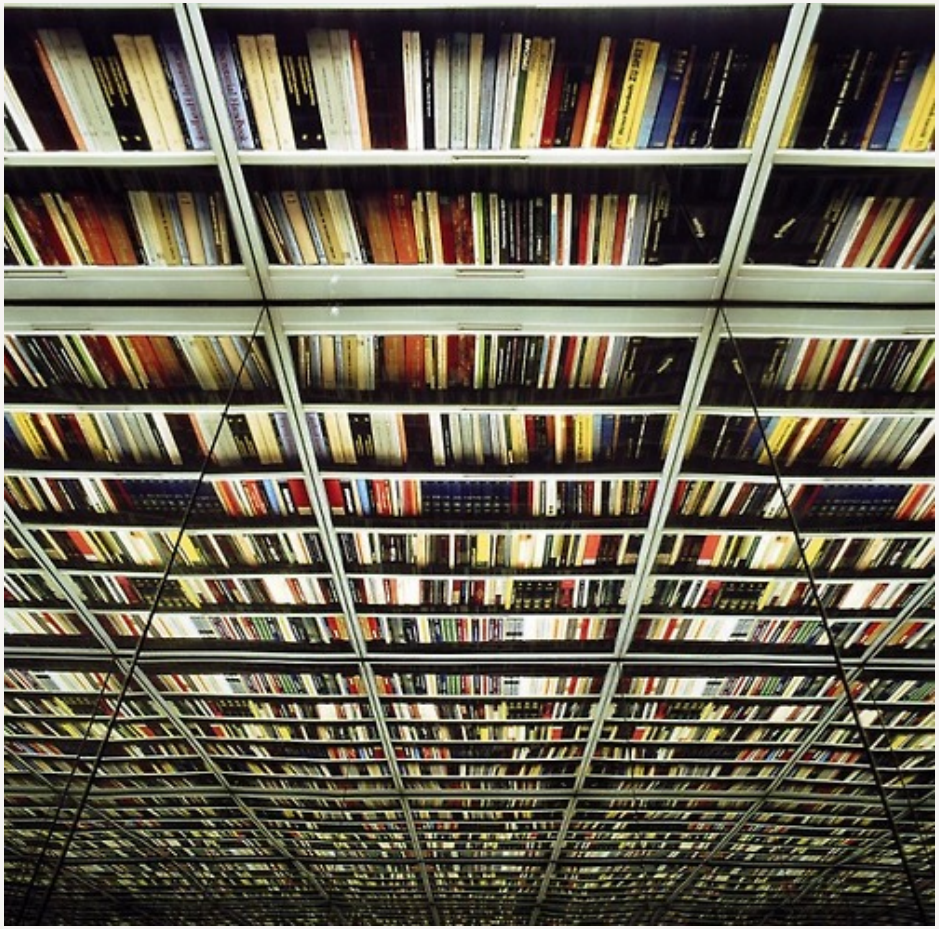
Natural Language Processing and Information Retrieval - Major Tasks and Applications

Danas Zuokas
Vilnius University

The Problem 1

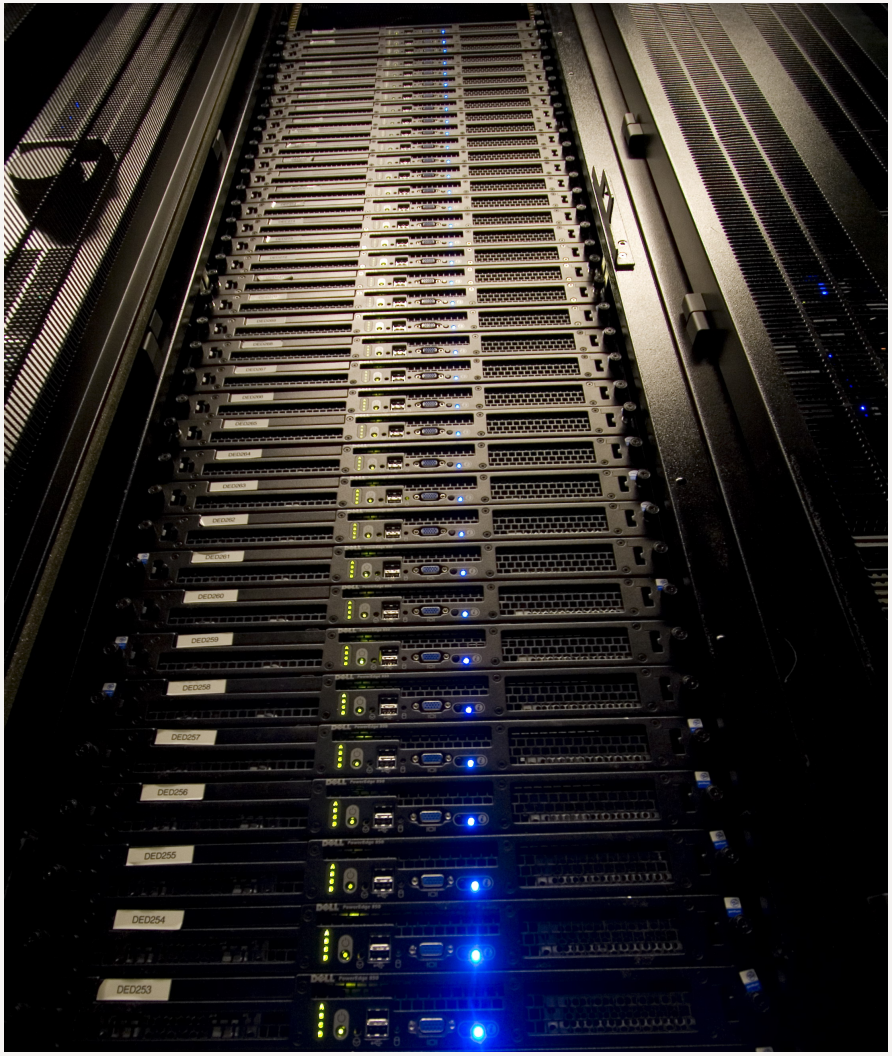


- The amount of data produced and stored keeps growing exponentially.
- Majority of that data is available in an semi- and unstructured form (text).
- Textual data contains significant business value.
- Effectively analyzing this type of data is of paramount importance.
- Traditional Information Technology techniques are not sufficient.

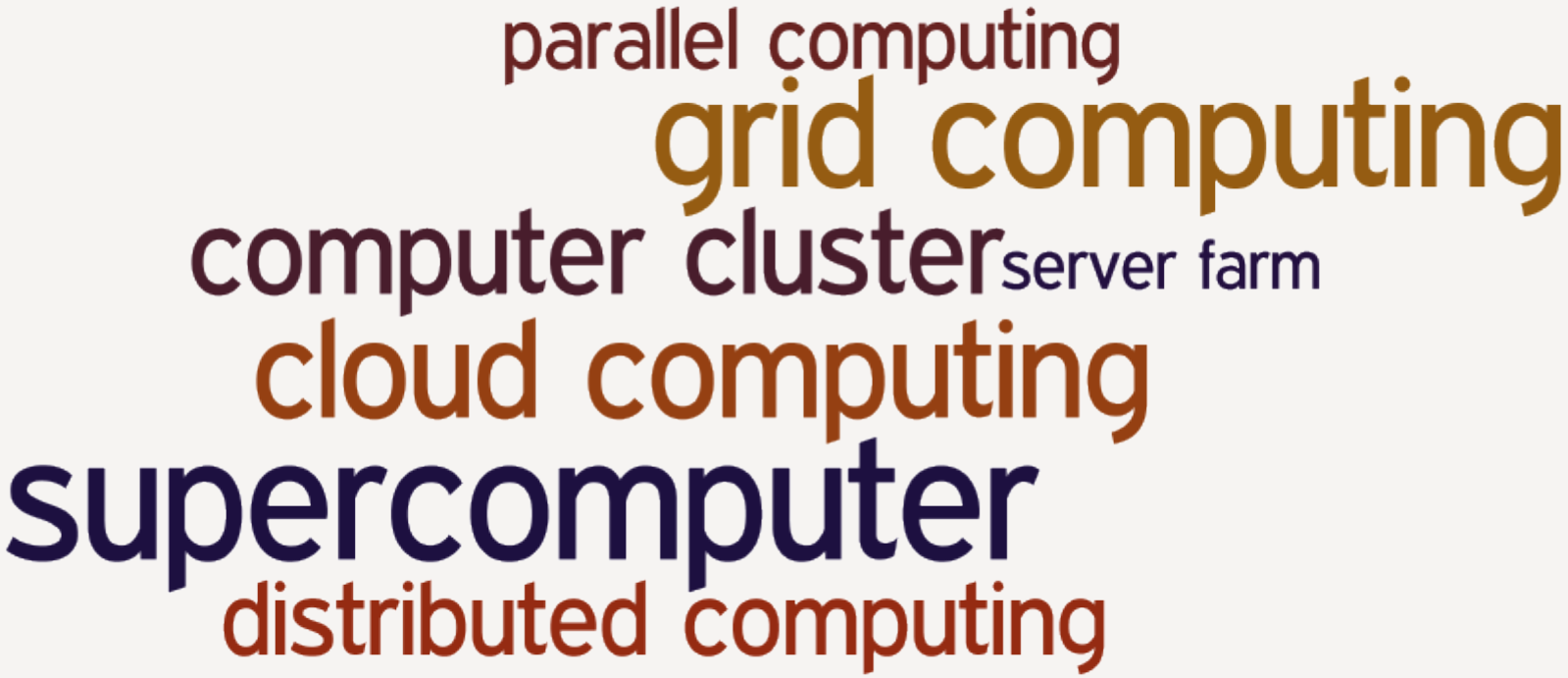
The Solution

The Problem 2

- The volume of data that is being employed by text analytics methods is vast.
- This requires a huge storage and computing infrastructure.
- It is very complex and expensive for a separate user to implement and maintain such systems.
- Actors with the advanced state-of-the-art scientific methodology knowledge and computing infrastructure are capable of providing text analytics service.



The Solution



A Model System

- Consider a system which deals with

  events as they are reported on the Web.
- The event is a segment of information describing

  who does what and when.
- On the one side a **service provider** would create a data pool of events.
- On the other side a user (**SME, public or other organization**)

  would get the **query-relevant events** with additional service:
  - **credibility** – a measure of importance of the event,
  - **sentiment** – opinions gathered and analyzed after the event was reported,
  - **associations** – a network of related events for further exploration.

2 Examples

view AUSTRALIA

Photo G

ATURES TIPS GUIDES DIRECTORY | NEWSLETTER

PRODUCT NEWS

Nikon Unveils D5100

April 5, 2011: Nikon has just released details of its latest DX-format DSLR camera, the D5100.

Replacing the D5000 and sitting between the D3100 and D7000, the new camera is 10% smaller than its predecessor. It shares the same 16.2-megapixel sensor as the D700 and boasts the next-generation EXPEED 2 image processing engine. The shutter mechanism is rated for 100,000 cycles. It's targeted at entry-level photographers who want a higher-featured camera than the D3100. The D5100 boasts a new vari-angle monitor with a side hinge that provides greater flexibility for shooting from different angles.

washingtonpost.com > World

» THIS STORY: READ + | WATCH + | Comments

Obama tells troops in Afghanistan: Success is within reach

By Ernesto Londono and Perry Bacon Jr.
Washington Post Foreign Service
Friday, December 3, 2010; 8:54 PM

KABUL - President Obama told U.S. troops in Afghanistan that success is within reach during a brief, unexpected visit Friday night that came amid a series of embarrassing setbacks in the effort to turn around the faltering war.

THIS STORY
- » Obama tells troops in Afghanistan: Success is within reach
- President Obama makes surprise visit to Afghanistan
- Obama, in Afghanistan, says U.S. will succeed

GALLERY

President Obama makes surprise visit to Afghanistan

On an unannounced visit, Obama praised U.S. troops Friday and said his administration's surge over the past year has weakened the Taliban.

» LAUNCH PHOTO GALLERY

Obama landed almost exactly a year to the day after he authorized a 30,000-troop surge, and as his advisers finalize a comprehensive assessment of whether the expanded U.S. military presence is working.

With few decisive signs of progress to show at a time when many Americans are growing impatient with the war, the administration is once again reexamining its strategy to marginalize extremist groups and prop up the anemic Afghan government.

top Network News PROFILE

Recommend

Tweet 0

View More Activity

TOOLBOX

Resize Print E-mail

Named Entity Recognition

COGNITIVE COMPUTATION GROUP
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Research People Software Demos Publications Resources Schedule

DEMOS

Problems? Email mssammon@illinois.edu

Named Entity Recognition Demo Results

The Named Entity Recognizer has identified the following named entities.

[PER Obama] tells troops in [MISC Afghanistan: Success] is within reach
By [PER Ernesto Londono] and [PER Perry Bacon Jr] .
[ORG Washington Post Foreign Service]
Friday, December 3, 2010; 8:54 PM
KABUL - President [PER Obama] told [LOC U.S.] troops in [LOC Afghanistan] that success is within reach during a brief, unexpected visit Friday night that came amid a series of embarrassing setbacks in the effort to turn around the faltering war.
[PER Obama] landed almost exactly a year to the day after he authorized a 30,000-troop surge, and as his advisers finalize a comprehensive assessment of whether the expanded [LOC U.S.] military presence is working.
With few decisive signs of progress to show at a time when many [MISC Americans] are growing impatient with the war, the administration is once again reexamining its strategy to marginalize extremist groups and prop up the anemic [MISC Afghan] government.

Key:

- PER - Person
- ORG - Organization
- LOC - Location
- MISC - Miscellaneous

Event Extraction

When: April 5, 2011

Who: Nikon

Did what: has just released details

About what: of its latest DX-format DSLR camera, the D5100.

Who: President Obama

Did what: makes surprise visit

Where: to Afghanistan

When: at Friday, December 3, 2010

Sentiment Analysis

04-05-2011 04:02 PM #7

KmH ○

TPF Junkie!

Join Date:	Apr 2009
Location:	Iowa
Posts:	12,800
My Gallery:	(0)
My Photos Are OK to Edit	
Liked:	179 times

The D5100 has a lower quality AA filter than the D7000 has.

Not only does the D7000 have better high ISO performance, it will also have better image quality than the D5100.

Interestingly you failed to mention two of the D7000's most important features, it's Multi-CAM 4800 DX, 39 focus point (9 cross-type) auto focus module, and it's new 2016 pixel, 3D, RGB metering sensor.

The D5100 is still using the 420 pixel RGB metering sensor, and the Multi-CAM 1000, 11 focus point (1 cross-type point) auto focus module.

Share

. .Keith. . . .

Check out my TPF - ***Spring Cleaning Sale*** - Nikon, Sekonic, Wacom, X-RITE (Only the Nikon stuff is still available, the rest

04-05-2011 07:48 PM #10

blackxthink ○

TPF Noob!

Join Date:	Apr 2011
Location:	Quebec, canada
Posts:	10
My Gallery:	(0)
My Photos Are NOT OK to Edit	
Liked:	0 times

d7000 + tokina 11-16mm 2.8 for extra wide

Depends on how much money you wanna spend.

The better gear you get at start , it will take longer for you to upgrade it.

I'd still get d7000 over d5100 Only because it's a higher end camera.

But if you have a really restricted budget , id get d5100 + cheaper lens

Share

Event Grouping

Obama landed almost exactly a year to the day after he authorized a 30,000-troop surge, and as his advisers finalize a comprehensive assessment of whether the expanded U.S. military presence is working.

the Taliban.
» LAUNCH PHOTO GA

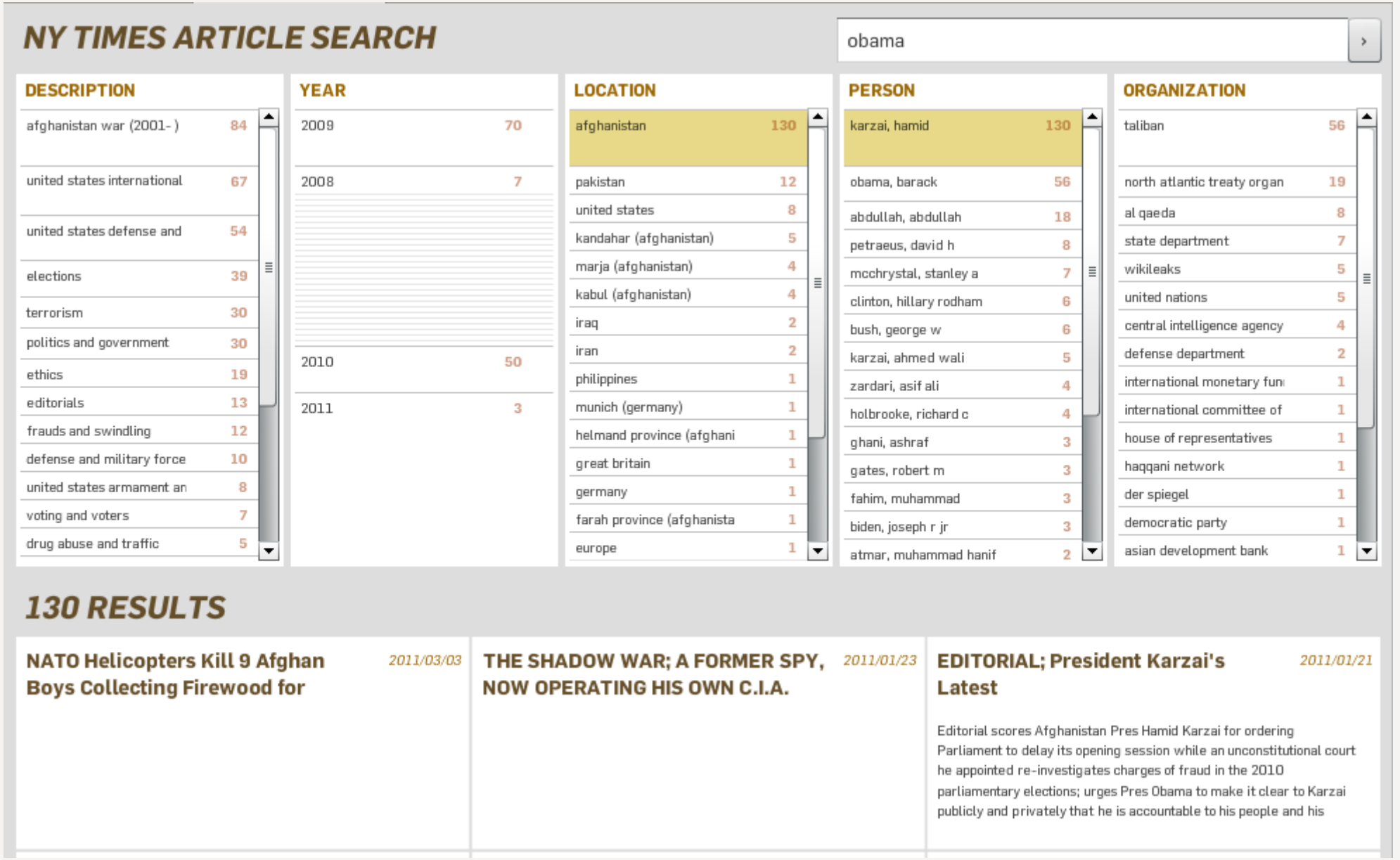
The U.S. effort has been marred by reversals and uncomfortable revelations in recent weeks. Last month, Afghan President Hamid Karzai angered U.S. officials after he said in an interview that U.S. troops were being overly aggressive in their pursuit of the Taliban.

Friday's visit followed by a day the release of U.S. diplomatic cables that chronicled rampant corruption in the Afghan government and the toll that has taken during a war that is in its 10th year.

The weather in and around Kabul was slightly hazy Friday afternoon, but it appeared to clear up by nightfall, when the president landed. Karzai and Obama met for roughly an hour last month in Lisbon during a NATO summit.

The White House expects to complete the process by Dec. 18, before Obama travels to Hawaii for the holidays.

Faceted Search



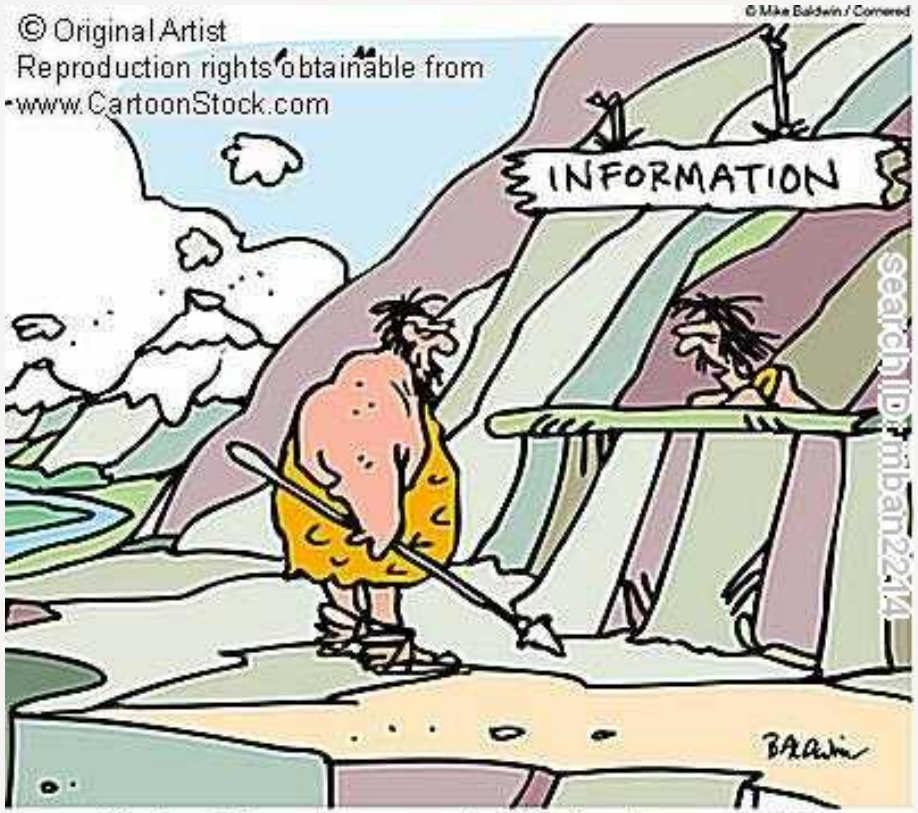
Impact

The user would benefit in saving operational costs when exploring web content. The service would benefit SME's, public and other organizations. To give a few examples:

- **Content and news providers** would use data pool to add functionality to their publications.
- The tool would help **governmental bodies** to evaluate the feedback from the citizens after the decisions are made.
- **Marketing/Public Relations companies** would use event pool to evaluate outcomes of their campaigns.
- **Predictive analytics companies** would use event pool to create predictive models (e.g. portfolio manager). Some events change sentiment on the market which then moves asset price.

Conclusions

- **Similar services** that would use text analytics tools also are and will be developed.
- Such systems are very **complex and expensive** in order to be implemented and maintained by separate users on their own.
- Only actors with the knowledge of advanced state-of-the-art **scientific methodology and computing infrastructure** are capable of creating such systems.
- Users would identify the demand for **specific functionalities** in such a way determining “the shape” of a service.



"I don't have any yet. We just opened."