Classification Tasks in CiteSeer

Compiled by Sujatha Das G & Cornelia Caragea

August 19, 2014

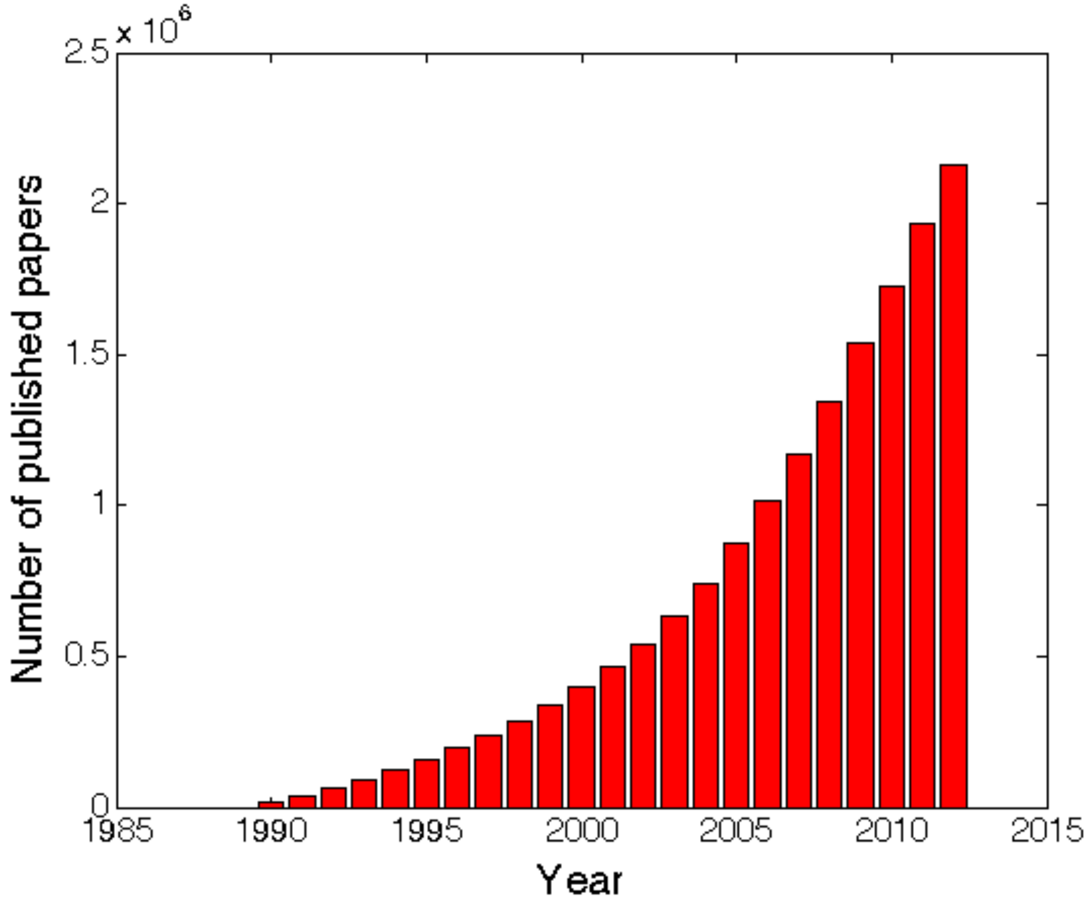
Various classification tasks in CiteSeer

- Is a crawled webpage useful to CiteSeer?
  - Researcher homepage
  - Group publication pages
  - Departmental technical reports page
- Is a crawled PDF document
  - Research document or not
- Is a research document on

  Data Mining or Computer Networks or Computer Architecture...
- Is a citation
  - Extending, refuting, crediting a given paper

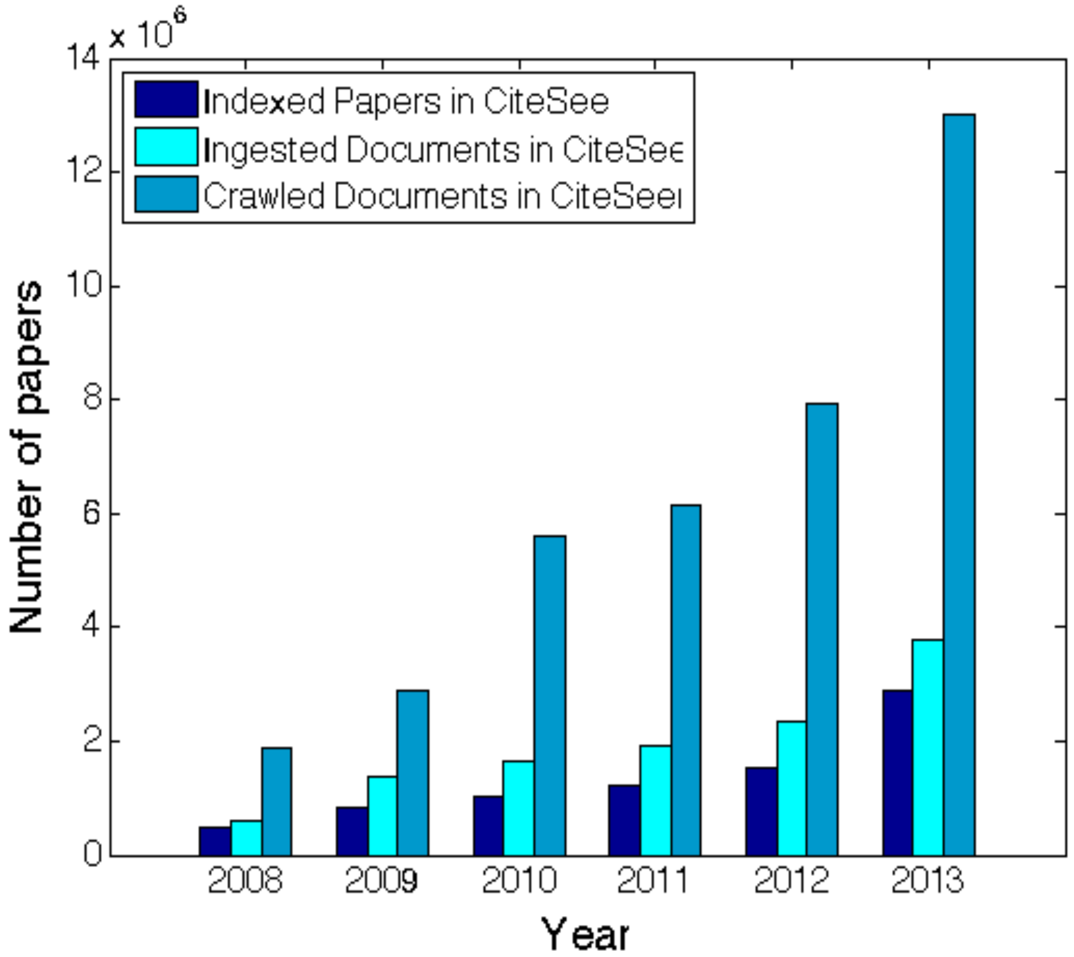
Challenges

- Large number of scholarly documents on the Web



The growth in the number of research papers published between 1990 and 2011,
extracted from DBLP.

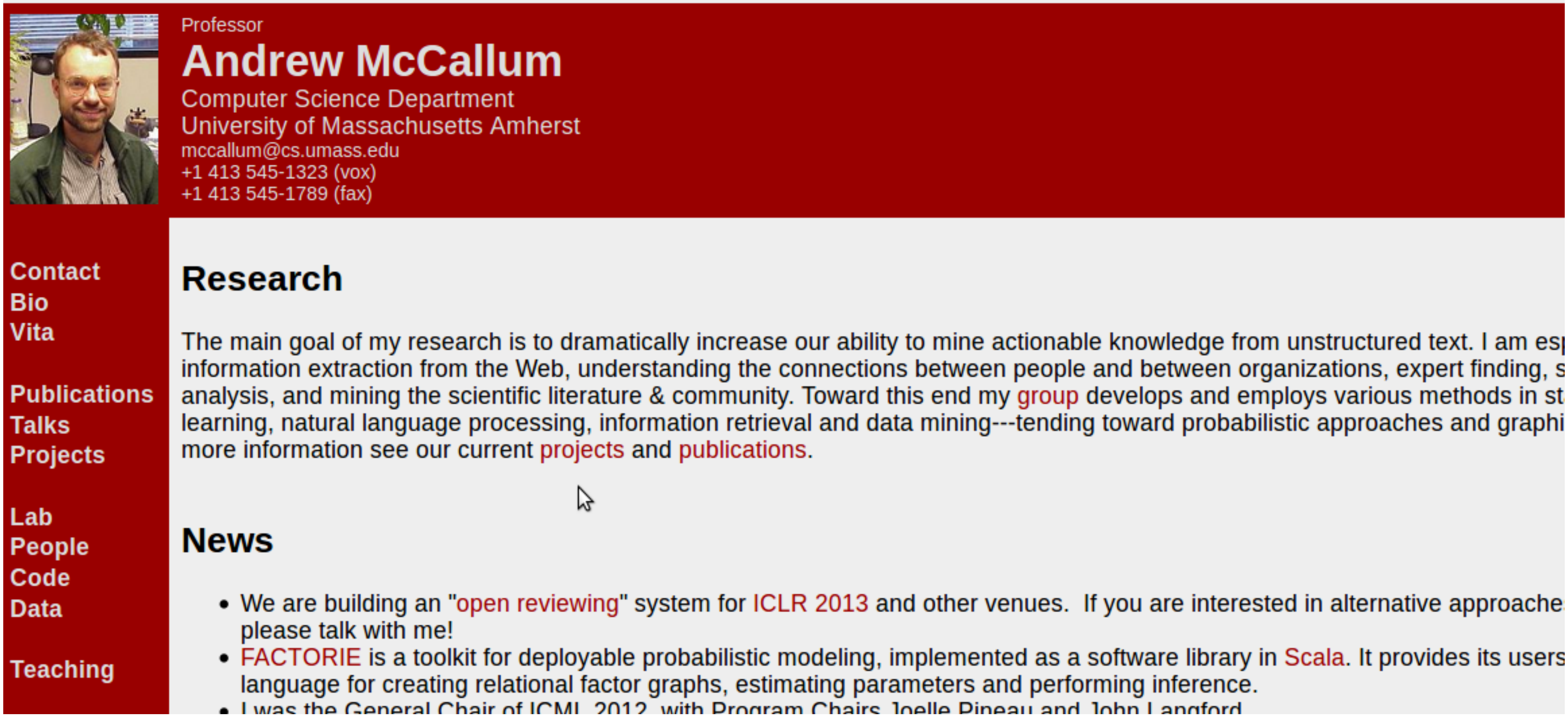
Lot of “junk” needs to be filtered



The growth in the number of crawled documents as well as in the number of research papers indexed by CiteSeerX between ‘08 and ‘13.

Researcher homepage classification

- Researcher homepages are
  - The target of “researcher name” queries on the Web.
  - An important resource for CiteSeer due to metadata and publication links.



Professor

Andrew McCallum

Computer Science Department
University of Massachusetts Amherst
mccallum@cs.umass.edu
+1 413 545-1323 (vox)
+1 413 545-1789 (fax)

Research

The main goal of my research is to dramatically increase our ability to mine actionable knowledge from unstructured text. I am es
information extraction from the Web, understanding the connections between people and between organizations, expert finding, s
analysis, and mining the scientific literature & community. Toward this end my group develops and employs various methods in st
learning, natural language processing, information retrieval and data mining---tending toward probabilistic approaches and graphi
more information see our current projects and publications.

News

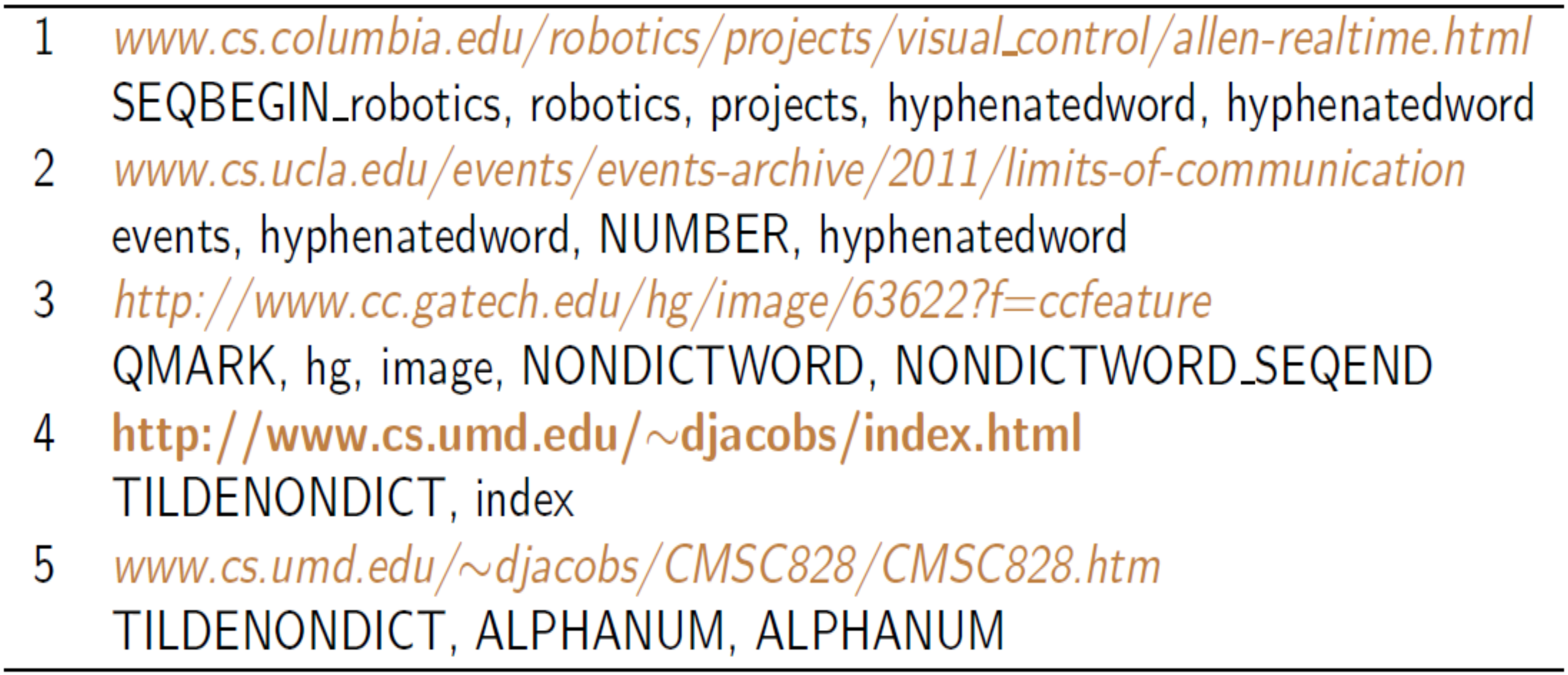
- We are building an "open reviewing" system for ICLR 2013 and other venues. If you are interested in alternative approache
please talk with me!
- FACTORIE is a toolkit for deployable probabilistic modeling, implemented as a software library in Scala. It provides its users
language for creating relational factor graphs, estimating parameters and performing inference.
- I was the General Chair of ICML 2012, with Program Chairs Joelle Pineau and John Langford

Lack of labeled negative pages

- Available labeled datasets do not cover current-day academic content encountered while crawling. Example such pages include
  - colloquia, seminars, lectures, publications, papers, talks, slides.
  - code, widgets, scripts, datasets.
  - department activities such as picnics, pages with embedded photos, and personal pages.
  - information on news, events, highlights, faq, forms.
  - alumni-related information, job and contest calls.

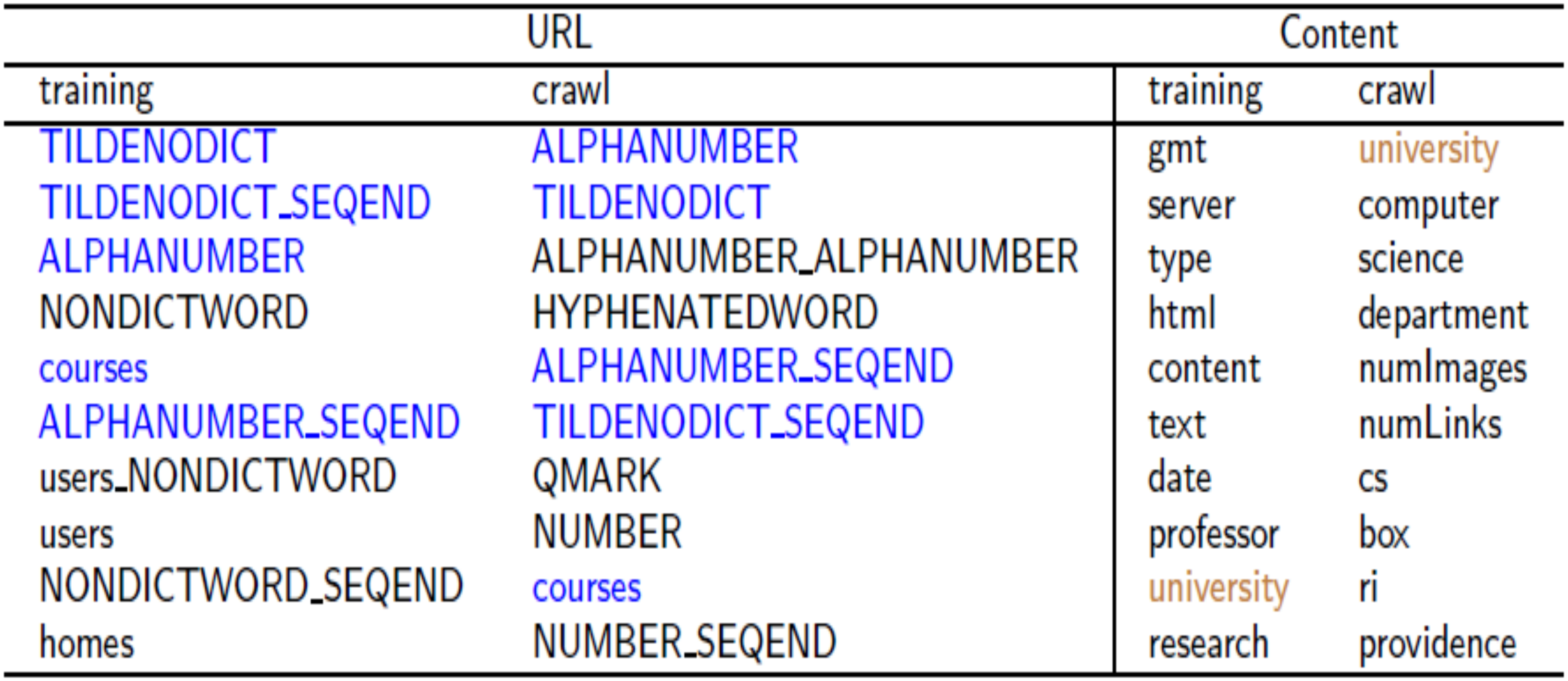
URL features

- Content (term features) not very effective due to lack of proper labeled data but URL features consistent across labeled dataset and crawled pages



1	*www.cs.columbia.edu/robotics/projects/visual_control/allen-realtime.html* SEQBEGIN_robotics, robotics, projects, hyphenatedword, hyphenatedword
2	*www.cs.ucla.edu/events/events-archive/2011/limits-of-communication* events, hyphenatedword, NUMBER, hyphenatedword
3	*http://www.cc.gatech.edu/hg/image/63622?f=ccfeature* QMARK, hg, image, NONDICTWORD, NONDICTWORD_SEQEND
4	**http://www.cs.umd.edu/∼djacobs/index.html** TILDENONDICT, index
5	*www.cs.umd.edu/∼djacobs/CMSC828/CMSC828.htm* TILDENONDICT, ALPHANUM, ALPHANUM

Note the overlap in the discriminative URL features from training and crawl datasets and hardly any overlap in the content features!



URL		Content	
training	crawl	training	crawl
TILDENODICT	ALPHANUMBER	gmt	university
TILDENODICT_SEQEND	TILDENODICT	server	computer
ALPHANUMBER	ALPHANUMBER_ALPHANUMBER	type	science
NONDICTWORD	HYPHENATEDWORD	html	department
courses	ALPHANUMBER_SEQEND	content	numImages
ALPHANUMBER_SEQEND	TILDENODICT_SEQEND	text	numLinks
users_NONDICTWORD	QMARK	date	cs
users	NUMBER	professor	box
NONDICTWORD_SEQEND	courses	university	ri
homes	NUMBER_SEQEND	research	providence

We still need a good text-based classifier

http://john.blitzer.com/

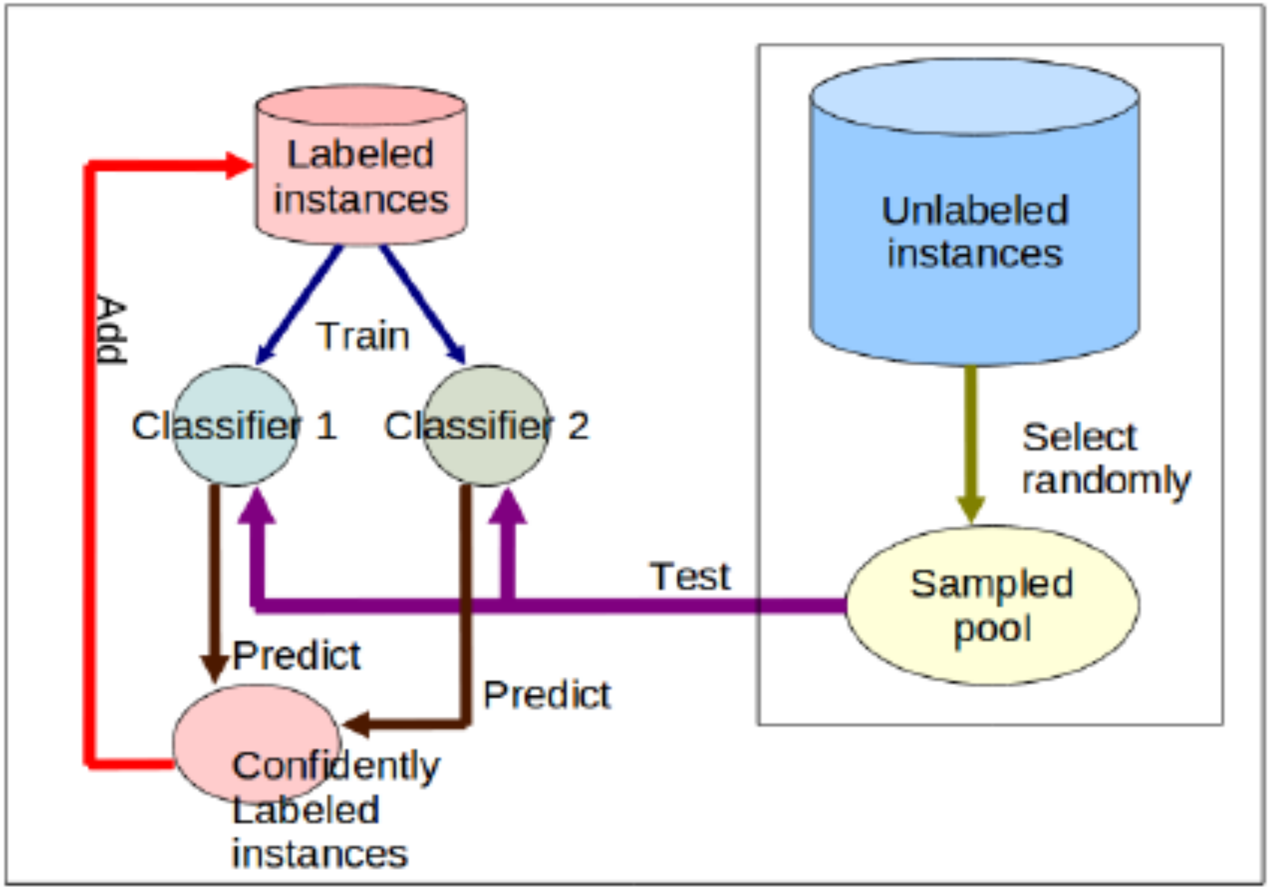
http://clgiles.ist.psu.edu/

http://ben.adida.net/

- URL features cannot be extracted always
- In our experiments, we could not extract URL features for about 27% of the training instances

Can we combine the evidence from the two sources (URL and content) to learn a better classifier?

Use Co-training!

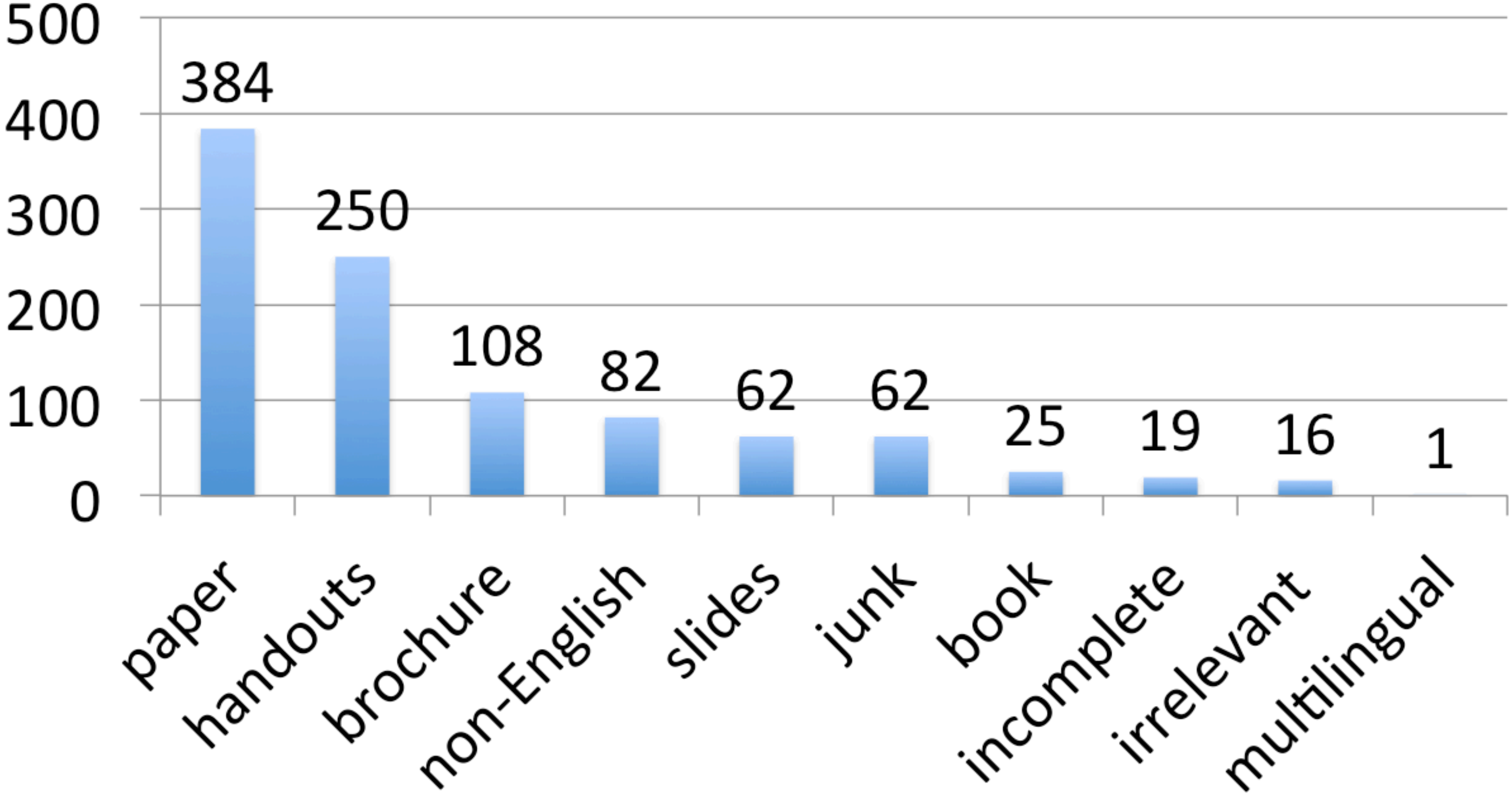


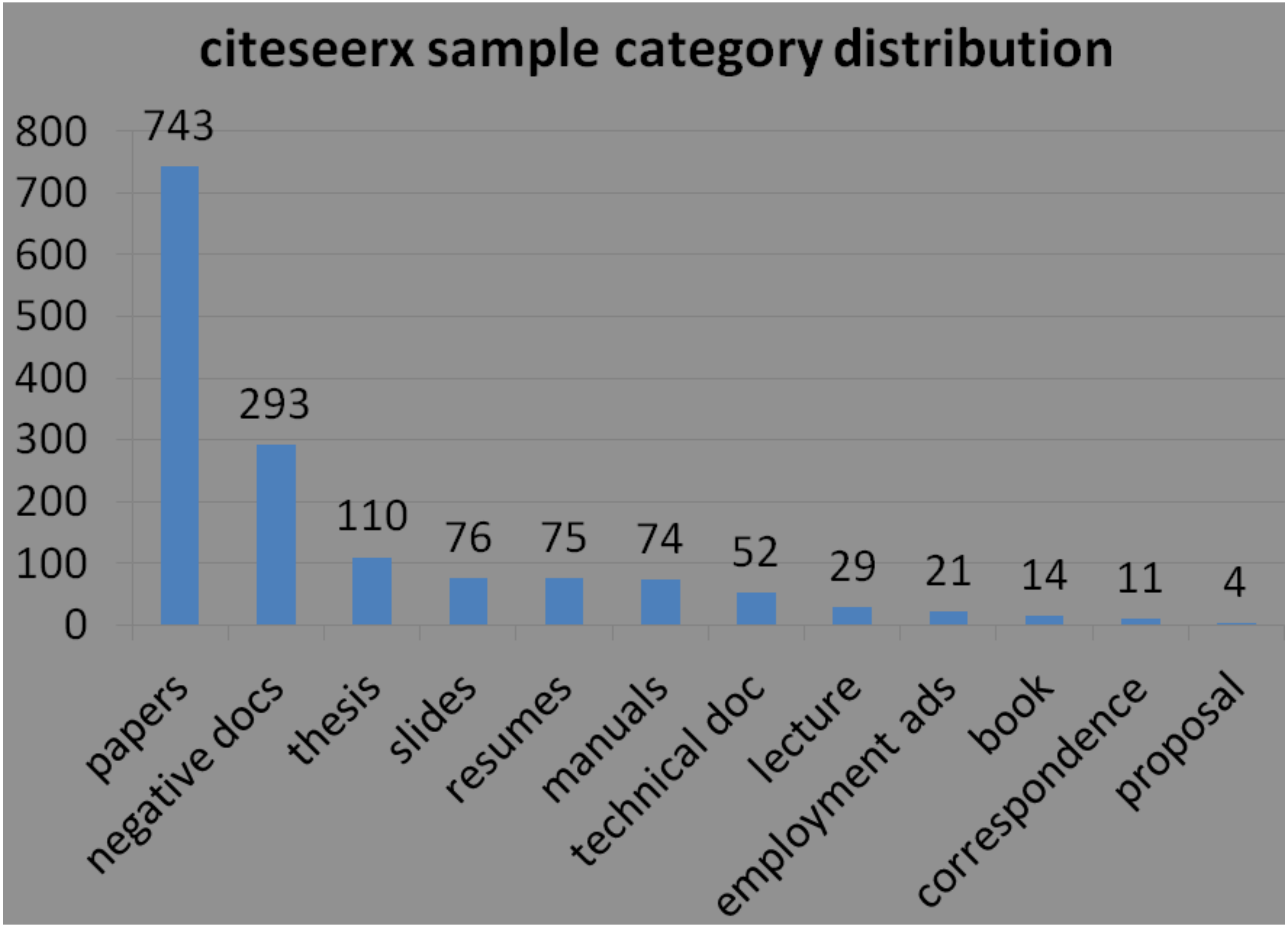
Modified from *http://web.cs.gc.cuny.edu/~zhengchen/papers/naacl09-bootstrap-slides.ppt*

Automatic Research Article Classification Methodology

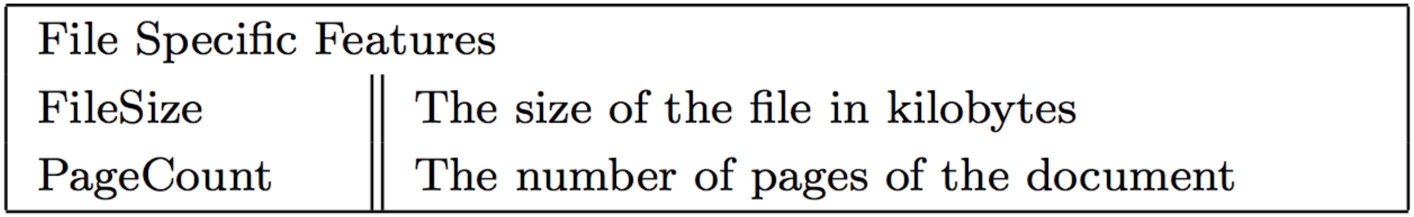
- Classify documents as *research* if they contain any of the words *references* or *bibliography* in text
  - Current method in CiteSeer
    - Will mistakenly classify documents such as CV or slides as research articles if they contain *references* in them
    - Will miss to identify research articles that do not contain any of the two words
- Classify documents using a "bag of words" approach
    - May not capture the specifics of research articles, e.g., due to the diversity of the topics covered in CiteSeerX.
    - For example, an article in HCI may have a different vocabulary space compared to a paper in IR, but some essential terms may persist across papers.
- Better methods?

crawl sample category distribution

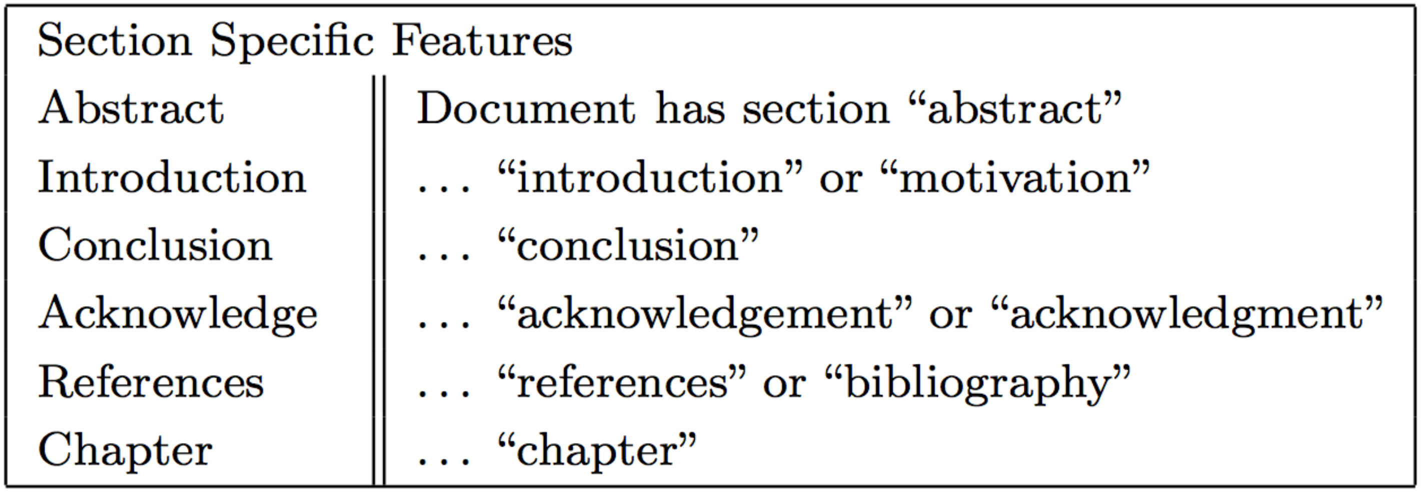



citeseerx sample category distribution
800
700
600
500
400
300
200
100
0
743
293
110
76
75
74
52
29
21
14
11
4
papers
negative docs
thesis
slides
resumes
manuals
technical doc
lecture
employment ads
book
correspondence
proposal

Possible Features for Research Article Identification



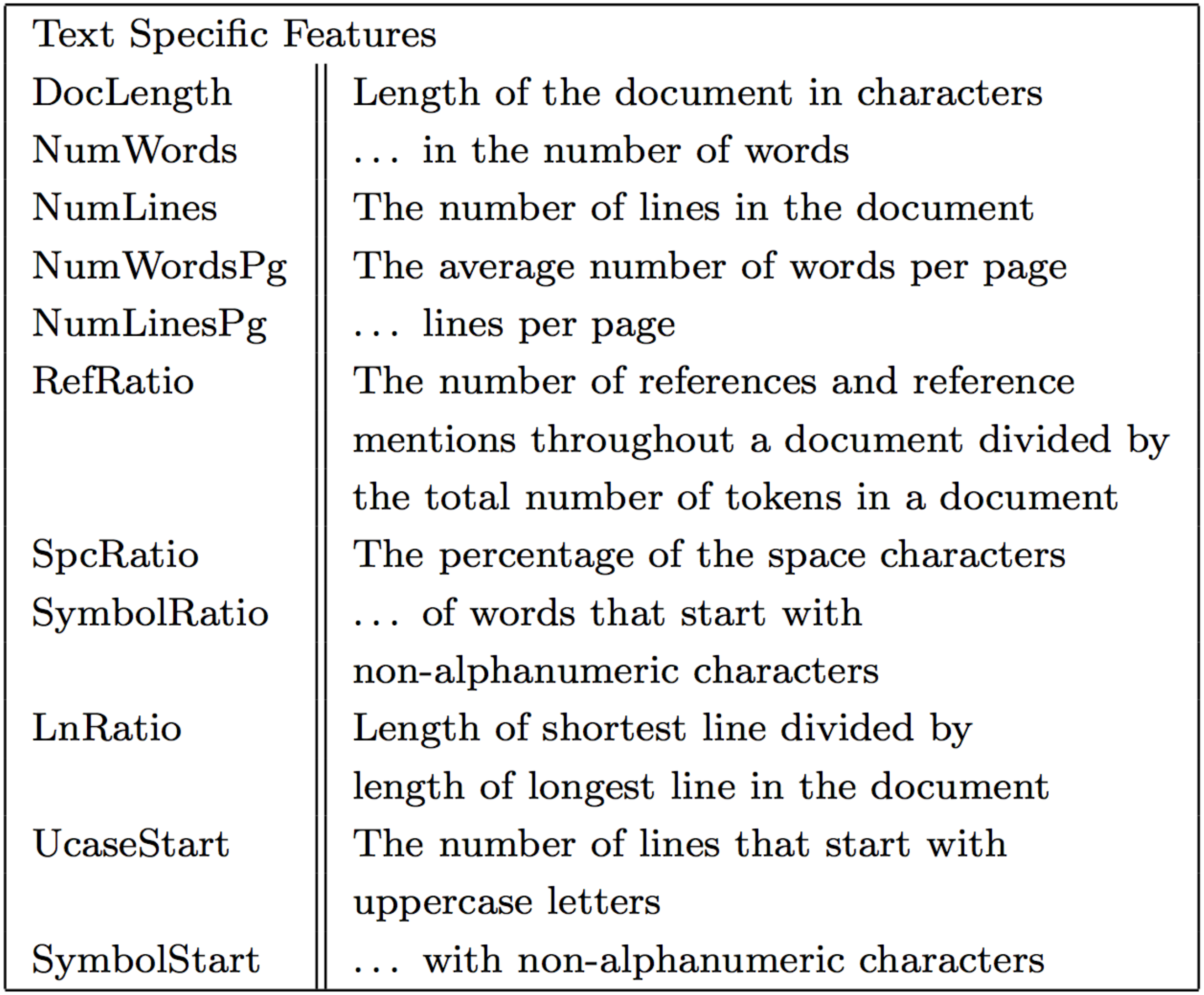
File Specific Features	
FileSize	The size of the file in kilobytes
PageCount	The number of pages of the document



Section Specific Features	
Abstract	Document has section “abstract”
Introduction	. . . “introduction” or “motivation”
Conclusion	. . . “conclusion”
Acknowledge	. . . “acknowledgement” or “acknowledgment”
References	. . . “references” or “bibliography”
Chapter	. . . “chapter”

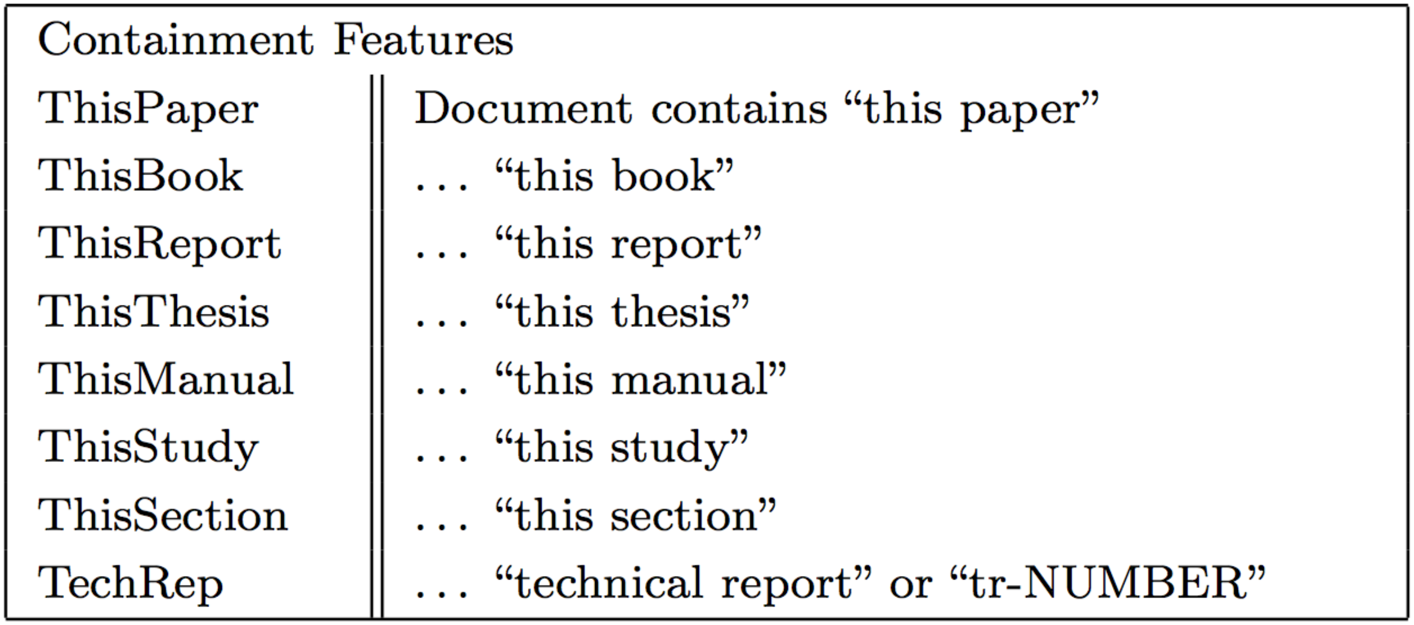
Data derived from PDFBox text

Structural Features



Text Specific Features	
DocLength	Length of the document in characters
NumWords	. . . in the number of words
NumLines	The number of lines in the document
NumWordsPg	The average number of words per page
NumLinesPg	. . . lines per page
RefRatio	The number of references and reference mentions throughout a document divided by the total number of tokens in a document
SpcRatio	The percentage of the space characters
SymbolRatio	. . . of words that start with non-alphanumeric characters
LnRatio	Length of shortest line divided by length of longest line in the document
UcaseStart	The number of lines that start with uppercase letters
SymbolStart	. . . with non-alphanumeric characters

Textual Features



Containment Features	
ThisPaper	Document contains “this paper”
ThisBook	. . . “this book”
ThisReport	. . . “this report”
ThisThesis	. . . “this thesis”
ThisManual	. . . “this manual”
ThisStudy	. . . “this study”
ThisSection	. . . “this section”
TechRep	. . . “technical report” or “tr-NUMBER”

Conclusions

- The classification tasks in CiteSeer are challenging
  - Although we deal with textual content, text classification algorithms/features don't work directly
  - Obtaining labeled data is difficult due to changing types and manual effort, so semi-supervised and unsupervised methods are desirable
  - Harvesting “domain-specific” knowledge in designing features is a must for accurate models
  - Need fast and adaptive models that can incorporated during crawls!

References

1. Sujatha Das Gollapalli, Cornelia Caragea, Prasenjit Mitra, C. Lee Giles: Researcher homepage classification using unlabeled data. WWW 2013
2. Cornelia Caragea, Jian Wu, Alina Maria Ciobanu, Kyle Williams, Juan Pablo Fernández Ramírez, Hung-Hsuan Chen, Zhaohui Wu, C. Lee Giles: CiteSeer x : A Scholarly Big Dataset. ECIR 2014
3. Cornelia Caragea, Jian Wu, Kyle Williams, Sujatha Das Gollapalli, Madian Khabsa, Pradeep Teregowda, and C. Lee Giles. "Automatic Identification of Research Articles from Crawled Documents." WSC workshop at WSDM 2014