Real Time Web Text Classification and Analysis of Reading Difficulty

Eleni Miltsakaki & Audrey Troutt

University of Pennsylvania

Summary

- Two system components to support struggling readers
  - Read-X: searches for text on the web, classifies it thematically and analyzes its reading difficulty
  - TOREADOR: takes text as input, highlights words predicted to be difficult given the reader's prior familiarity with thematic area

Outline

- Motivation and design principles
- Description of Read-X
  - Web search
  - Text extraction
  - Text classification
  - Readability formulas
- Modeling the reader
  - Prior thematic familiarity
  - Description of TOREADOR
- Related work
- Future work

Outline

- Motivation and design principles
- Description of Read-X
  - Web search
  - Text extraction
  - Text classification
  - Readability formulas
- Modeling the reader
  - Description of TOREADOR
- Related work
- Future work

Struggling readers

- 29% of high school seniors below basic achievement in reading in 2005

  (U.S. Department of Education 2005)

- Lack of age and interest appropriate reading material for adolescent and low level readers.

- Look at the web!

Design principles

- Easy access and use
  - Run from the web, no installation, no manual, no fees
- Text retrieval and analysis in real time
  - Accessing the web directly
  - No need for updating pre-processed database
- Empowering the educator
  - by providing a tool that s/he can use to build his/her own curriculum

Outline

- Motivation
- Applications
- Read-X
  - Web search
  - Text extraction
  - Readability analysis
  - Text classification
- Modeling the reader
  - Description of TOREADOR
- Related work
- Future work

Read-X

- Searches the web on keyword prompt
- Extracts the human-readable text from the html, xml, doc or PDF document stored at each URL.
- Returns text classified:
  - Thematically
  - Reading level

Text classification

- Corpus
  - 3.4 m words, mostly from NetTrekker, manually tagged
- Experimented with three classifiers
  - Naïve Bayes
  - MaxEnt
  - MIRA *(Crammer et al 2008)*
- Three levels of granularity
  - 3-way (3 themes)
  - 8-way (8 themes)
  - 41-way (41 themes)

NB and MaxEnt using MALLET's tools (http://mallet.cs.umass.edu)

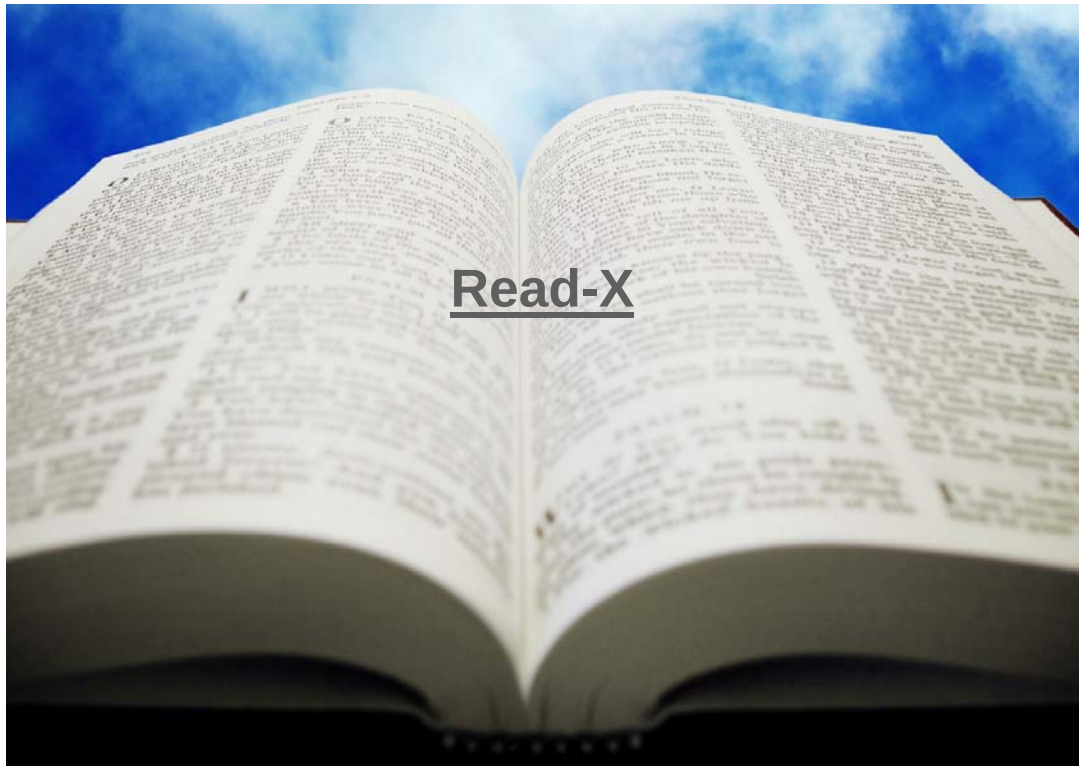
Results

Classifier	3-way	8-way	41-way
Naïve Bayes	88.23%	66%	30%
MaxEnt	92.86%	78%	66%
MIRA	N/A	76%	58%

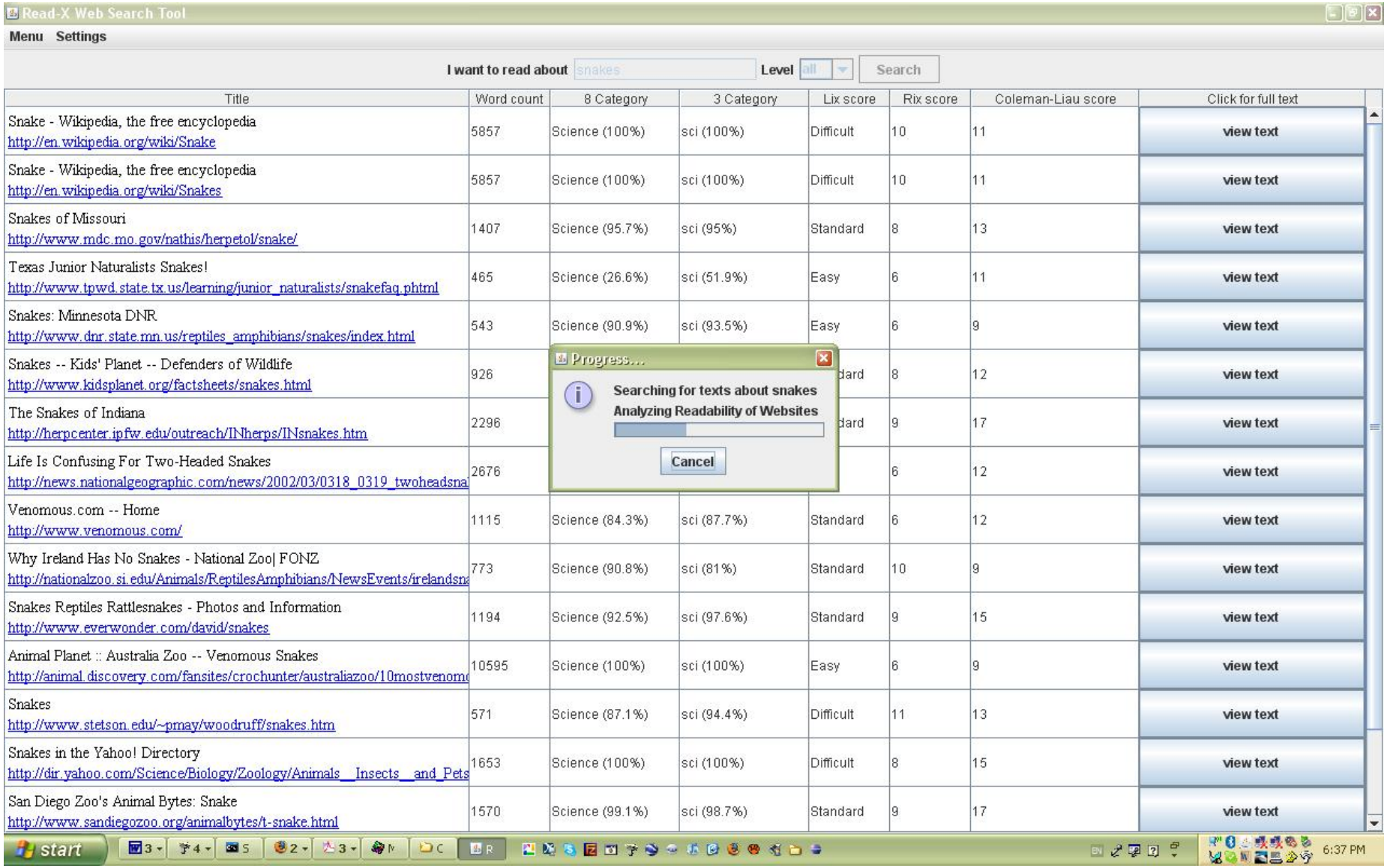
Readability analysis

- This version of Read-X computes traditional readability formulas
  - Lix, Rix, Coleman-Liau
    (Anderson 1983, Coleman & Liau 1975)

Demo



http://net-read.blogspot.com/


Read-X Web Search Tool
Menu Settings
I want to read about snakes Level all Search
Title
Word count
8 Category
3 Category
Lix score
Rix score
Coleman-Liau score
Click for full text
Snake - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Snake
5857
Science (100%)
sci (100%)
Difficult
10
11
view text
Snake - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Snakes
5857
Science (100%)
sci (100%)
Difficult
10
11
view text
Snakes of Missouri
http://www.mdc.mo.gov/nathis/herpetol/snake/
1407
Science (95.7%)
sci (95%)
Standard
8
13
view text
Texas Junior Naturalists Snakes!
http://www.tpwd.state.tx.us/learning/junior_naturalists/snakefaq.phtml
465
Science (26.6%)
sci (51.9%)
Easy
6
11
view text
Snakes: Minnesota DNR
http://www.dnr.state.mn.us/reptiles_amphibians/snakes/index.html
543
Science (90.9%)
sci (93.5%)
Easy
6
9
view text
Snakes -- Kids' Planet -- Defenders of Wildlife
http://www.kidsplanet.org/factsheets/snakes.html
926
8
12
view text
The Snakes of Indiana
http://herpcenter.ipfw.edu/outreach/INherps/INsnakes.htm
2296
9
17
view text
Life Is Confusing For Two-Headed Snakes
2676
6
12
view text
Venomous.com -- Home
http://www.venomous.com/
1115
Science (84.3%)
sci (87.7%)
Standard
6
12
view text
Why Ireland Has No Snakes - National Zoo| FONZ
773
Science (90.8%)
sci (81%)
Standard
10
9
view text
Snakes Reptiles Rattlesnakes - Photos and Information
http://www.everwonder.com/david/snakes
1194
Science (92.5%)
sci (97.6%)
Standard
9
15
view text
Animal Planet :: Australia Zoo -- Venomous Snakes
http://animal.discovery.com/fansites/crochunter/australiazoo/10mostvenom
10595
Science (100%)
sci (100%)
Easy
6
9
view text
Snakes
http://www.stetson.edu/~pmay/woodruff/snakes.htm
571
Science (87.1%)
sci (94.4%)
Difficult
11
13
view text
Snakes in the Yahoo! Directory
http://dir.yahoo.com/Science/Biology/Zoology/Animals__Insects__and_Pets
1653
Science (100%)
sci (100%)
Difficult
8
15
view text
San Diego Zoo's Animal Bytes: Snake
http://www.sandiegozoo.org/animalbytes/t-snake.html
1570
Science (99.1%)
sci (98.7%)
Standard
9
17
view text
Progress...
Searching for texts about snakes
Analyzing Readability of Websites
Cancel
start
6:37 PM

Outline

- Motivation
- Applications
- Read-X
  - Web search
  - Text extraction
  - Readability analysis
  - Text classification
- Modeling the reader
  - Description of TOREADOR
- Related work
- Future work

Modeling the reader

- Some reader variables
  - Familiarity with thematic area
  - Reading ability
  - Interest/Motivation

Modeling familiarity

- *Word frequencies per thematic area*

- Typical word frequency indices used by test developers and educators are computed from corpora with mixed text.

Word frequencies

Arts		Career and Business		Literature		Philosophy		Science		SocialStudies		SportHealth		Technology	
Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq
musical	166	product	257	seemed	1398	argument	174	trees	831	behavior	258	players	508	software	584
leonardo	166	income	205	myself	1257	knowledge	158	bacteria	641	states	247	league	443	computer	432
instrument	155	market	194	friend	1255	augustine	148	used	560	psychoanalytic	222	player	435	site	333
horne	149	price	182	looked	1231	belief	141	growth	486	social	198	soccer	396	video	308
banjo	128	cash	178	things	1153	memory	130	acid	476	clemency	167	football	359	games	303
american	122	analysis	171	caesar	1059	truth	130	years	472	psychology	157	games	320	used	220
used	119	resources	165	going	1051	logic	129	alfalfa	386	psychotherapy	147	teams	292	systems	200
nature	111	positioning	164	having	1050	things	125	crop	368	united	132	national	273	programming	174
artist	104	used	153	asked	1023	existence	115	species	341	society	131	years	263	using	172
wright	98	sales	151	indeed	995	informal	113	acre	332	court	113	season	224	engineering	170

TOREADOR

- TOREADOR highlights words expected to be difficult for the reader.

- Predicts difficult words per grade level and thematic familiarity

TOREADOR

To read or not to read---that is the question!

Enter Text Here | Read Text Here

With Mentors at Their Sides, Girls in Need Write Their Stories and Find New Lives
By J. COURTNEY SULLIVAN

On Saturdays during the school year and all week in the summer, PinChang Huang, 16, leaves her home in Queens just after dawn and boards a crowded van bound for a nail salon on Long Island.

Through a long workday, she gives manicures, pedicures and massages, and observes her clients at the Aroma Spa with a careful eye. Older women are most prone to yell if you make a mistake. Customers who read books tend to tip the best.

PinChang has not seen her mother since she came to New York with her father and brother four years ago from a small village in China. She spoke no English and had no friends, and all the buildings looked the same to her, so she often walked into the wrong apartment complex on her way home from school. To ward off frustration and loneliness, she started keeping a journal.

"I wrote down everything I saw, everything that made me happy or upset," she said. "I wrote the things I wished I could say out loud." In front of a packed

Reader's grade level: 2 3 4 5 6 7 8 9 10 11 12 13+
Familiar subjects to reader: Literature Sports Science

WordNet Search - 3.0 - WordNet home page - Glossary - Help

Word to search for: manicure
Search WordNet
Display Options: (Select option to change)
Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Noun

- S: (n) **manicure** (professional care for the hands and fingernails)

Verb

- S: (v) **manicure** (trim carefully and neatly) *"manicure fingernails"*
- S: (v) **manicure** (care for (one's hand) by cutting and shaping the nails, etc.)

WordNet home page

Done

TOREADOR

To read or not to read---that is the question!

Enter Text Here | **Read Text Here**

With Mentors at Their Sides, Girls in Need Write Their Stories and Find New Lives By J. COURTNEY SULLIVAN On Saturdays during the school year and all week in the summer, PinChang Huang, 16, leaves her home in Queens just after dawn and boards a crowded van bound for a nail salon on Long Island. Through a long workday, she gives manicures, pedicures and massages, and observes her clients at the Aroma Spa with a careful eye. Older women are most prone to yell if you make a mistake. Customers who read books tend to tip the best. PinChang has not seen her mother since she came to New York with her father and brother four years ago from a small village in China. She spoke no English and had no friends, and all the buildings looked the same to her, so she often walked into the wrong apartment complex on her way home from school. To ward off frustration and loneliness, she started keeping a journal. "I wrote down everything I saw, everything that made me happy or upset," she said. "I wrote the things I wished I could say out loud." In front of a packed auditorium at the New School in Manhattan one night recently, she got her chance. PinChang and her mentor, Deborah Kolben, a former managing editor at The Village Voice, read an essay they wrote together about PinChang's getting her first manicure. PinChang spoke about the peculiar sensation of being on the receiving end of an exchange that often makes her feel "like a slave." The reading was hosted by a nonprofit group that pairs high school girls from disadvantaged backgrounds who want to be writers with women who are authors, journalists, playwrights, poets and editors. The group produces an anthology of student writing each spring, and puts on several public readings. The readings are often the first chance girls get to read their own words in front of an audience. "I was so nervous when I stepped onstage," recalled PinChang, a junior at Flushing International High School. "I was shaking. But now I feel like I can say or do anything." Maya Nussbaum, 31, helped found the group, Girls Write Now, 10 years ago when she was a senior majoring in creative writing at

WordNet Search - 3.0 - WordNet home page - Glossary - Help

Word to search for: manicure

Search WordNet

Display Options: (Select option to change)

Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Noun

- S: (n) **manicure** (professional care for the hands and fingernails)

Verb

- S: (v) **manicure** (trim carefully and neatly) *"manicure fingernails"*
- S: (v) **manicure** (care for (one's hand) by cutting and shaping the nails, etc.)

WordNet home page

Done

Related systems

- NetTrekker *(commercial product)*
  - manually built database
  - organized by readability, grade level and subject area
- REAP Tutor *(Heilman et al 2006/8)*
  - Intelligent tutoring system
  - Off-line pre-processed database
  - Includes exercises to support ESL vocabulary learning

Current and future work

- Evaluation studies
  - Reading eye-tracking to evaluate predictions of difficult vocabulary
  - Self-paced reading + comprehension question for overall reading difficulty

Current and future work

- Deeper understanding of reading difficulty
  - Syntactic/semantic complexity
    - Syntactic ‘signature’ of different levels of complexity
    - Syntactic proxies for propositional density
  - Perceived coherence
    - Number of introduced topics and length of elaboration
    - Rhetorical structure and intensity of inferencing

Thank you!