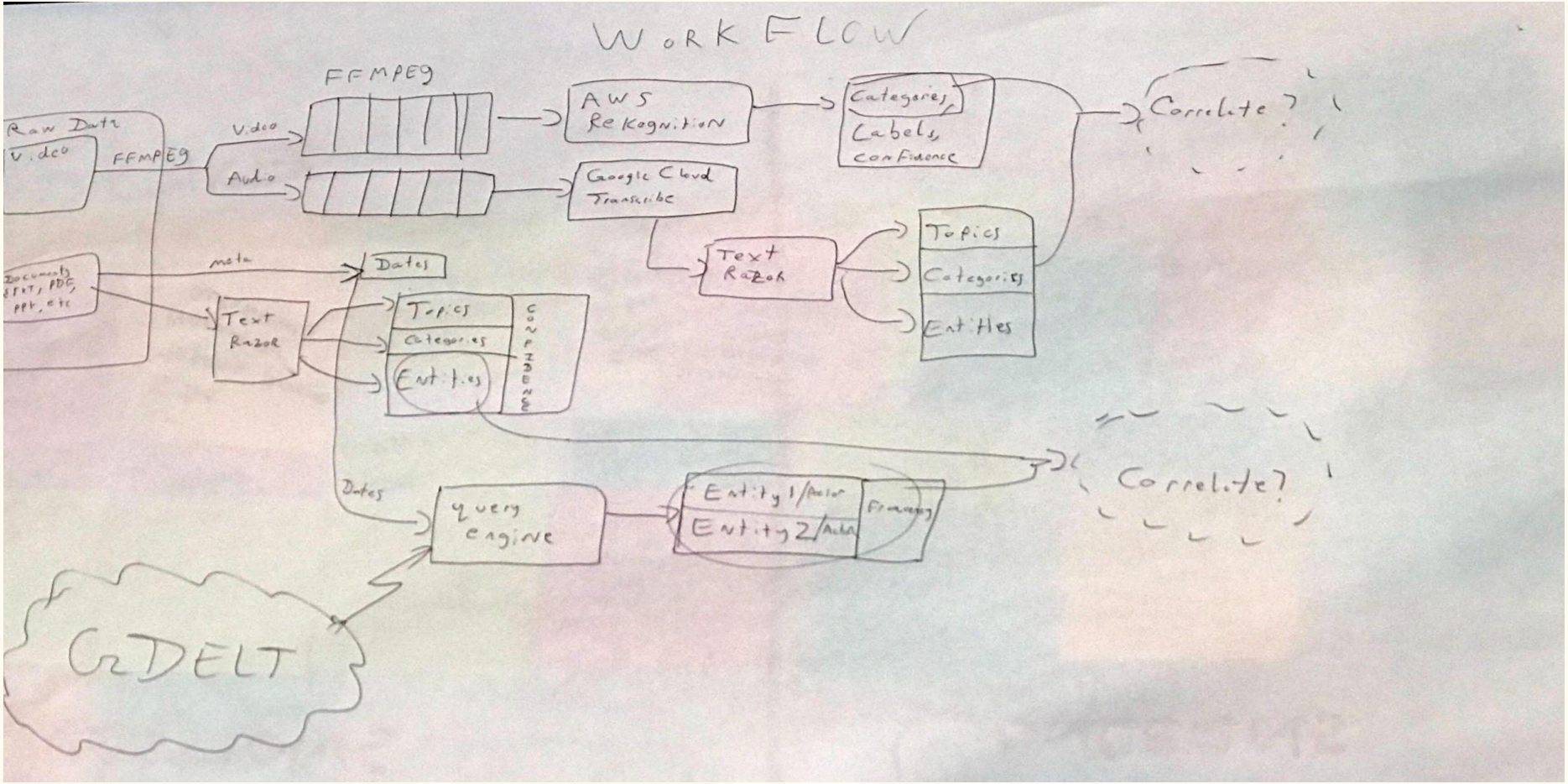
Data Science Futures Hackathon

Muntaser Syed
Jim Cannaliato
Sally Tan
Kent Huang

Workflow


WORKFLOW
Raw Data
Video
FFMPEG
Video
Audio
FFMPEG
AWS Rekognition
Categories
Labels, confidence
Correlate?
Google Cloud Transcribe
Text Razor
Topics
Categories
Entities
Documents (TXT, PDF, PPT, etc
meta
Dates
Text Razor
Topics
Categories
Entities
CONFIDENCE
Correlate?
Dates
query engine
Entity1/Actor
Entity2/Actor
Frames
GraphDB

Processing Times

- Video Processing:
  - 2 hrs 19 mins to create 20 second snapshots (one thread)
  - 44 minutes to perform label and text recognition (two threads)

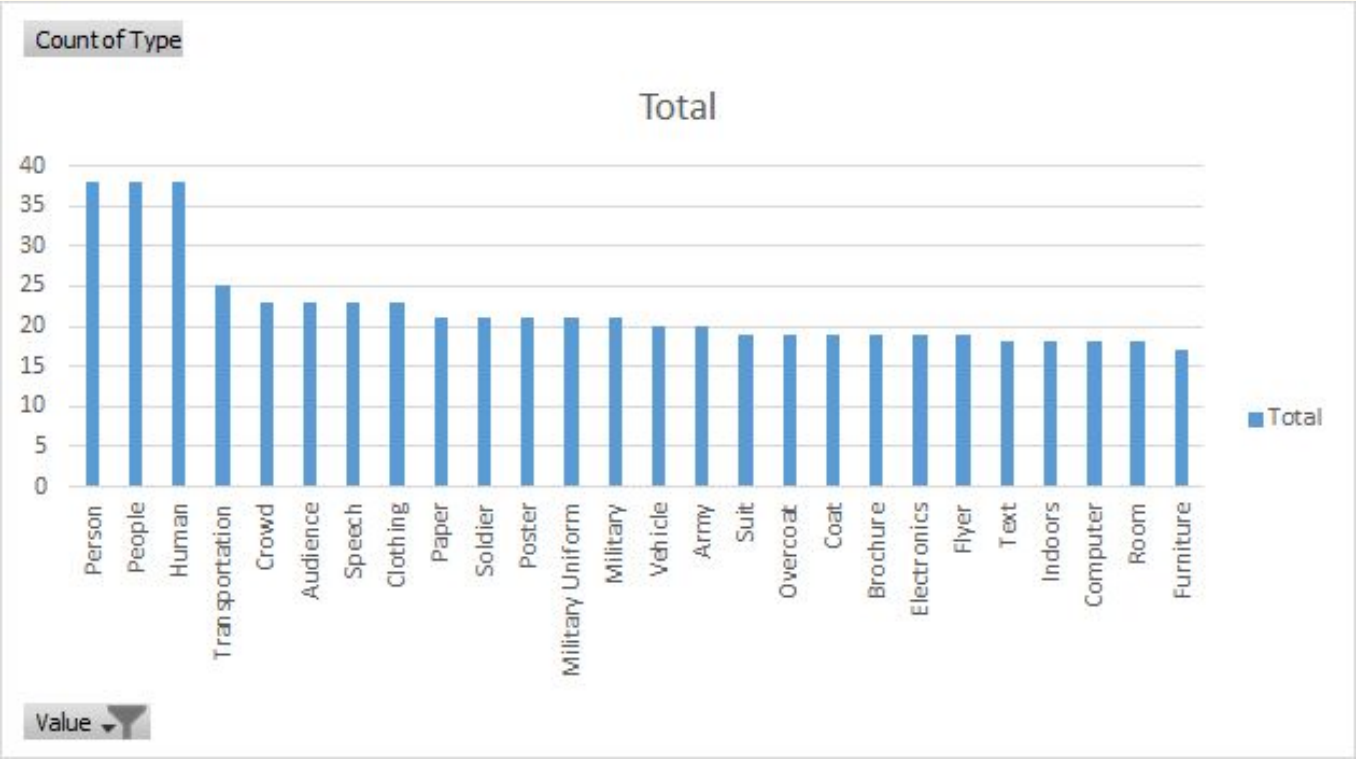
- Google Cloud Speech API
  - 380 seconds runtime for 18 minutes audio (actual)

- IBM Watson Speech to Text
  - 808 seconds runtime for 18 minutes audio (actual)

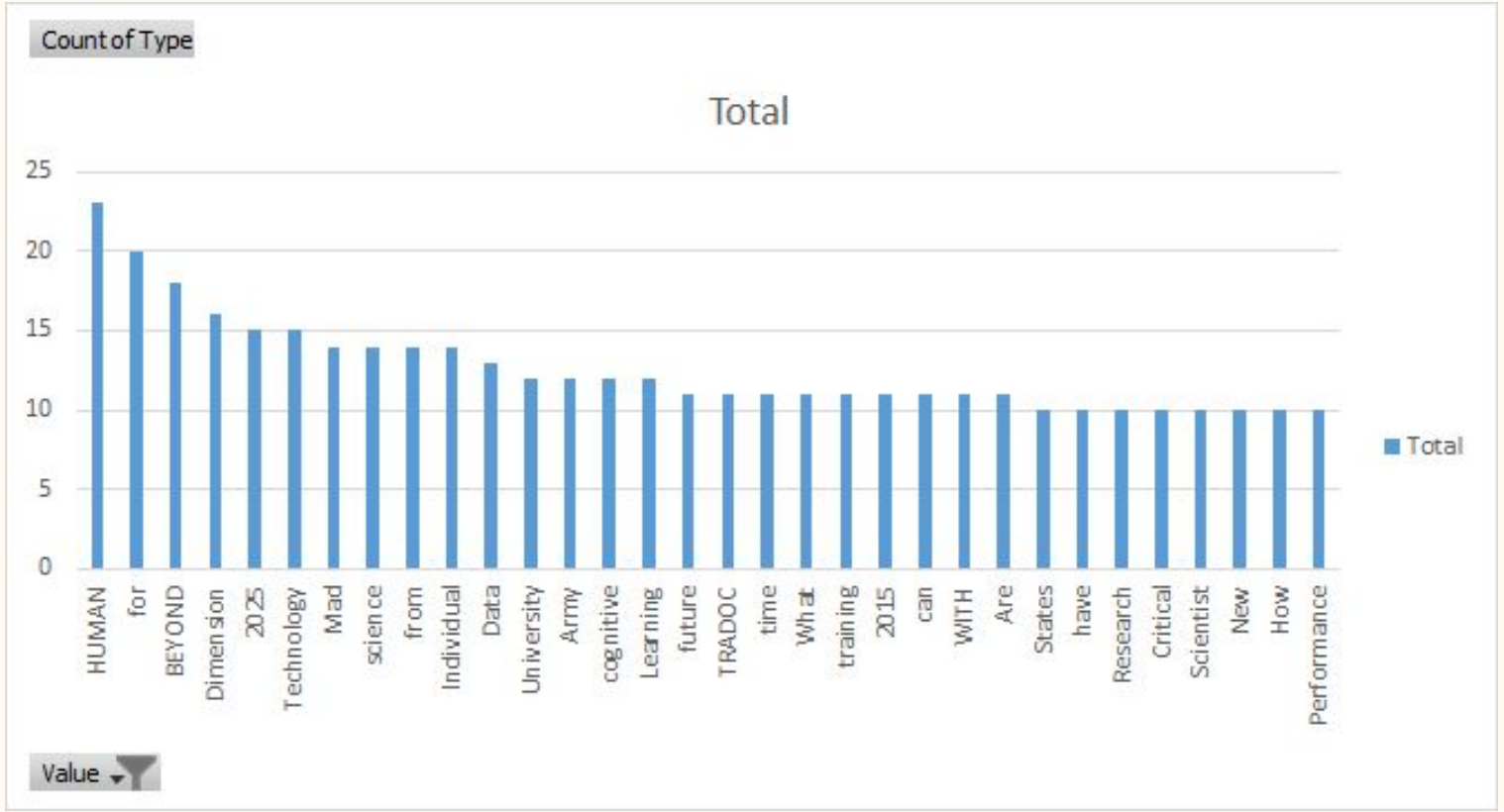
- Google Extrapolated
  - Total run time (actual) = 12 hours (= 720 min)
  - Transcription run time = 253 minutes = 4.2 hours

Sphinx run time (actual) >12 hours

Top Labels from Video Snapshot Processing



Top Text Found in Video Snapshots



Other interesting phrases found in the video

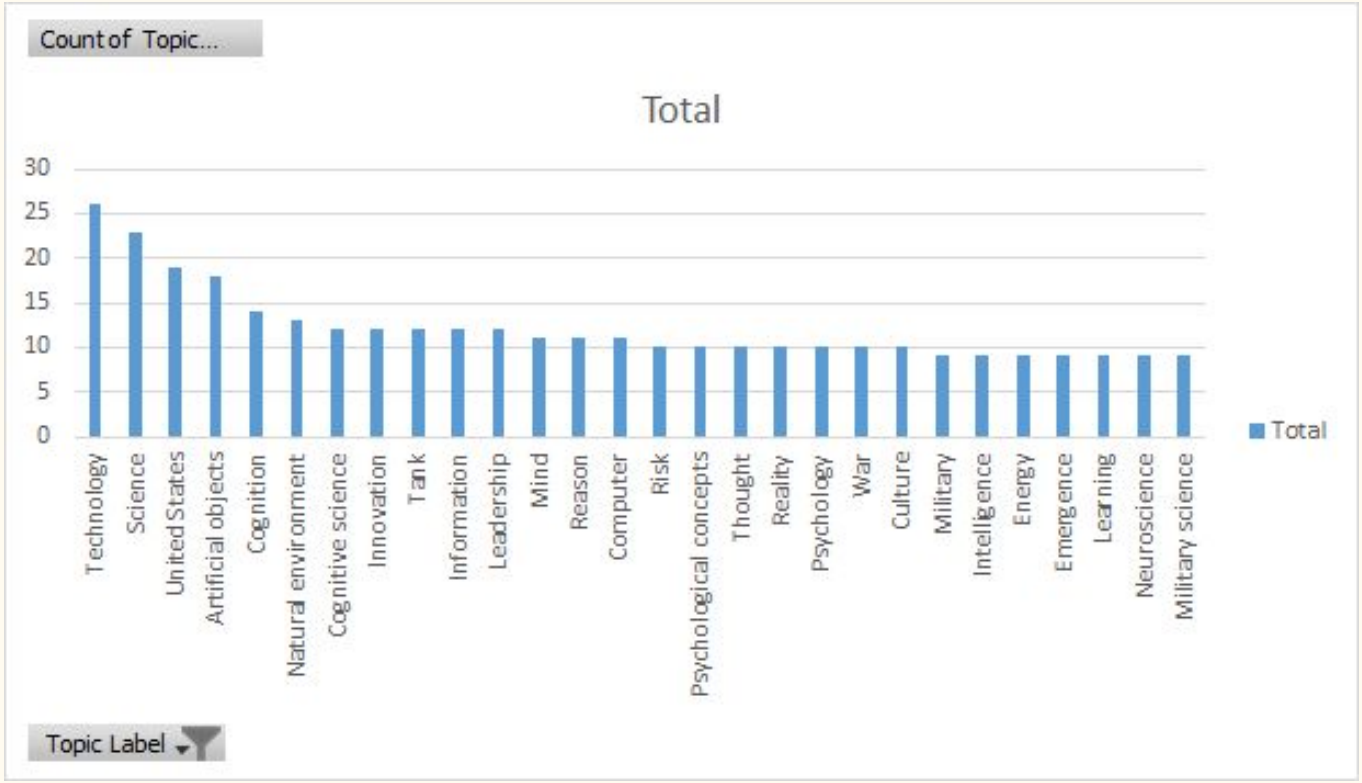
Topics/Titles

Minerals in Afghanistan Potential al value ron $420.9bil.
Lead for the Mad Scientist Initiative TRADOC
Robotics and artificial intelligence
Decision Science
ODNI Futures for Afghanistan (2008)
Chronic Fatigue Syndrome and Myalgic Encephalomyelitis
Patient-Focused Development Initiative
Biological Motion Perception

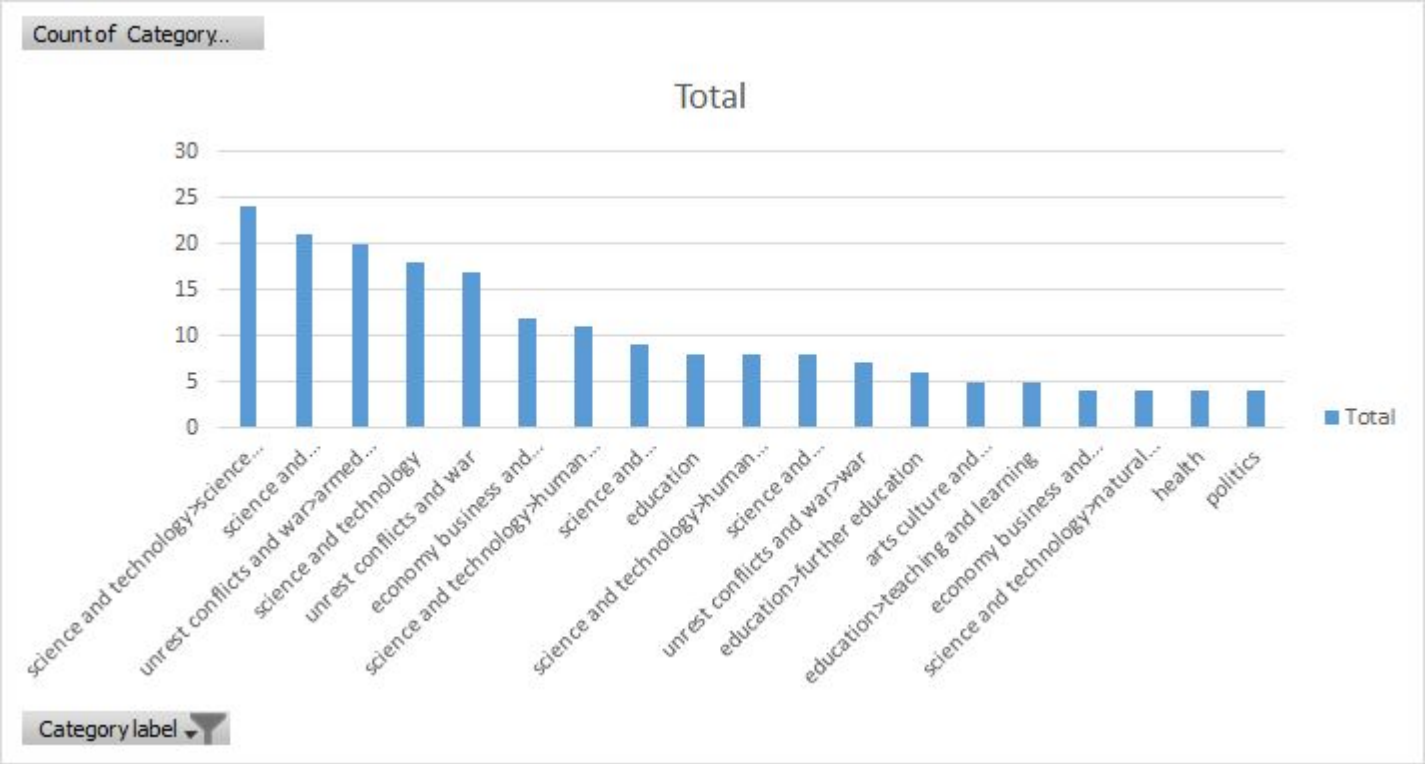
People's Names

General David G. Perkins
Dr AIbert Palazzo
Eliabethl Chalecii PhD
originally developed by Barry B. Hughes
Dr. Baruch Fischhoff
von Neumann & Morgenstern (Princeton)
BY PAUL F. LAZARSFELD
Howard Heinz Professor
Alan Baddeley
Norman Mackworth
H. M. VERNON M.A M.D.
BG (Ret) Pete Palmer
Dr. Brett Piekarski

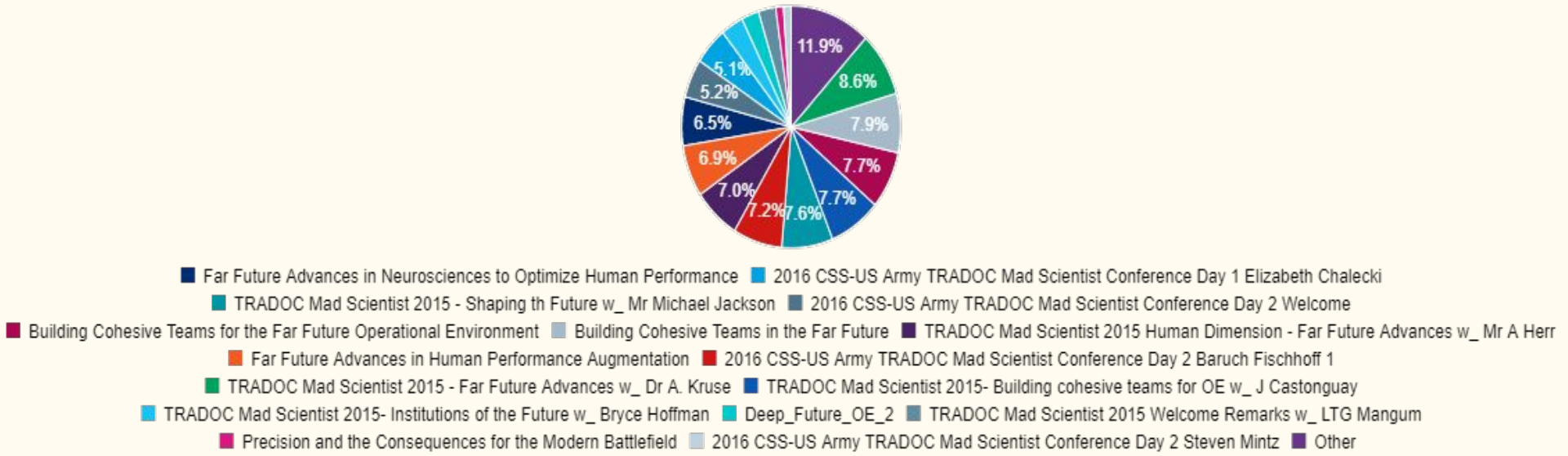
Top Topics found with TextRazor



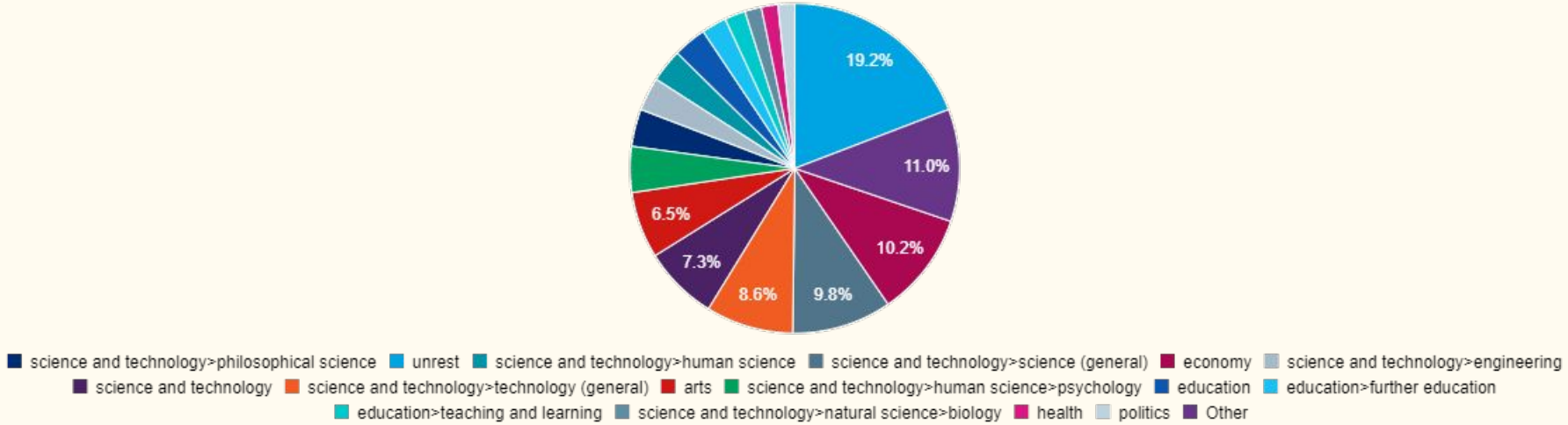
Top Categories found with TextRazor



Video Tag Pie Chart



Category Analysis From Transcripts



GDELT Date Analysis Example

Video: INFORMATION PAPER Mad Scientist Megacities (final)_20151221.pdf

Date: 2015-12-21

Entities/Actors for that date:

['UNITED STATES', 'POLICE', 'GOVERNMENT', 'UNITED KINGDOM', 'ISRAEL', 'PRESIDENT', 'CHINA', 'RUSSIA', 'COMPANY', 'SYRIA', 'SCHOOL', 'NIGERIA', 'AUSTRALIA', 'CANADA', 'FRANCE', 'TURKEY', 'STUDENT', 'IRAN', 'AFGHANISTAN', 'AFRICA', 'BUSINESS', 'AMERICAN', 'IRAQ', 'MEDIA', 'OBAMA', 'COMMUNITY', 'PAKISTAN', 'RUSSIAN', 'ISRAELI', 'PHILIPPINE']

['UNITED STATES', 'GOVERNMENT', 'POLICE', 'ISRAEL', 'SYRIA', 'UNITED KINGDOM', 'CHINA', 'RUSSIA', 'SCHOOL', 'PRESIDENT', 'COMPANY', 'NIGERIA', 'AUSTRALIA', 'IRAQ', 'TURKEY', 'AFGHANISTAN', 'CANADA', 'STUDENT', 'BUSINESS', 'AMERICAN', 'SYRIAN', 'MUSLIM', 'AFRICA', 'FRANCE', 'IRAN', 'ISLAMIC', 'COMMUNITY', 'MEDIA', 'PAKISTAN', 'HOSPITAL']

Going Further

Smart snapshots using scene detection

Text cleanup from video extraction

Filtering data from GDELT

Correlation with filtered GDELT data

Correlation with video and audio extraction

Predictive analysis