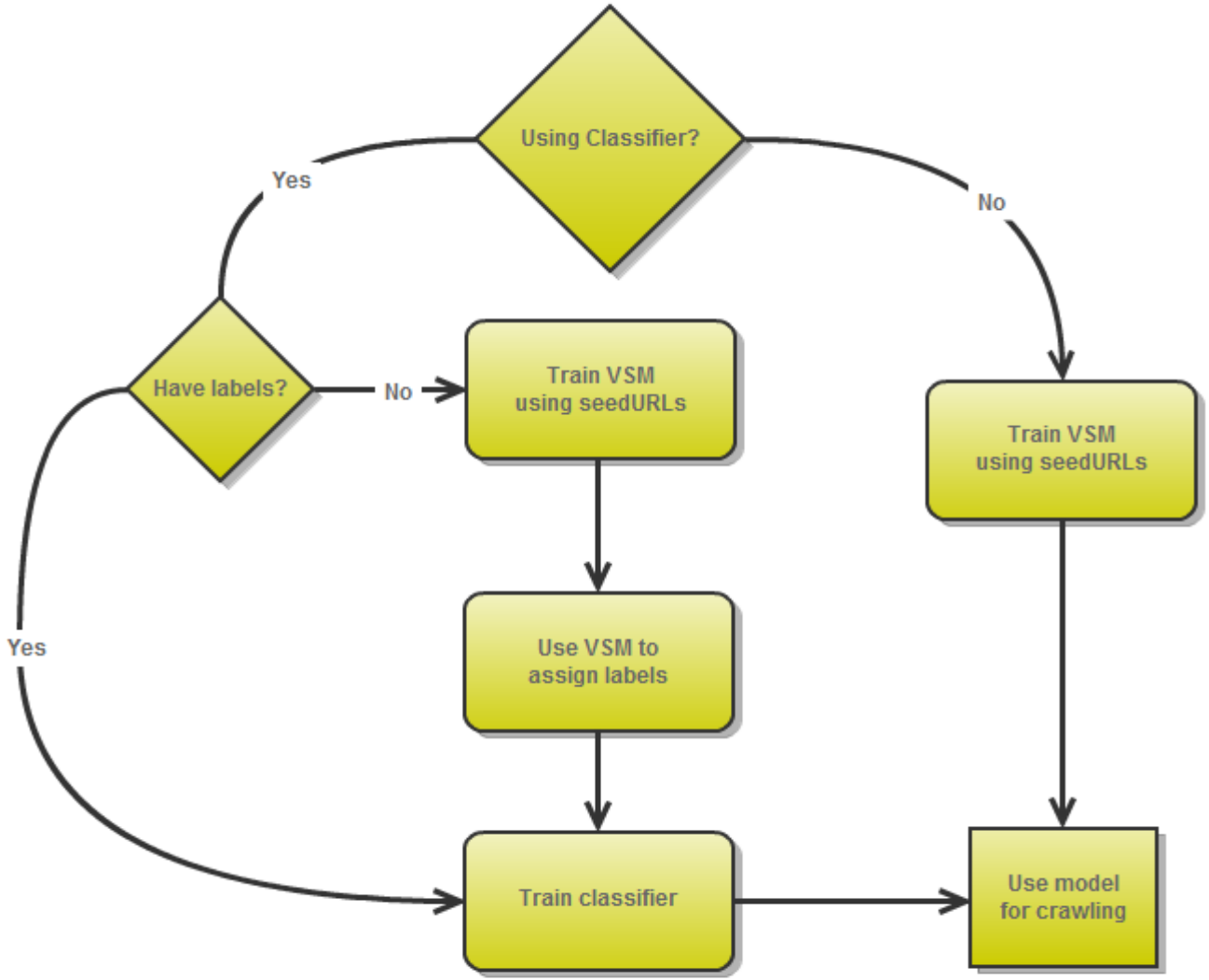
D_2="This doctor enjoys eating anything orange"

D_3="You can't compare an apple and an orange"

	D1	D2	D3
"Apple"	0.707	0	0.707
"Doctor"	0.707	0.707	0
"Orange"	0	0.707	0.707

$$\vec{D}_1 \bullet \vec{D}_2 = 0.5$$

Model Training Process



Current State

- Improved Relevance Calculations
- Enabled use of Naive Bayesian and SVM classifiers
- Limited use of tf-idf and lsi as labellers
- Configuration and documentation

Current State

- Error Handling
- Improved Search Algorithm
- Real-time model updates

	actual class (observation)	
predicted class (expectation)	**tp** (true positive) Correct result	**fp** (false positive) Unexpected result
	fn (false negative) Missing result	**tn** (true negative) Correct absence of result

Recall: $\frac{tp}{tp + fn}$

Precision: $\frac{tp}{tp + fp}$

- Initial: Recall: .4821 Precision: .2842
- Final: Recall: .4252 Precision: .9908

Demo Video

CTRnet Collection

http://wilcollins.com/ctrnet

Future Plans

- Seed generation
- Improve recall
- Speed/Memory improvements
- Better web text extraction
- Better training documents

Questions?