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

Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology

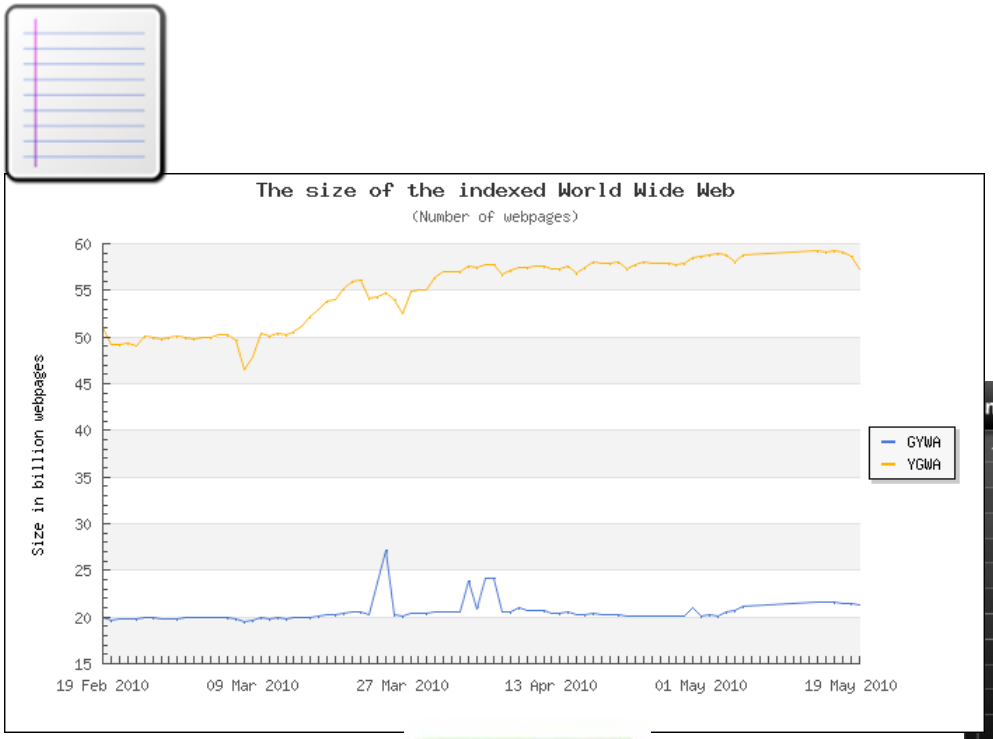
Outline

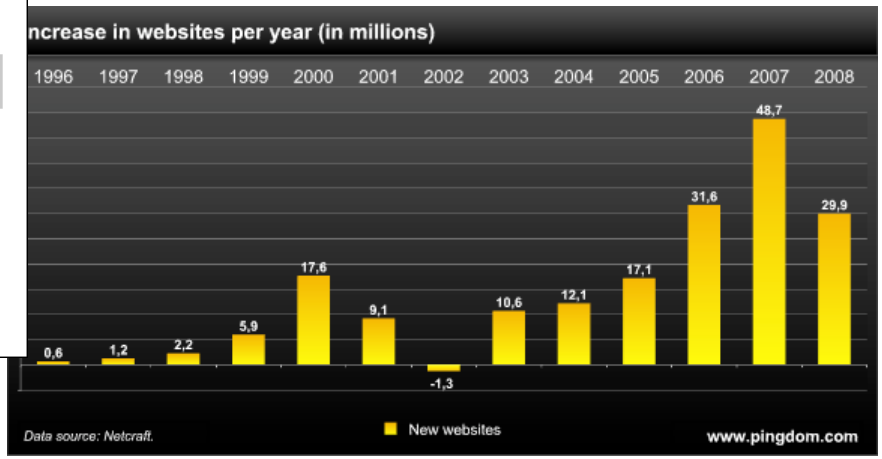
- The Information Landscape
- Importance of Content Detection
- Challenges
- Approaches
- Introduction to Apache Tika
- Using Apache Tika
- Wrap-up

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

The Information Landscape



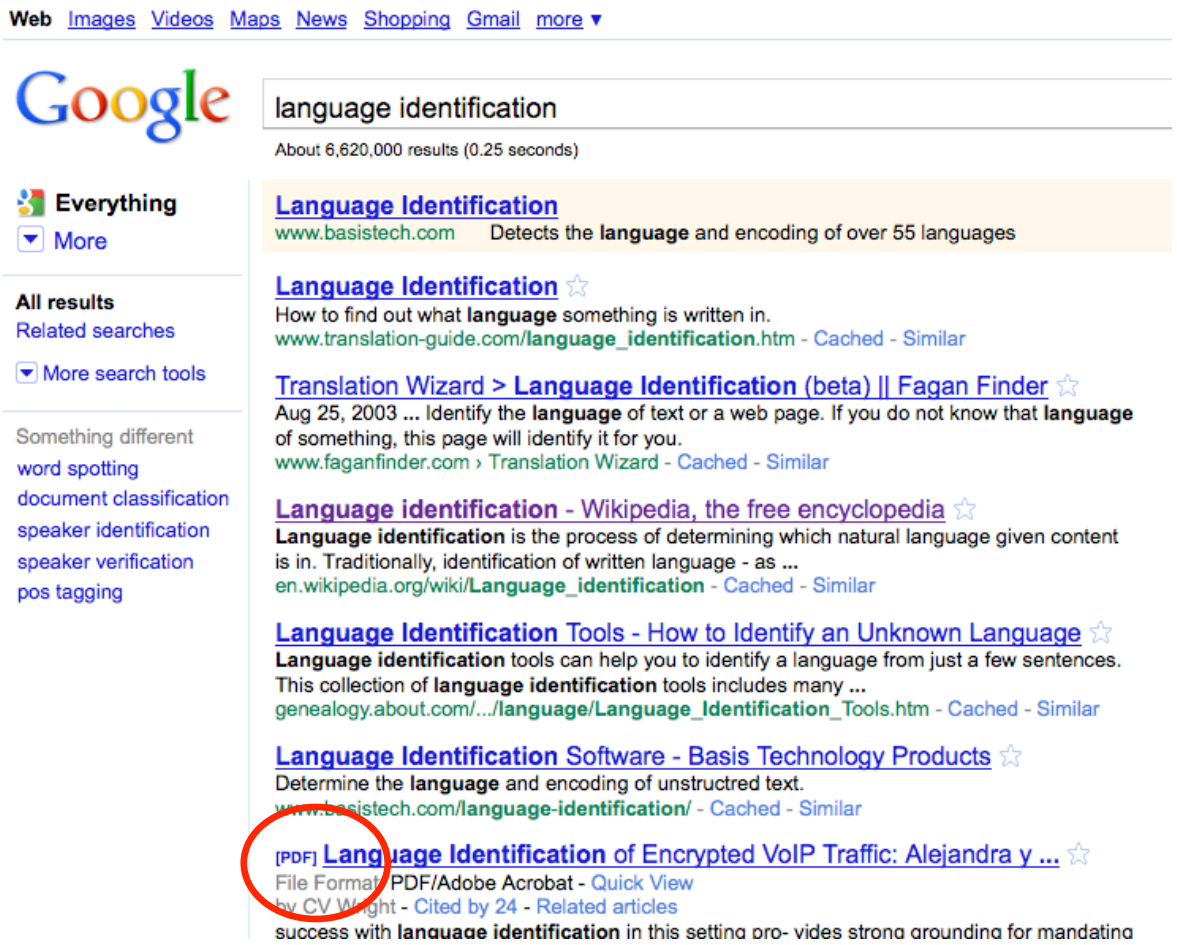


Proliferation of content types available

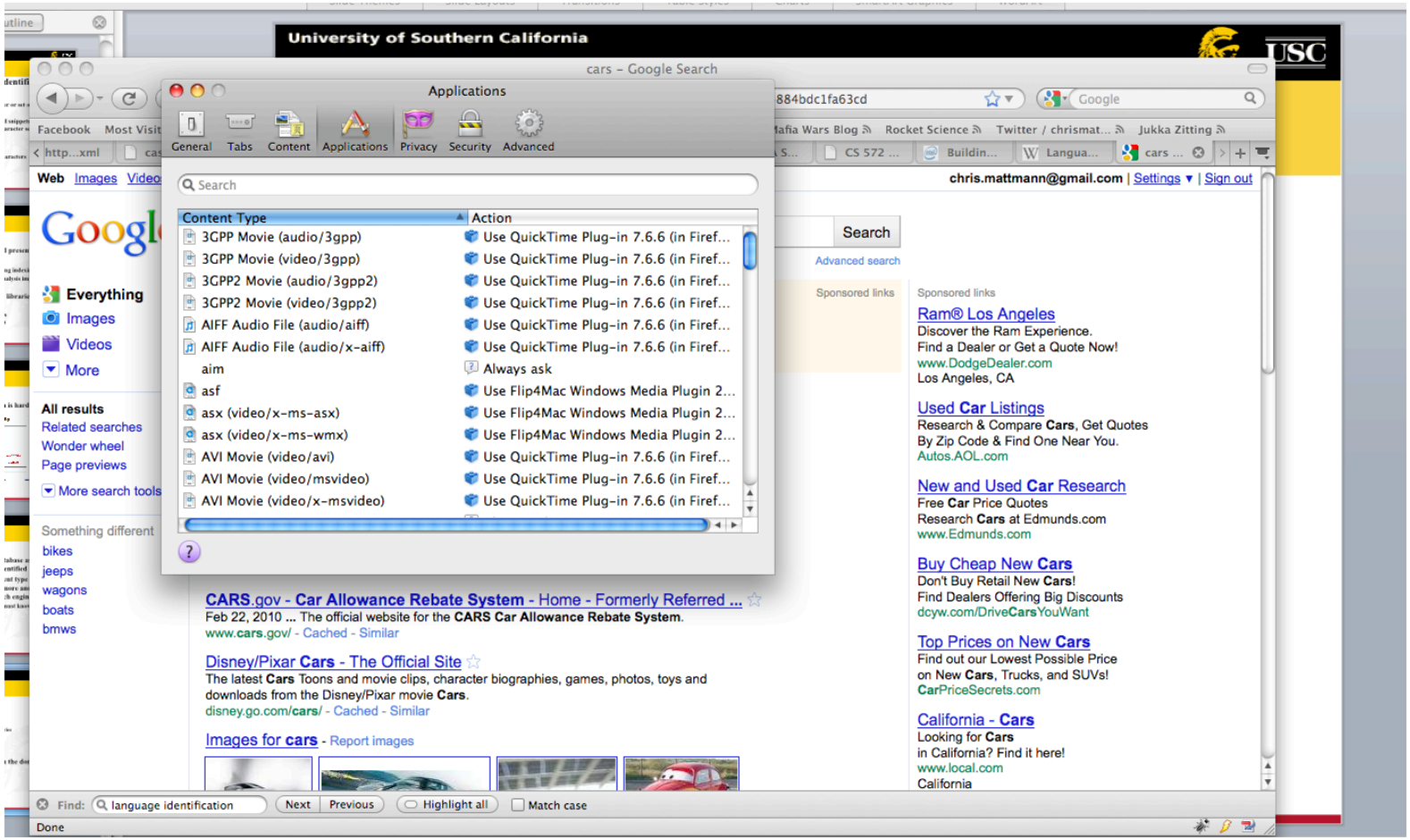
- By some accounts, 16K to 51K content types*
- What to do with content types?
  - Parse them
    - How?
    - Extract their text and structure
  - Index their metadata
    - In an indexing technology like Lucene, Solr, or Compass, or in Google Appliance
  - Identify what language they belong to
   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

*http://filext.com/

Importance of content types



Importance of content type detection



Search Engine Architecture