A Concept-based Model for Enhancing Text Categorization

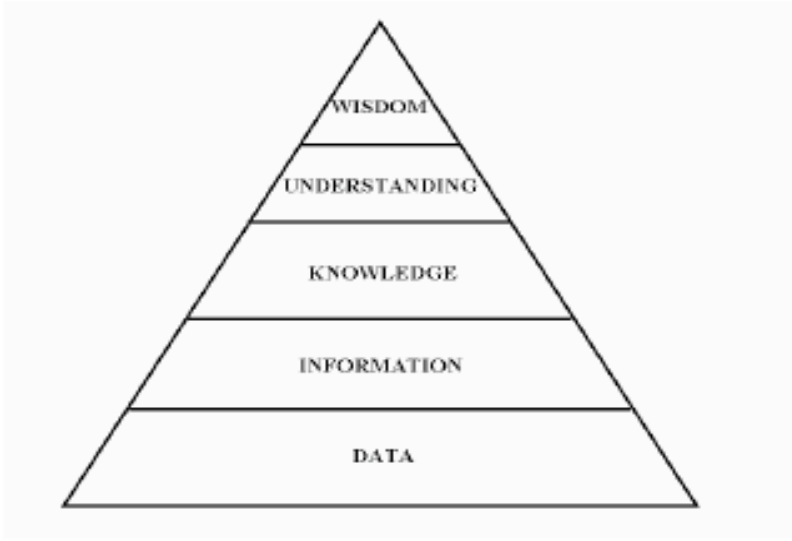
- MATH3346 - Fernando Figueiredo
-
- Research Paper by:
- Shady Shehata - Fakhri Karray - Mohamed Kamel
- Pattern Analysis and Machine Intelligence Research Group
- Electrical and Computer Engineering Department
- University of Waterloo
- Canada

Presentation Outline

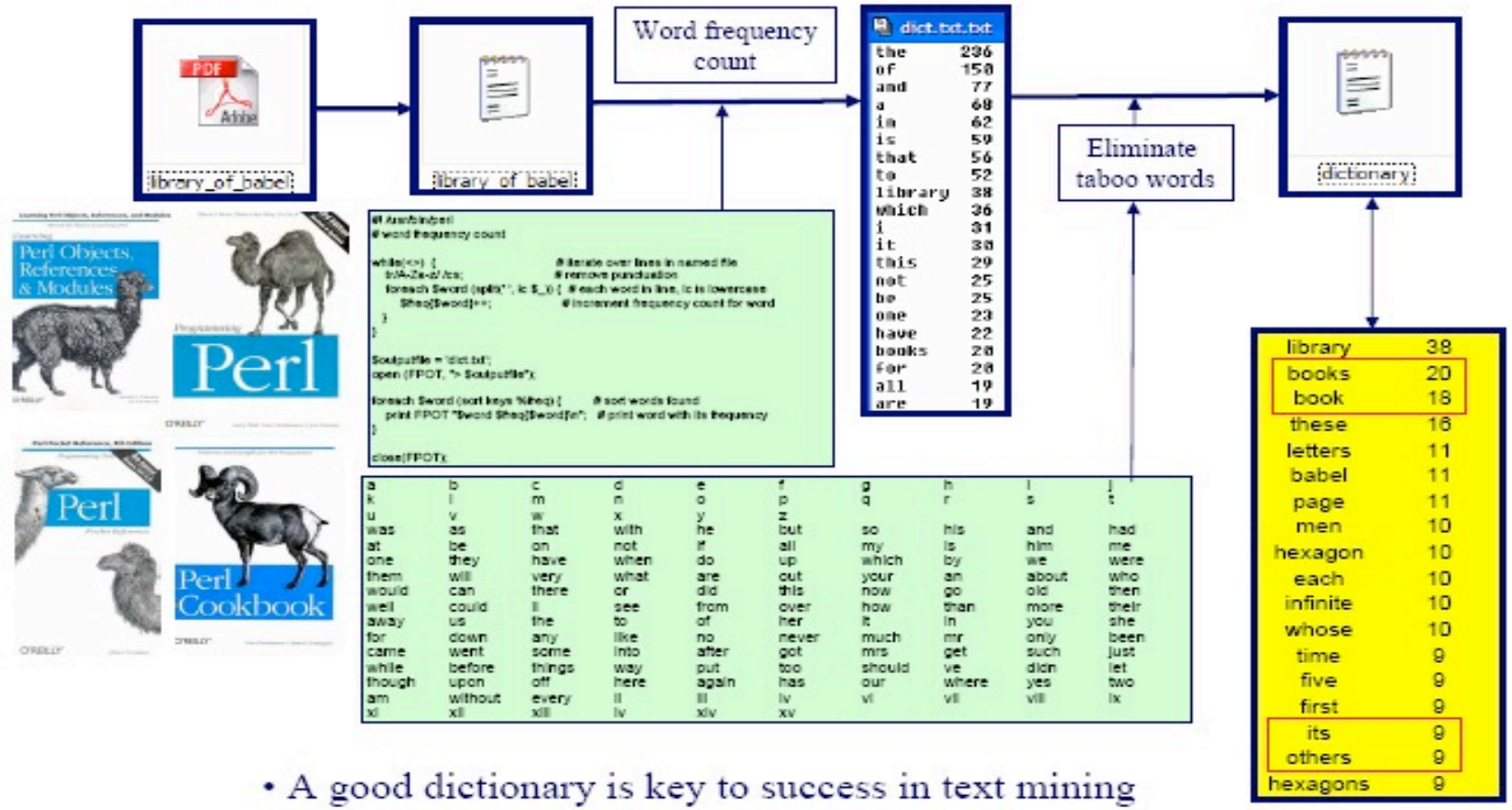
- Overview of text mining.
- Concept based Model framework.
  - Statistical Analyzer.
  - Conceptual Ontological Graph.
  - Concept Extractor.
  - Experimental results.
- Critique
- Acknowledgments
- Questions

What is Text Mining?

- Text mining is the automated retrieval of novel, interesting and possibly intelligent information from one single document or from many documents
- Combines machine learning, pattern recognition, data mining and linguistics
- Issues:
  - Text is a very compact representation
  - Texts can be considered as unstructured data
  - How to convert text into numbers?
  - Making dictionaries
  - Language and text
- Applications:
  - Text classification
  - Case study retrieval in medicine and law
  - Automated translation
  - Ontologies
  - Document identification and retrieval
  - Discovering trends and intelligence (e.g. time-based text mining)
  - Document summarization
  - Spam filtering
  - Author/Language identification

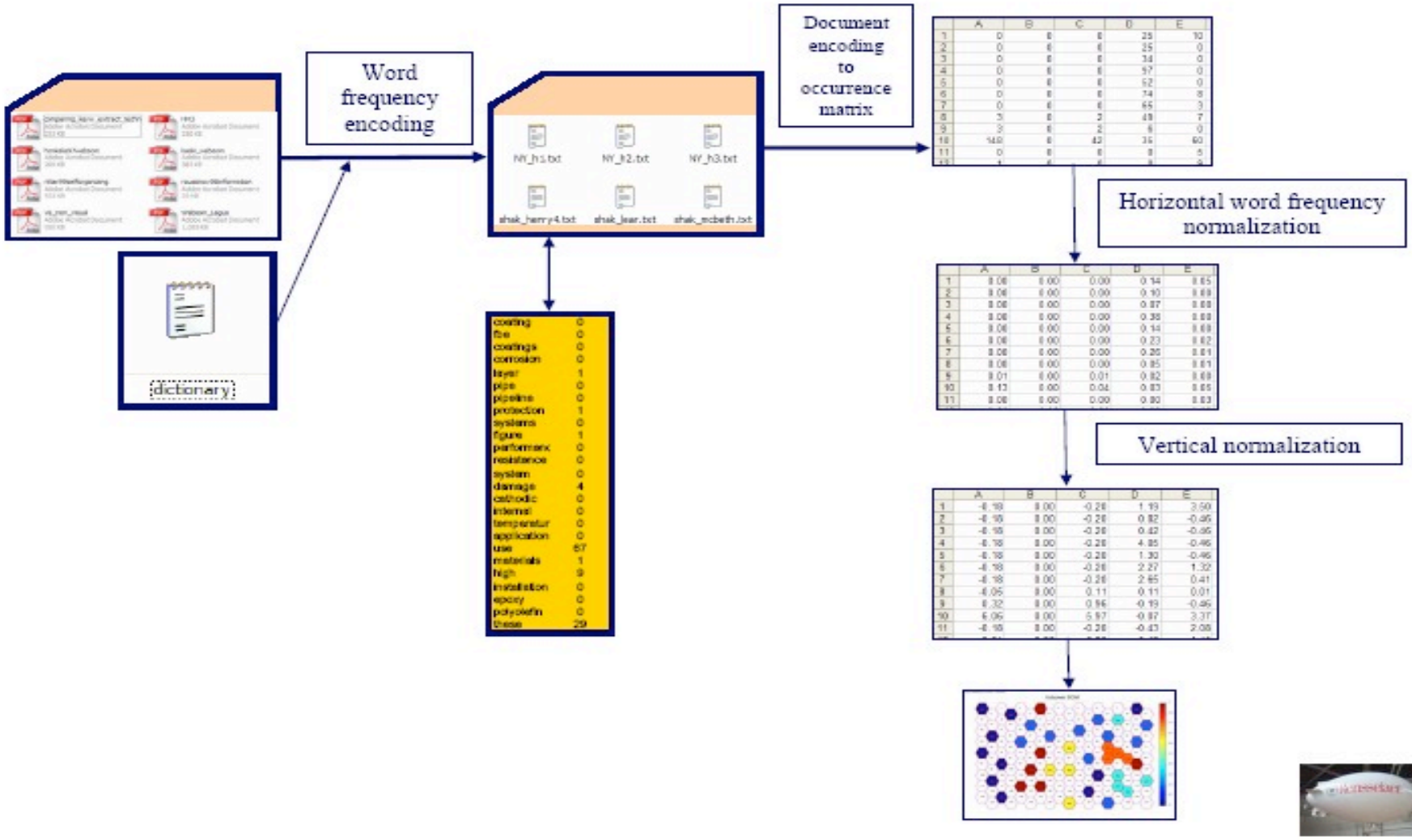


Text mining - step #1: Generating a dictionary from text file(s)



- A good dictionary is key to success in text mining
  - dictionary from single or multiple documents
  - consider number of words, different dictionaries (e.g., emotional dictionary)
  - concatenation/wordstemming (e.g., lumping book and books, walk and walking)

Text mining - step #2: Converting texts to numbers



tf-idf Encoding

- tf-idf: term frequency – inverse document frequency (use occurrence matrix)
- How important is a word to a document in a collection or corpus?
  - Importance ~ # times word appears in document
  - Offset by frequency by which word appears in the corpus of all documents
- Variations of td-idf weighting used by search engines to score relevance to query
- tf: Term frequency = number of times a term appears in a particular document

$$tf_i = \frac{n_i}{\sum_k n_k}$$

- idf: Inverse document frequency = importance of term by dividing the number of all documents by the number of documents containing the term

$$idf_i = \log \frac{|D|}{|\{d : d \ni t_i\}|}$$

$|D|$ total number of ducuments in corpus

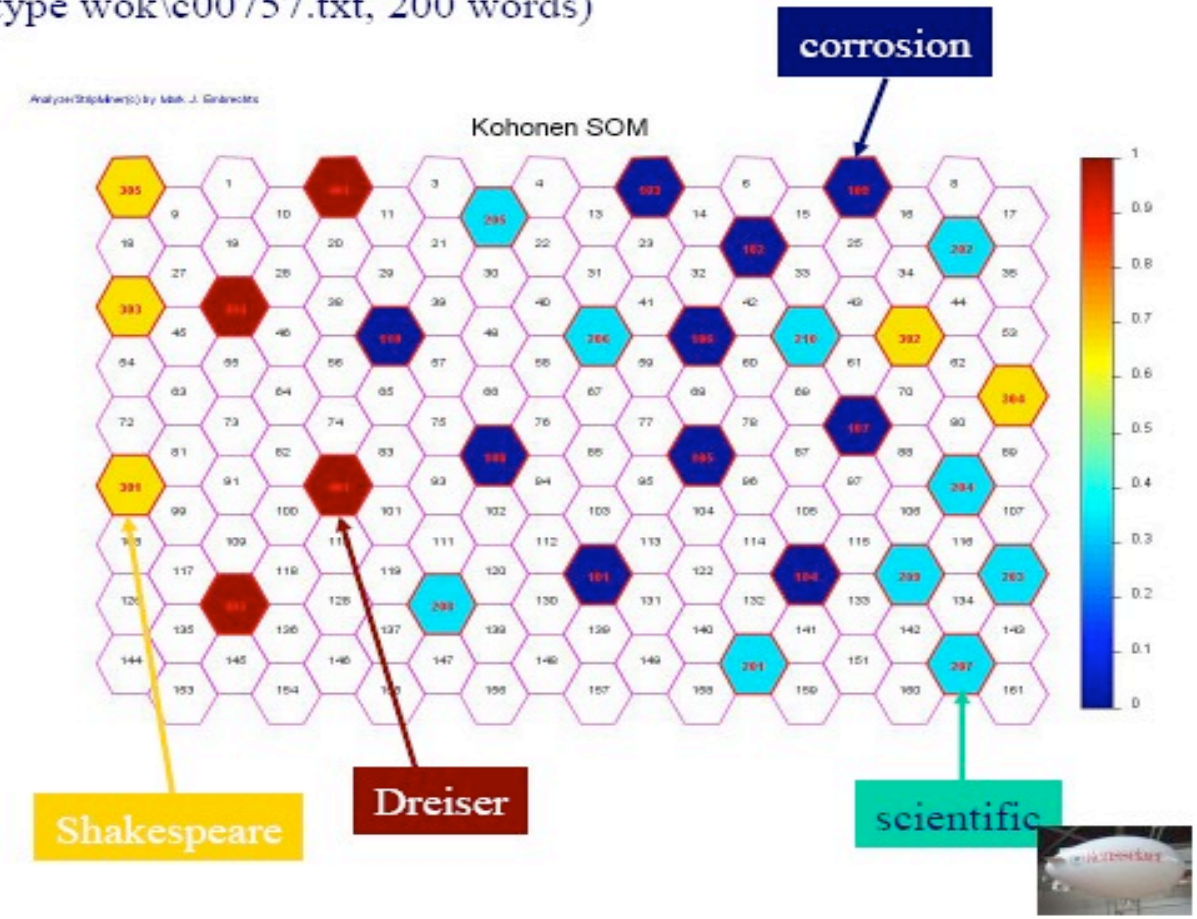
$|\{d : d \ni t_i\}|$ number of documents where the term t_i appears

$$tfidf = tf_i \cdot idf_i$$

Example: 4 text classes, 29 files

- Make first a dictionary, by going into dictionary tab
  - Make_Dictionary (e.g., type wok\c00757.txt, 200 words)
  - Activate_Dictionary
- Prepare
  - Word_Freqs
  - Make_Meta_File
  - Run_SOM

c00050.txt	1	101
c00131.txt	1	102
c00182.txt	1	103
c00192.txt	1	104
c00494.txt	1	105
c00496.txt	1	106
c00498.txt	1	107
c00627.txt	1	108
c00672.txt	1	109
c00757.txt	1	110
IEEE_Buda.txt	2	209
AE_using_wavelet_6.txt	2	210
comparing_keyw_extract_techn_1.txt	2	201
IEEE_May2000_Kewley_3.txt	2	202
JNWatson_8.txt	2	203
KB-ME-PLS_4.txt	2	204
Kewley_2002_Battle_5.txt	2	205
KKSLP04_9.txt	2	206
roussinov98information_7.txt	2	207
Websom_Lagus_10.txt	2	208
shak_hamlet.txt	3	301
shak_lear.txt	3	302
shak_mcbeth.txt	3	303
shak_caesar.txt	3	304
shak_henry4.txt	3	305
dreicer_twelve_men.txt	4	401
dreicer_titan.txt	4	402
dreicer_sister_carrie.txt	4	403
dreicer_financier.txt	4	404



KATHOLIEKE UNIVERSITEIT
LEUVEN

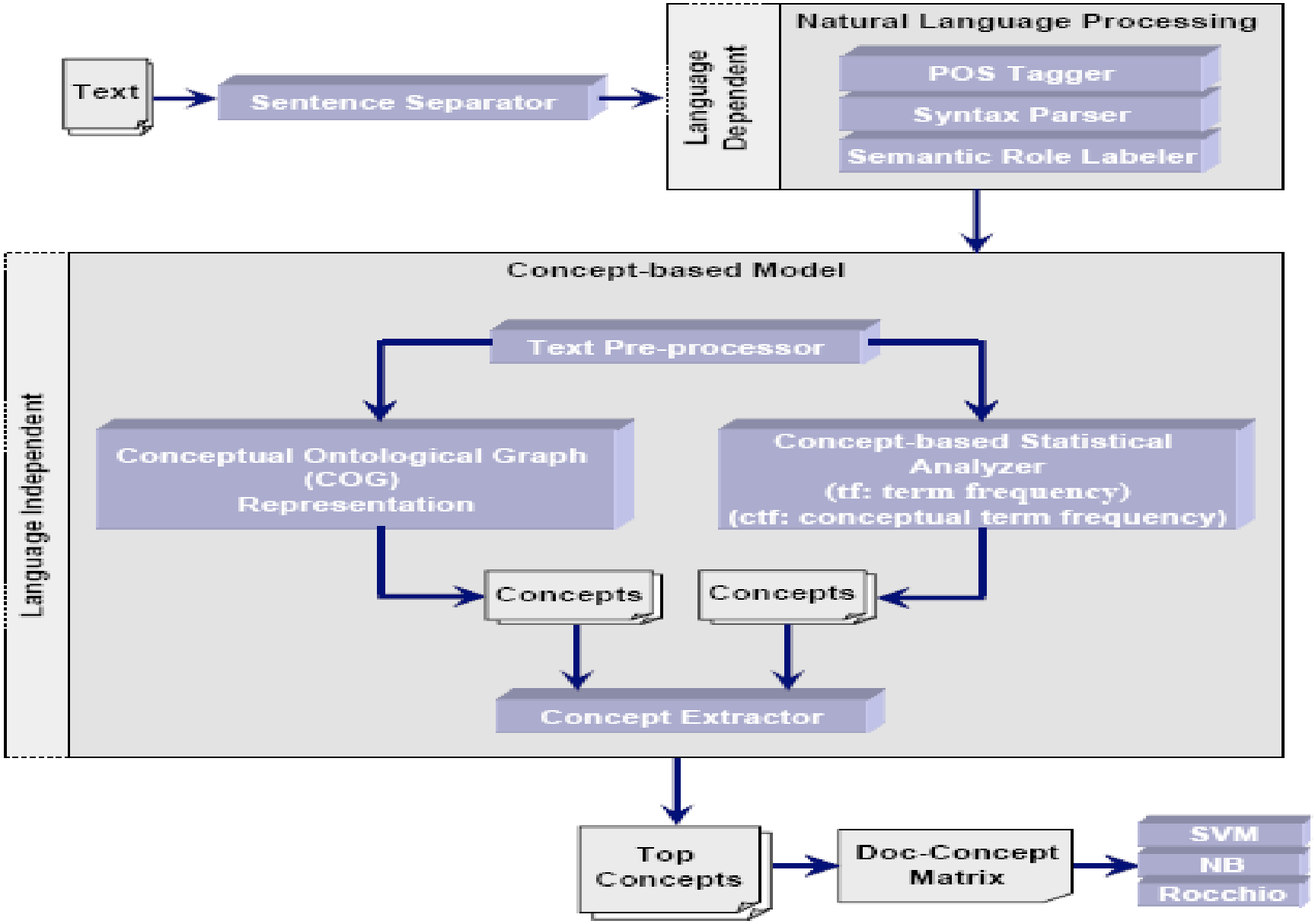


Figure 1: Concept-based Model

Concept-based Statistical Analyzer

$$weight_{stat_i} = tfweight_i + ctfweight_i \quad (1)$$

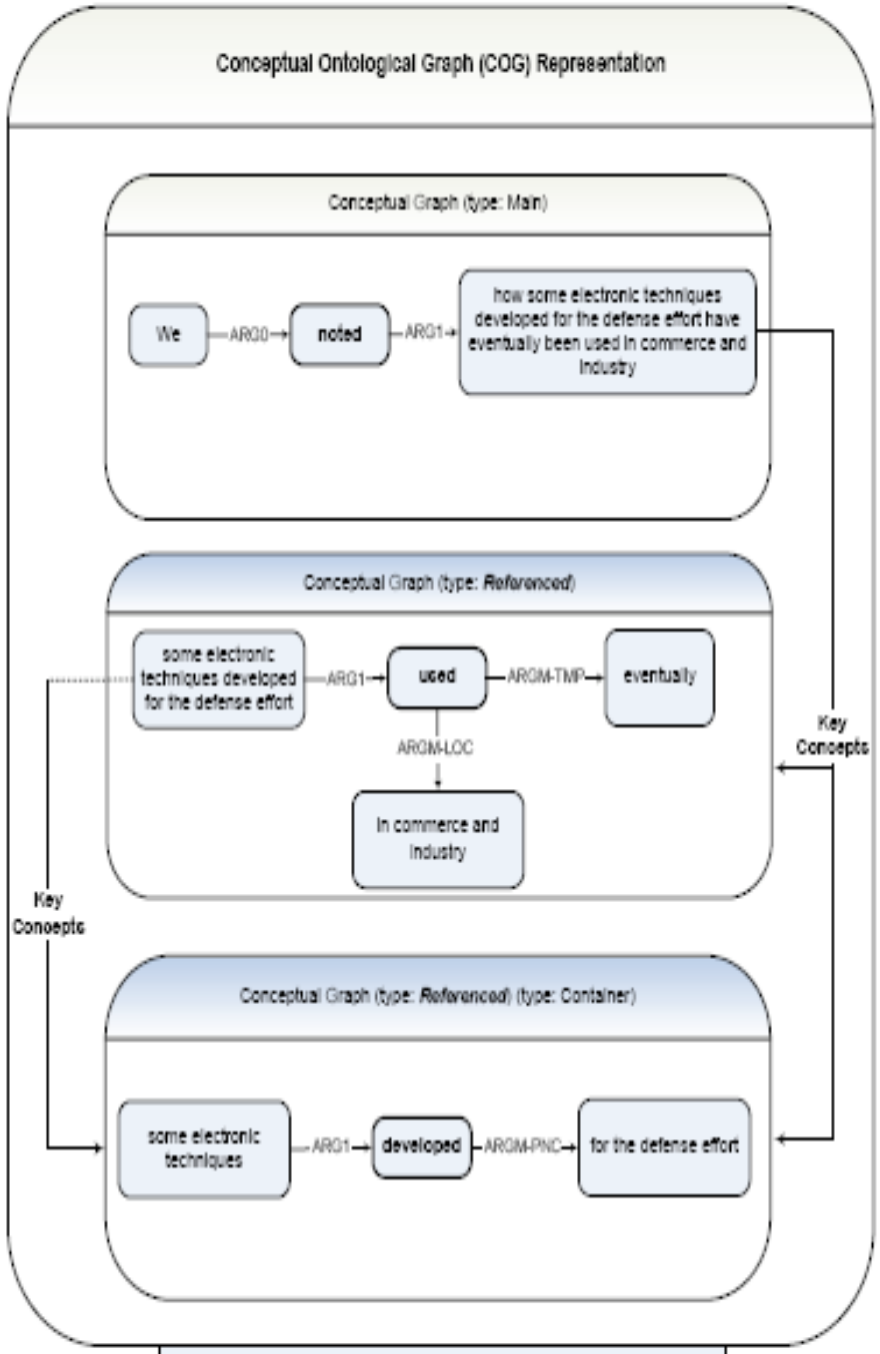
$$tfweight_i = \frac{tf_{ij}}{\sqrt{\sum_{j=1}^{cn} (tf_{ij})^2}}, \quad (2)$$

$$ctfweight_i = \frac{ctf_{ij}}{\sqrt{\sum_{j=1}^{cn} (ctf_{ij})^2}}, \quad (3)$$

Conceptual Ontological Graph (COG)

$$weightcog_i = tfweight_i * Lcog_i \qquad (4)$$

- 1st term is calculated in equation 4.
- 2nd term is discussed in a separate paper by the authors "Enhancing text retrieval performance using COG".


Conceptual Ontological Graph (COG) Representation
Conceptual Graph (type: Main)
We
ARG0
noted
ARG1
how some electronic techniques developed for the defense effort have eventually been used in commerce and industry
Conceptual Graph (type: Referenced)
some electronic techniques developed for the defense effort
ARG1
used
ARGM-TMP
eventually
ARGM-LOC
in commerce and industry
Key Concepts
Key Concepts
Conceptual Graph (type: Referenced) (type: Container)
some electronic techniques
ARG1
developed
ARGM-PNC
for the defense effort

Concept Extractor

$$weight_{comb_i} = weight_{stat_i} * weight_{COG_i} \quad (5)$$

- Combines the term calculated by the concept based statistical analyzer and the term calculated by the COG representation.
- Selects the top concepts with the maximum weight value.

Experimental Results

Table 2: Text Classification Improvement using Combined Approach ($weight_{comb}$)

DataSet		*Single-Term*		*Concept-based*		
		Macro Avg(F1)	Avg Error	Macro Avg(F1)	Avg Error	*Improvement*
Reuters	SVM	0.7421	0.0871	0.8953	0.0121	+20.64%, -86.10%
	NB	0.6127	0.2754	0.8462	0.0342	+38.11%, -87.58%
	Rocchio	0.6513	0.1632	0.8574	0.0231	+31.64%, -85.84%
ACM	SVM	0.4973	0.1782	0.8263	0.0532	+66.15%, -70.14%
	NB	0.4135	0.4215	0.7964	0.0641	+92.59%, -84.79%
	Rocchio	0.4826	0.2733	0.7935	0.0635	+64.42%, -76.76%
Brown	SVM	0.6143	0.1134	0.8753	0.0211	+42.48%, -81.39%
	NB	0.5071	0.3257	0.8372	0.0341	+65.09%, -89.53%
	Rocchio	0.5728	0.2413	0.8465	0.0243	+47.78%, -89.92%

Critique

- Novel approach to text Categorization.
- Concept Ontological Graph explained in a different research paper.
- Experimental results used “editor” documents, where terms are, in general, well structured and clean.
- Difficult to explain when “concepts” are extracted either by the concept based statistical analysis, or the COG.
- It would be interesting to reproduce the results using different techniques and datasets.

Acknowledgment

- Text Mining Slides courtesy of:
- Mark J. Embrechts (embrem@rpi.edu)
- Department of Decision Sciences & Engineering Systems
- Department of Information Technology
- Rensselaer Polytechnic Institute, Troy, NY 12180
- Presented at ICANN 2007

Thank You

?