



Chris A. Mattmann, Ph.D.

Jet Propulsion Laboratory, California Institute of Technology /  
University of Southern California / Apache Software Foundation

# Content Detection and Analysis for Big Data

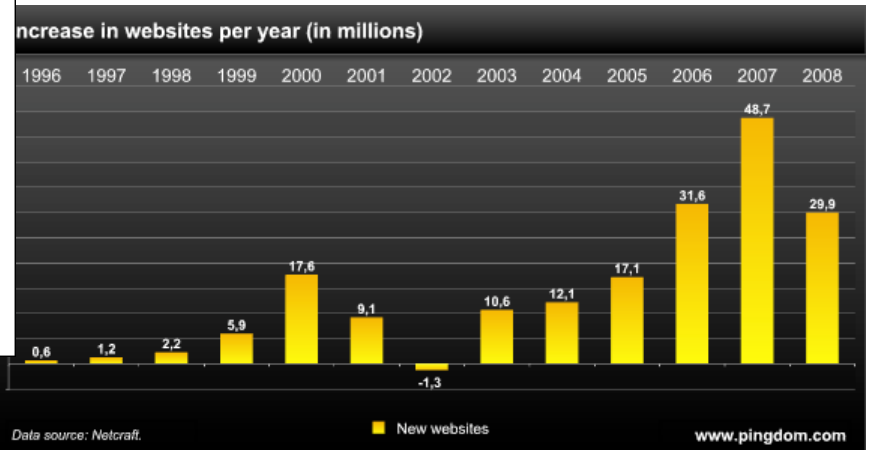
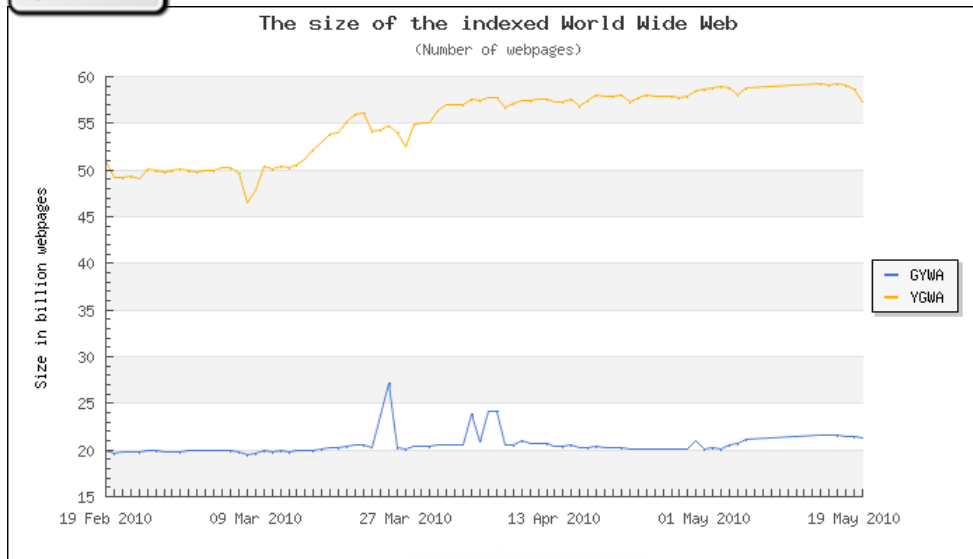
# Outline

- The Information Landscape
- Importance of Content Detection
- Challenges
- Approaches
- Introduction to Apache Tika
- Using Apache Tika
- Wrap-up

# Some notes

- Talk optimized for breadth, not depth
  - Depth can be found in my CSCI 572 course at USC on Search Engines and Information Retrieval
    - <http://www-scf.usc.edu/~csci572/>
- Encourage you to check the references at the end and feel free to ask questions
  - <http://twitter.com/chrismattmann>  
@chrismattmann
  - [chris.a.mattmann@nasa.gov](mailto:chris.a.mattmann@nasa.gov)

# The Information Landscape



# Proliferation of content types available

- By some accounts, 16K to 51K content types\*
- What to do with content types?
  - Parse them
    - How?
    - Extract their text and structure
  - Index their metadata
    - In an indexing technology like Lucene, Solr, or Compass, or in Google Appliance
  - Identify what language they belong to
    - Ngrams

\*<http://filext.com/>

# Importance of content types

Web Images Videos Maps News Shopping Gmail more ▼

Google

language identification

About 6,620,000 results (0.25 seconds)

Everything  
More

All results  
Related searches  
More search tools

Something different  
word spotting  
document classification  
speaker identification  
speaker verification  
pos tagging

**Language Identification**  
[www.basistech.com](http://www.basistech.com) Detects the language and encoding of over 55 languages

**Language Identification** ☆  
How to find out what language something is written in.  
[www.translation-guide.com/language\\_identification.htm](http://www.translation-guide.com/language_identification.htm) - Cached - Similar

**Translation Wizard > Language Identification (beta) || Fagan Finder** ☆  
Aug 25, 2003 ... Identify the language of text or a web page. If you do not know that language of something, this page will identify it for you.  
[www.faganfinder.com](http://www.faganfinder.com) > Translation Wizard - Cached - Similar

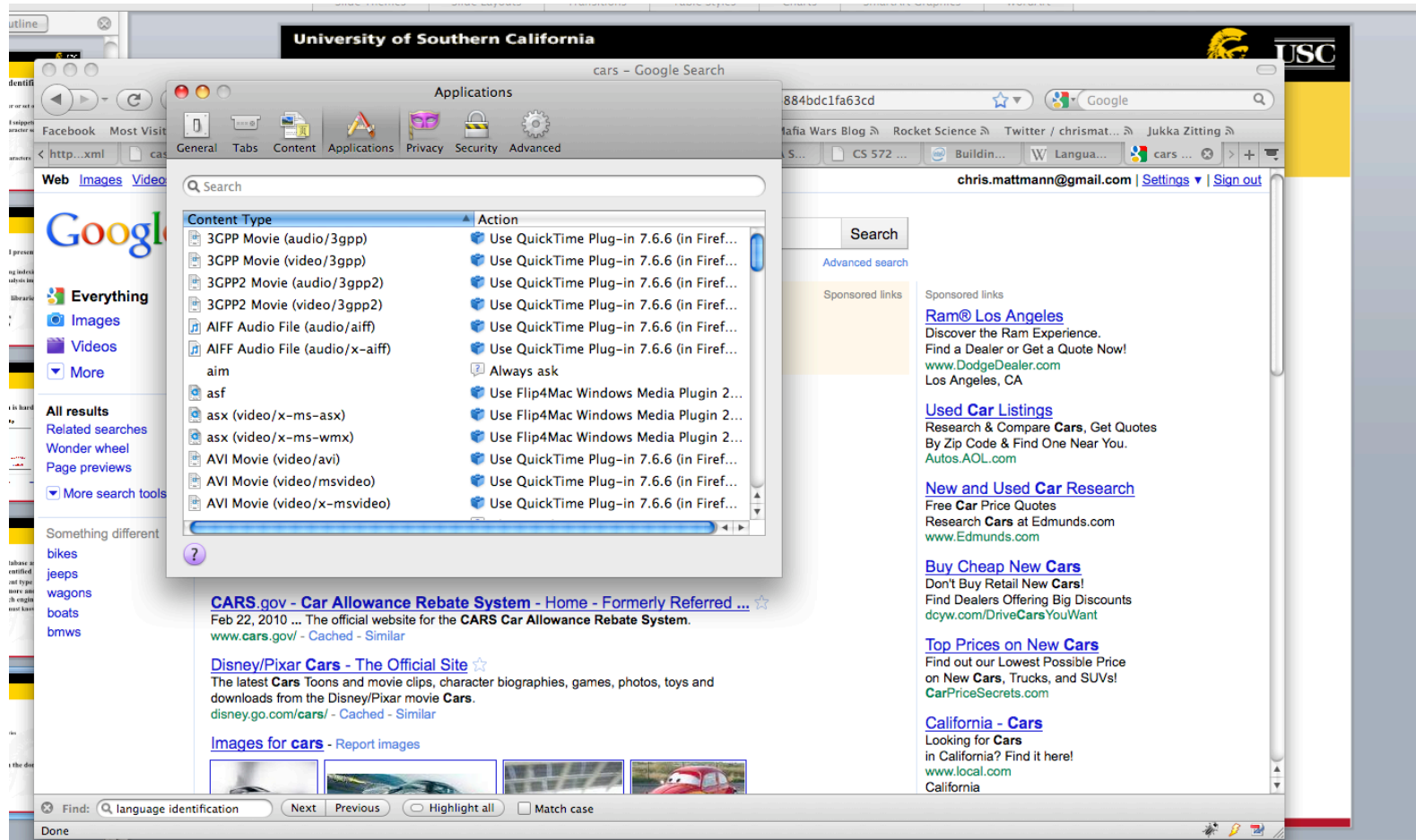
**Language identification - Wikipedia, the free encyclopedia** ☆  
Language identification is the process of determining which natural language given content is in. Traditionally, identification of written language - as ...  
[en.wikipedia.org/wiki/Language\\_identification](http://en.wikipedia.org/wiki/Language_identification) - Cached - Similar

**Language Identification Tools - How to Identify an Unknown Language** ☆  
Language identification tools can help you to identify a language from just a few sentences. This collection of language identification tools includes many ...  
[genealogy.about.com/.../language/Language\\_Identification\\_Tools.htm](http://genealogy.about.com/.../language/Language_Identification_Tools.htm) - Cached - Similar

**Language Identification Software - Basis Technology Products** ☆  
Determine the language and encoding of unstructured text.  
[www.basistech.com/language-identification/](http://www.basistech.com/language-identification/) - Cached - Similar

**[PDF] Language Identification of Encrypted VoIP Traffic: Alejandra y ...** ☆  
File Format: PDF/Adobe Acrobat - Quick View  
by CV Wright - Cited by 24 - Related articles  
success with language identification in this setting provides strong evidence for mandating

# Importance of content type detection



# Search Engine Architecture

