"Analyzing Big Datasets"

Acronyms

Team: Kevin Ding
Philipp Dumitrescu
Herman Leung

Mentor: Hossein Falaki

databricks

Overview

- **Overarching Goal**
  - Given an undefined acronym in a text, identify its meaning

- **Why is this an interesting problem?**

  Acronyms are
  - newly created all the time
  - not like regular words
  - ambiguous / opaque

- **Data science**
  - Unstructured text analysis on big datasets

Acronym Datasets

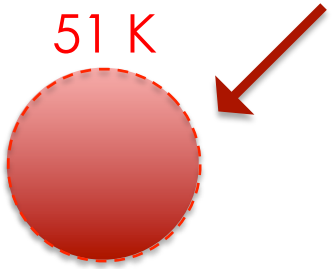
- Scraping / Regex Extraction

AF Acronym Finder

650 Web Pages

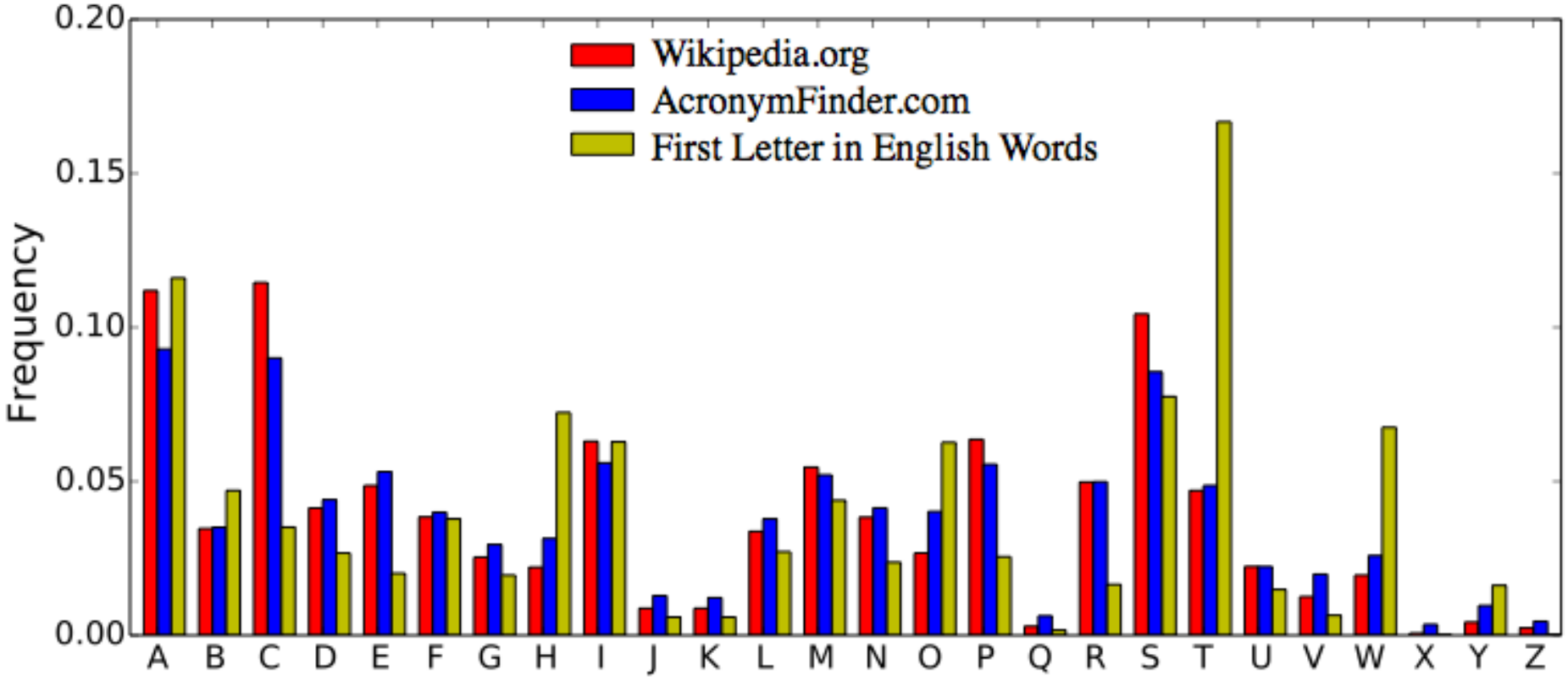
5 GB
raw text

Acronym-Definition Pairs
- San Francisco(SF)
- Quality Assurance (QA)



Unique Acronyms (ignoring definitions)

Letter Distribution



Usage

- Number of Articles Acronyms Appear In

- Top Acronyms

MP: 6743
NCAA: 1297
NFL: 1264
NHL: 1221

- Are English Words?

AF Acronym Finder 10 K

14 K

Simple Acronym Disambiguation

'SOA' voted most 'confusing acronym of the year'

By Joe McKendrick for Service Oriented | November 5, 2007

"This is a comprehensive tutorial that teaches fundamental and advanced SOA design principles, supplemented with detailed case studies and technologies used to implement SOA in the real world……"

Possible definitions

Sons of Anarchy

Service-oriented Architecture

State of the art

……

Content-based matching learning technqiue

Service-Oriented Architecture

Conclusions

- **Future development**
  - Identify more complex & new acronyms
  - Create a disambiguation engine
- **What we learned / challenges**
  - Data cleaning can drastically affect analysis
  - Exploration with parallel processing constraints
  - New tools (distributed computing and NLP)
- **Thanks to Hossein!**

TMA?

.....

Texas Medical Association

Transportation Management Area

Transparent Media Adapter

Tampa Museum of Art

Too Many Acronyms

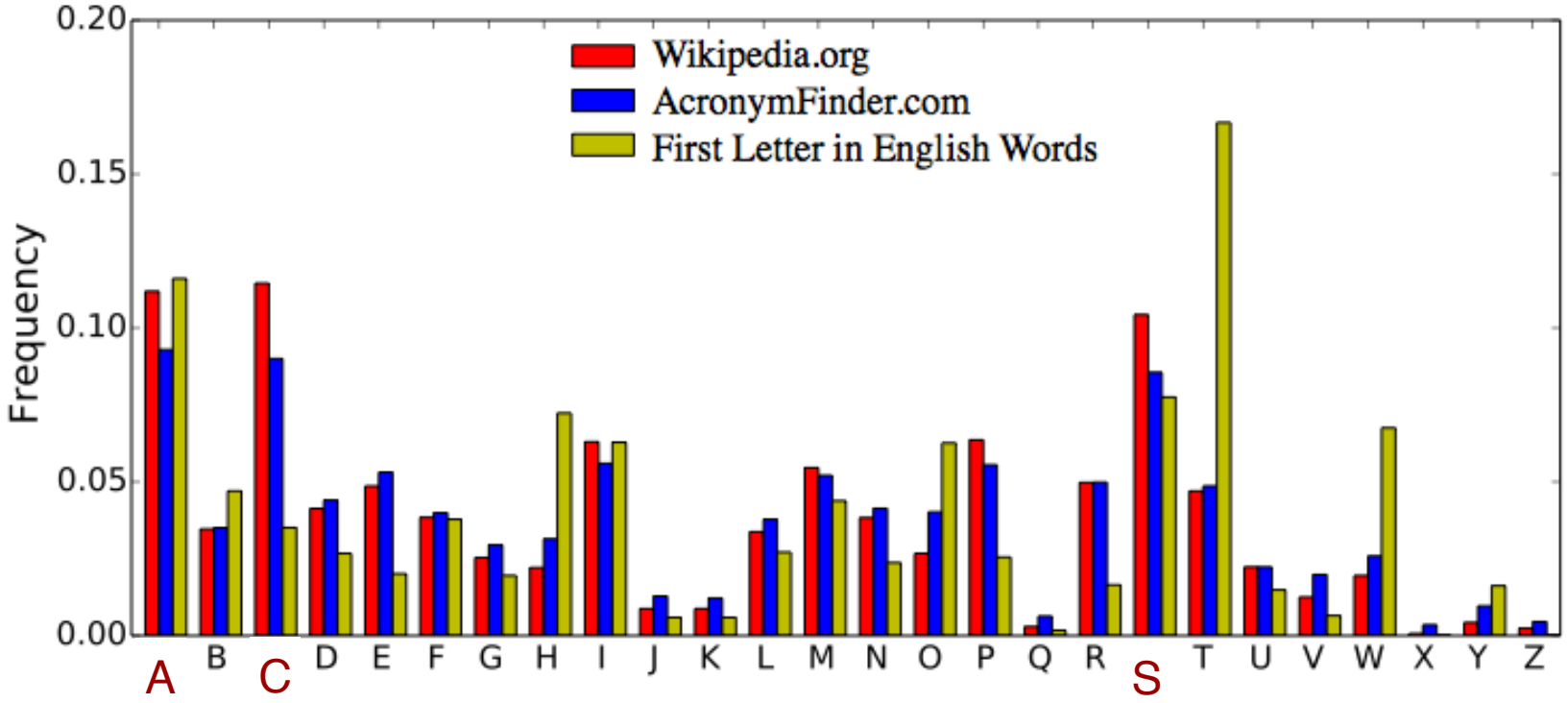
Tri Methyl Aluminum

Theoretical Mechanical Advantage

Total Material Assets

.....

Letter Distribution



Association
American
Air
……

Center
Council
College
,,,,,,

System
Society
School
…