News Categorization

Sarath Joseph

Saurabh Mehta

Index

- Introduction
- Project Description
- Semantic Representation of Text
- Topic Modelling
- Discover topics from a corpus
- LDA
- Tools
- Project Roadmap
- Architecture
- Topic Results
- Evaluation
- Future Work
- Bibliography

Introduction

- as more information becomes available, it becomes more difficult to find and discover what we need

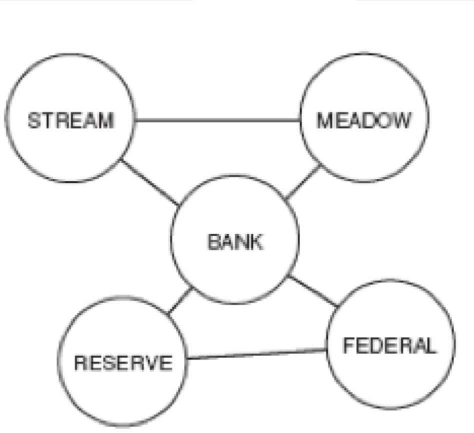
- we need a way to help us to organize and understand these vast amounts of information

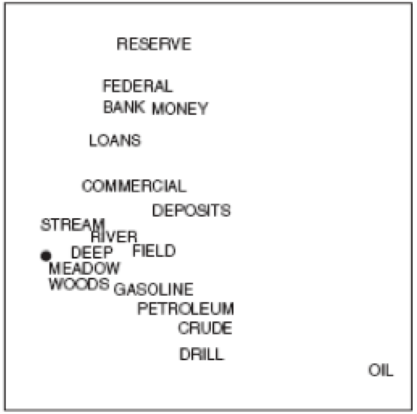
- ex. google books

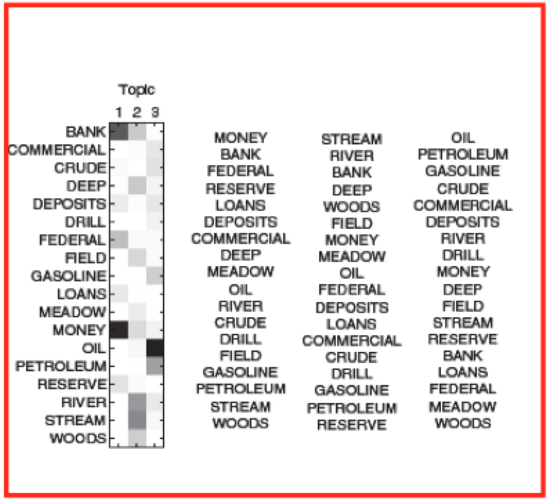
Project Description

- We plan to develop a system to aggregate and classify news from diverse sources using machine learning algorithms and techniques.
- We develop the system from the point view of an online news startup.
- As a potential startup that works to aggregate news from multiple sources, it is essential to categorize news coming in from various sources.
- There are several existing ways to achieve this ranging from automatically sorting scraped articles into categories based on keywords or scraping individual parts of the website for categories to obtaining tag based news using News APIs.
- However, dynamic tagging of news data can be challenging. The strong advent of social media platforms as news sources can also be exploited to obtain trending information and for buzz monitoring and machine learning techniques can be employed to generate easily accessible news ranging a wide span of topics.

Semantic Representation of Text







MONEY	STREAM	OIL
BANK	RIVER	PETROLEUM
FEDERAL	BANK	GASOLINE
RESERVE	DEEP	CRUDE
LOANS	WOODS	COMMERCIAL
DEPOSITS	FIELD	DEPOSITS
COMMERCIAL	MONEY	RIVER
DEEP	MEADOW	DRILL
MEADOW	OIL	MONEY
OIL	FEDERAL	DEEP
RIVER	DEPOSITS	FIELD
CRUDE	LOANS	STREAM
DRILL	COMMERCIAL	RESERVE
FIELD	CRUDE	BANK
GASOLINE	DRILL	LOANS
PETROLEUM	GASOLINE	FEDERAL
STREAM	PETROLEUM	MEADOW
WOODS	RESERVE	WOODS

Topic Modelling

Topic modelling provides methods for automatically organizing, understanding and summarizing large electronic archives.

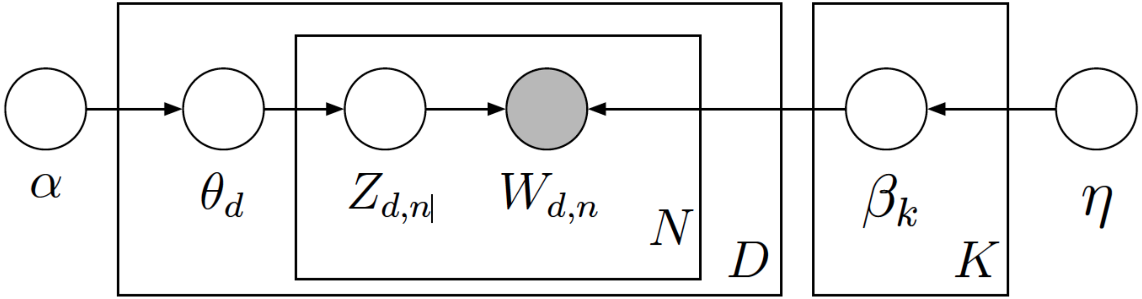
1. Discover the hidden themes that pervade the collection.
2. Annotate the documents according to those themes.
3. Use the annotations to organize and summarize the texts.

Documents are mixture of topics and a topic is a probability distribution over words.

Discover topics from a corpus

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

LDA



Tools

- Mallet (Java)
- Flask (Python Microframework)
- MongoDb

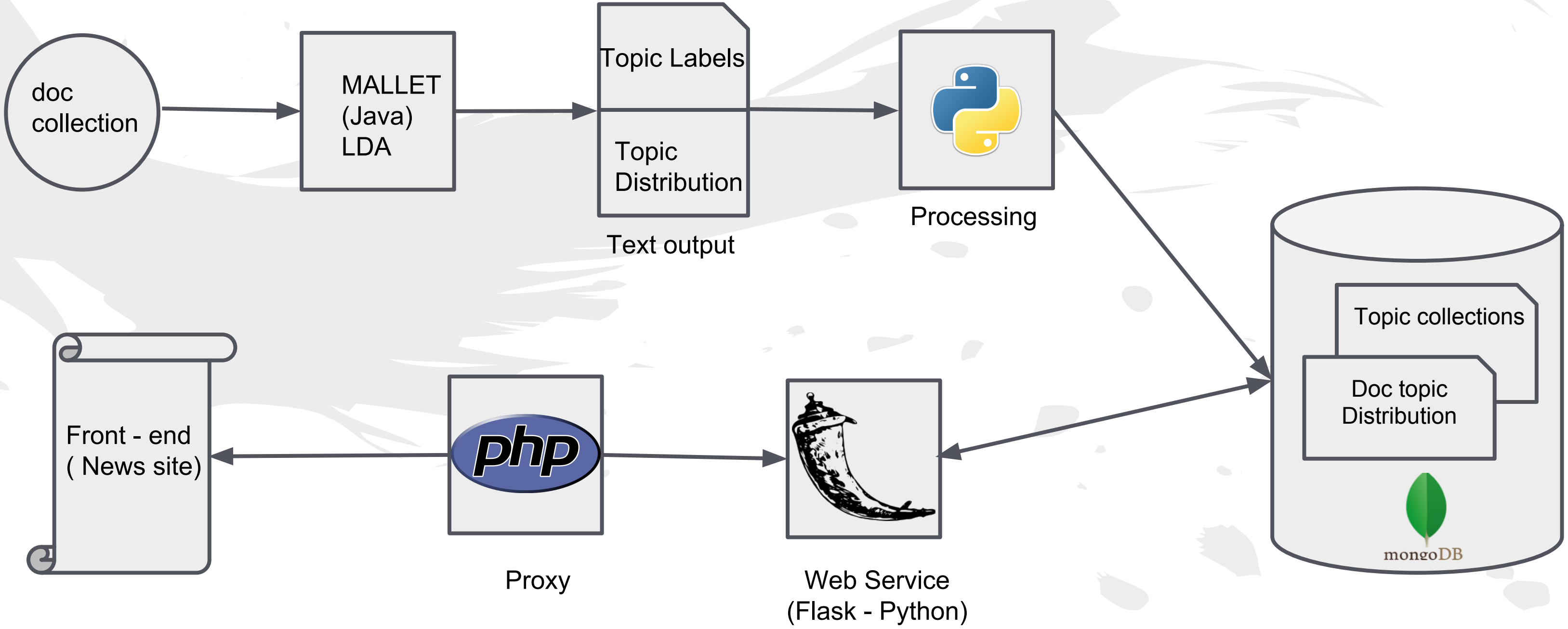
Python NewsPaper API

Pymongo for MongoDb

Roadmap

- Extraction of News articles using Python NewsPaper API (Text and metadata) and storing in Mongodb
- Feeding the article text into Mallet
- Parse the Mallet output using Python
- Create new collections in Mongodb with Parsed output
- Creating a Web service for topic based access
- Labelling of Topics
- Creating a client for consumption of news web-service

Architecture



Topic Results

Topic 0 - Middle East
Topic 1 - International
Topic 2 - Justice
Topic 3 - Showbiz
Topic 4 - Life
Topic 5 - Stories
Topic 6 - Travel
Topic 7 - Politics
Topic 8 - Education
Topic 9- The UK and India
Topic 10- Religion
Topic 11- Food
Topic 12- Health
Topic 14- Trends
Topic 15- Economy
Topic 16- Sports
Topic 18- Theatre and Art

Topic 19 - Housing and Design
Topic 21- Business
Topic 22 - China
Topic 23- Technology
Topic 24 - Opinion
Topic 25 - Books and Authors
Topic 26 - Spanish
Topic 27- Flight and Travel
Topic 28 - US Crime
Topic 29- Ebola

Miscellaneous

Topic 13 - Spam
Topic 17 - Decisions
Topic 20 - Blogs

Evaluation

40 articles from the categorization results for 5 topics were observed and the precision was noted

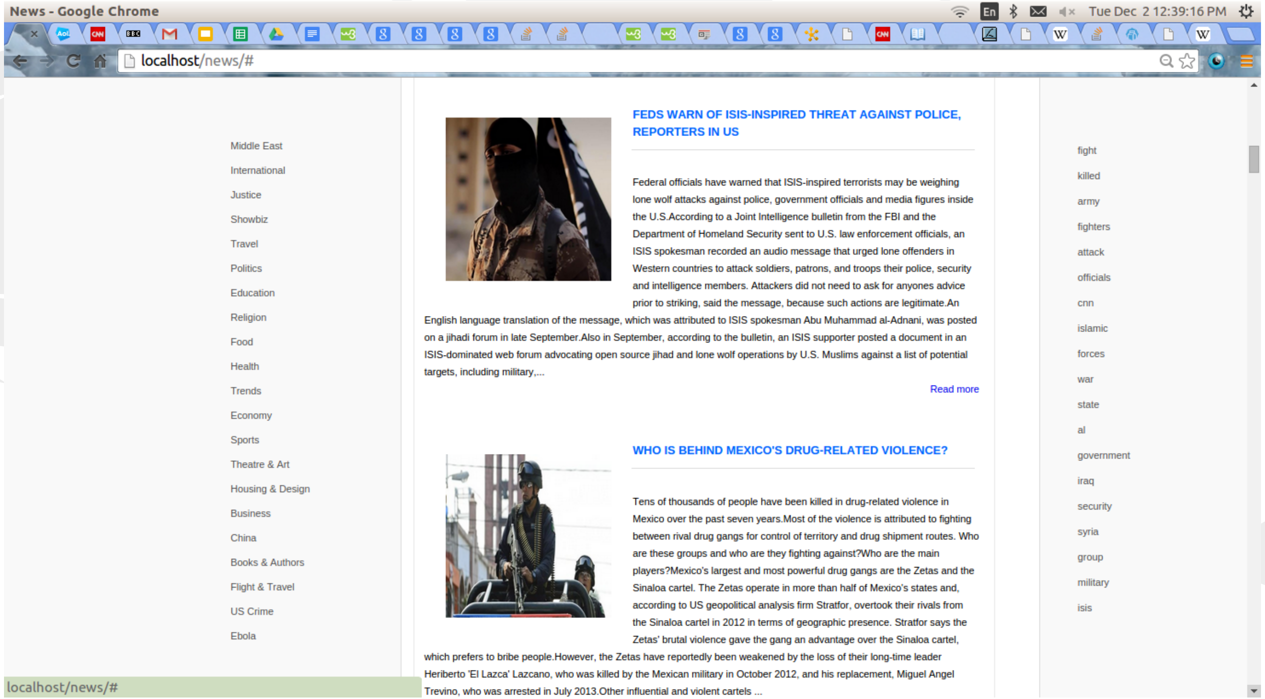
Evaluation (contd)

Topic 0

Middle East

30/40

Precision ~ 0.75



Middle East
International
Justice
Showbiz
Travel
Politics
Education
Religion
Food
Health
Trends
Economy
Sports
Theatre & Art
Housing & Design
Business
China
Books & Authors
Flight & Travel
US Crime
Ebola

FEDS WARN OF ISIS-INSPIRED THREAT AGAINST POLICE, REPORTERS IN US

Federal officials have warned that ISIS-inspired terrorists may be weighing lone wolf attacks against police, government officials and media figures inside the U.S.According to a Joint Intelligence bulletin from the FBI and the Department of Homeland Security sent to U.S. law enforcement officials, an ISIS spokesman recorded an audio message that urged lone offenders in Western countries to attack soldiers, patrons, and troops their police, security and intelligence members. Attackers did not need to ask for anyones advice prior to striking, said the message, because such actions are legitimate.An English language translation of the message, which was attributed to ISIS spokesman Abu Muhammad al-Adnani, was posted on a jihadi forum in late September.Also in September, according to the bulletin, an ISIS supporter posted a document in an ISIS-dominated web forum advocating open source jihad and lone wolf operations by U.S. Muslims against a list of potential targets, including military,...

Read more

WHO IS BEHIND MEXICO'S DRUG-RELATED VIOLENCE?

Tens of thousands of people have been killed in drug-related violence in Mexico over the past seven years.Most of the violence is attributed to fighting between rival drug gangs for control of territory and drug shipment routes. Who are these groups and who are they fighting against?Who are the main players?Mexico's largest and most powerful drug gangs are the Zetas and the Sinaloa cartel. The Zetas operate in more than half of Mexico's states and, according to US geopolitical analysis firm Stratfor, overtook their rivals from the Sinaloa cartel in 2012 in terms of geographic presence. Stratfor says the Zetas' brutal violence gave the gang an advantage over the Sinaloa cartel, which prefers to bribe people.However, the Zetas have reportedly been weakened by the loss of their long-time leader Heriberto 'El Lazca' Lazcano, who was killed by the Mexican military in October 2012, and his replacement, Miguel Angel Trevino, who was arrested in July 2013.Other influential and violent cartels ...

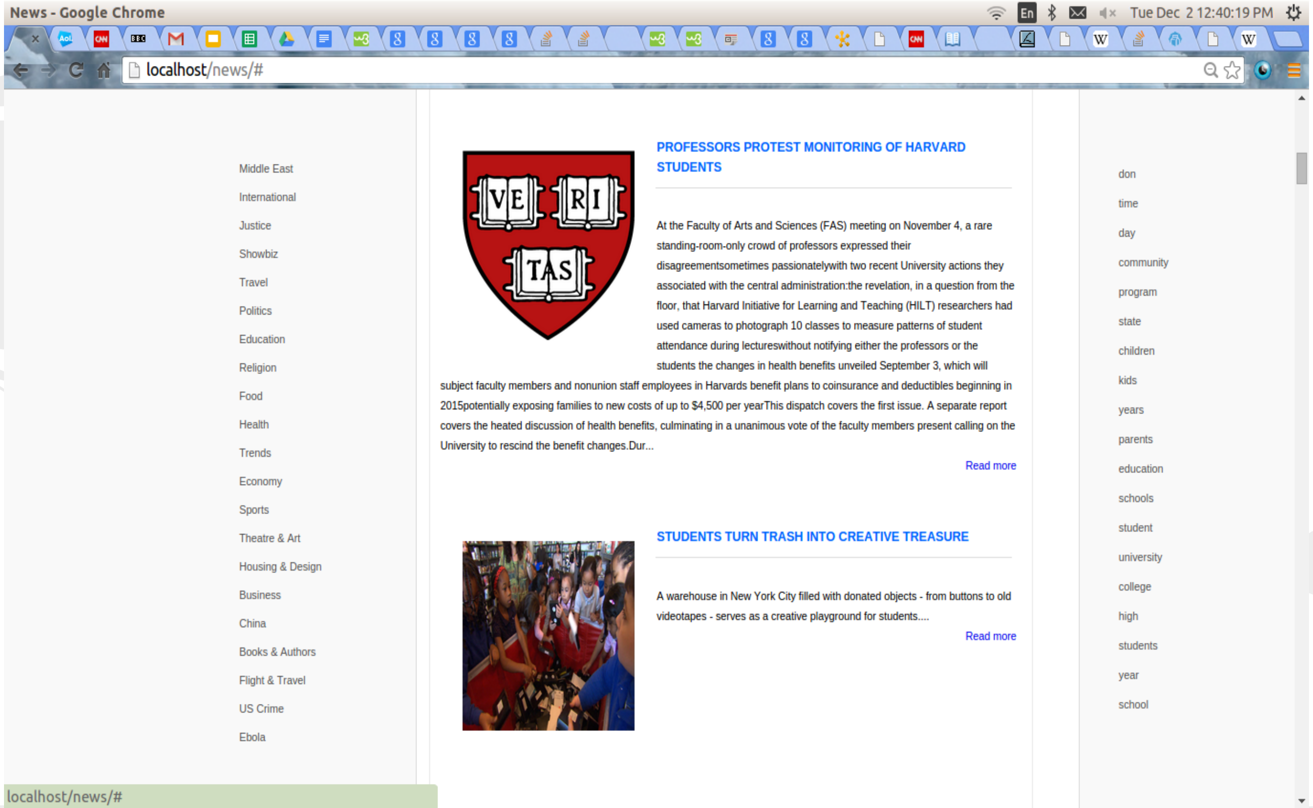
fight
killed
army
fighters
attack
officials
cnn
islamic
forces
war
state
al
government
iraq
security
syria
group
military
isis

Evaluation (contd)

Topic 8 Education

34/40

Precision ~ 0.86



PROFESSORS PROTEST MONITORING OF HARVARD STUDENTS

At the Faculty of Arts and Sciences (FAS) meeting on November 4, a rare standing-room-only crowd of professors expressed their disagreementsometimes passionatelywith two recent University actions they associated with the central administration:the revelation, in a question from the floor, that Harvard Initiative for Learning and Teaching (HILT) researchers had used cameras to photograph 10 classes to measure patterns of student attendance during lectureswithout notifying either the professors or the students the changes in health benefits unveiled September 3, which will subject faculty members and nonunion staff employees in Harvards benefit plans to coinsurance and deductibles beginning in 2015potentially exposing families to new costs of up to $4,500 per yearThis dispatch covers the first issue. A separate report covers the heated discussion of health benefits, culminating in a unanimous vote of the faculty members present calling on the University to rescind the benefit changes.Dur...

Read more

STUDENTS TURN TRASH INTO CREATIVE TREASURE

A warehouse in New York City filled with donated objects - from buttons to old videotapes - serves as a creative playground for students....

Read more

don
time
day
community
program
state
children
kids
years
parents
education
schools
student
university
college
high
students
year
school

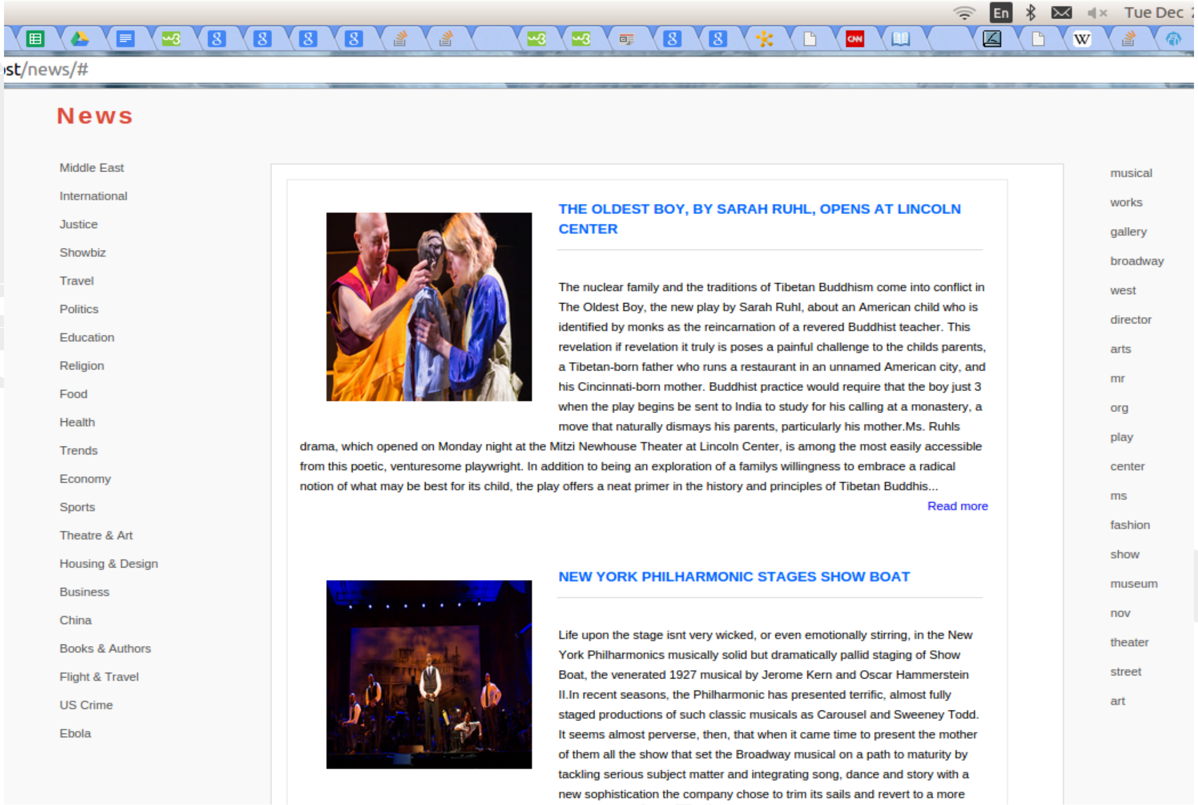
Evaluation (contd)

Topic 18

Theatre & Art

35/40

Precision ~ 0.88



THE OLDEST BOY, BY SARAH RUHL, OPENS AT LINCOLN CENTER

The nuclear family and the traditions of Tibetan Buddhism come into conflict in The Oldest Boy, the new play by Sarah Ruhl, about an American child who is identified by monks as the reincarnation of a revered Buddhist teacher. This revelation if revelation it truly is poses a painful challenge to the childs parents, a Tibetan-born father who runs a restaurant in an unnamed American city, and his Cincinnati-born mother. Buddhist practice would require that the boy just 3 when the play begins be sent to India to study for his calling at a monastery, a move that naturally dismays his parents, particularly his mother.Ms. Ruhls drama, which opened on Monday night at the Mitzi Newhouse Theater at Lincoln Center, is among the most easily accessible from this poetic, venturesome playwright. In addition to being an exploration of a familys willingness to embrace a radical notion of what may be best for its child, the play offers a neat primer in the history and principles of Tibetan Buddhis...

Read more

NEW YORK PHILHARMONIC STAGES SHOW BOAT

Life upon the stage isnt very wicked, or even emotionally stirring, in the New York Philharmonics musically solid but dramatically pallid staging of Show Boat, the venerated 1927 musical by Jerome Kern and Oscar Hammerstein II.In recent seasons, the Philharmonic has presented terrific, almost fully staged productions of such classic musicals as Carousel and Sweeney Todd. It seems almost perverse, then, that when it came time to present the mother of them all the show that set the Broadway musical on a path to maturity by tackling serious subject matter and integrating song, dance and story with a new sophistication the company chose to trim its sails and revert to a more

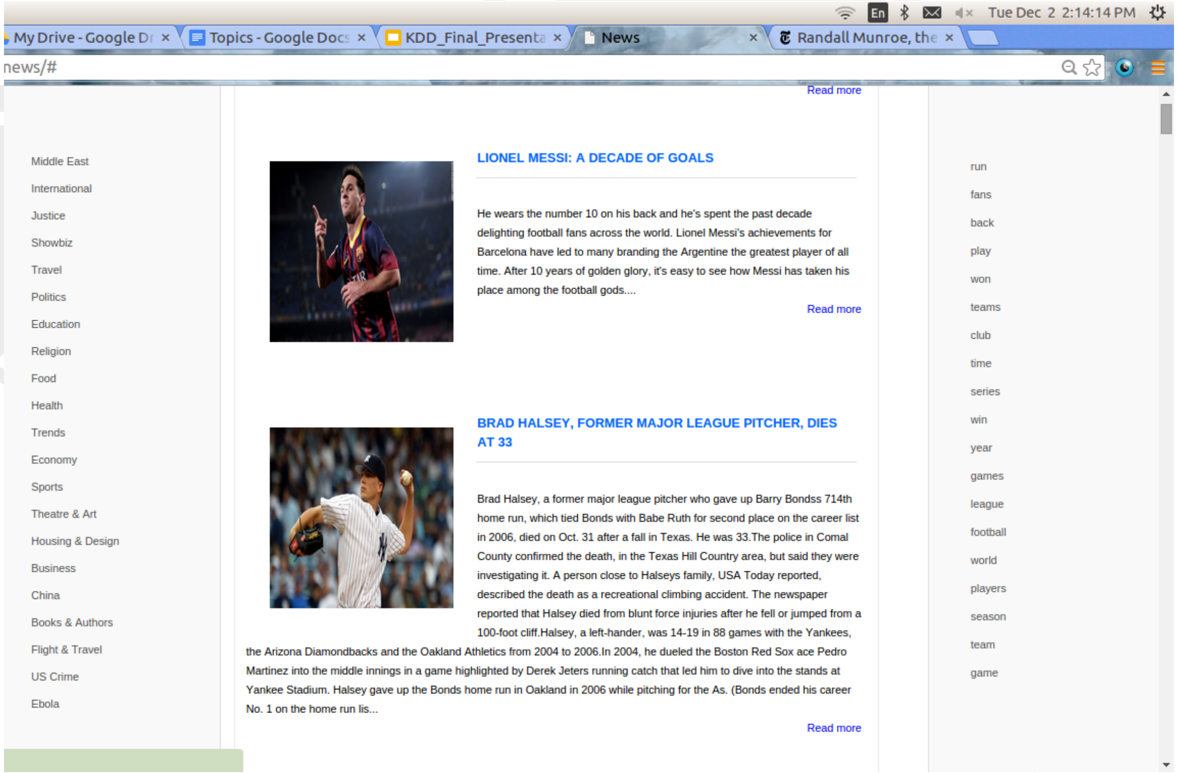
Evaluation (contd)

Topic 16

Sports

40/40

Precision ~ 1.00



LIONEL MESSI: A DECADE OF GOALS

He wears the number 10 on his back and he's spent the past decade delighting football fans across the world. Lionel Messi's achievements for Barcelona have led to many branding the Argentine the greatest player of all time. After 10 years of golden glory, it's easy to see how Messi has taken his place among the football gods....

Read more

BRAD HALSEY, FORMER MAJOR LEAGUE PITCHER, DIES AT 33

Brad Halsey, a former major league pitcher who gave up Barry Bondss 714th home run, which tied Bonds with Babe Ruth for second place on the career list in 2006, died on Oct. 31 after a fall in Texas. He was 33.The police in Comal County confirmed the death, in the Texas Hill Country area, but said they were investigating it. A person close to Halseys family, USA Today reported, described the death as a recreational climbing accident. The newspaper reported that Halsey died from blunt force injuries after he fell or jumped from a 100-foot cliff.Halsey, a left-hander, was 14-19 in 88 games with the Yankees, the Arizona Diamondbacks and the Oakland Athletics from 2004 to 2006.In 2004, he dueled the Boston Red Sox ace Pedro Martinez into the middle innings in a game highlighted by Derek Jeters running catch that led him to dive into the stands at Yankee Stadium. Halsey gave up the Bonds home run in Oakland in 2006 while pitching for the As. (Bonds ended his career No. 1 on the home run lis...

Read more

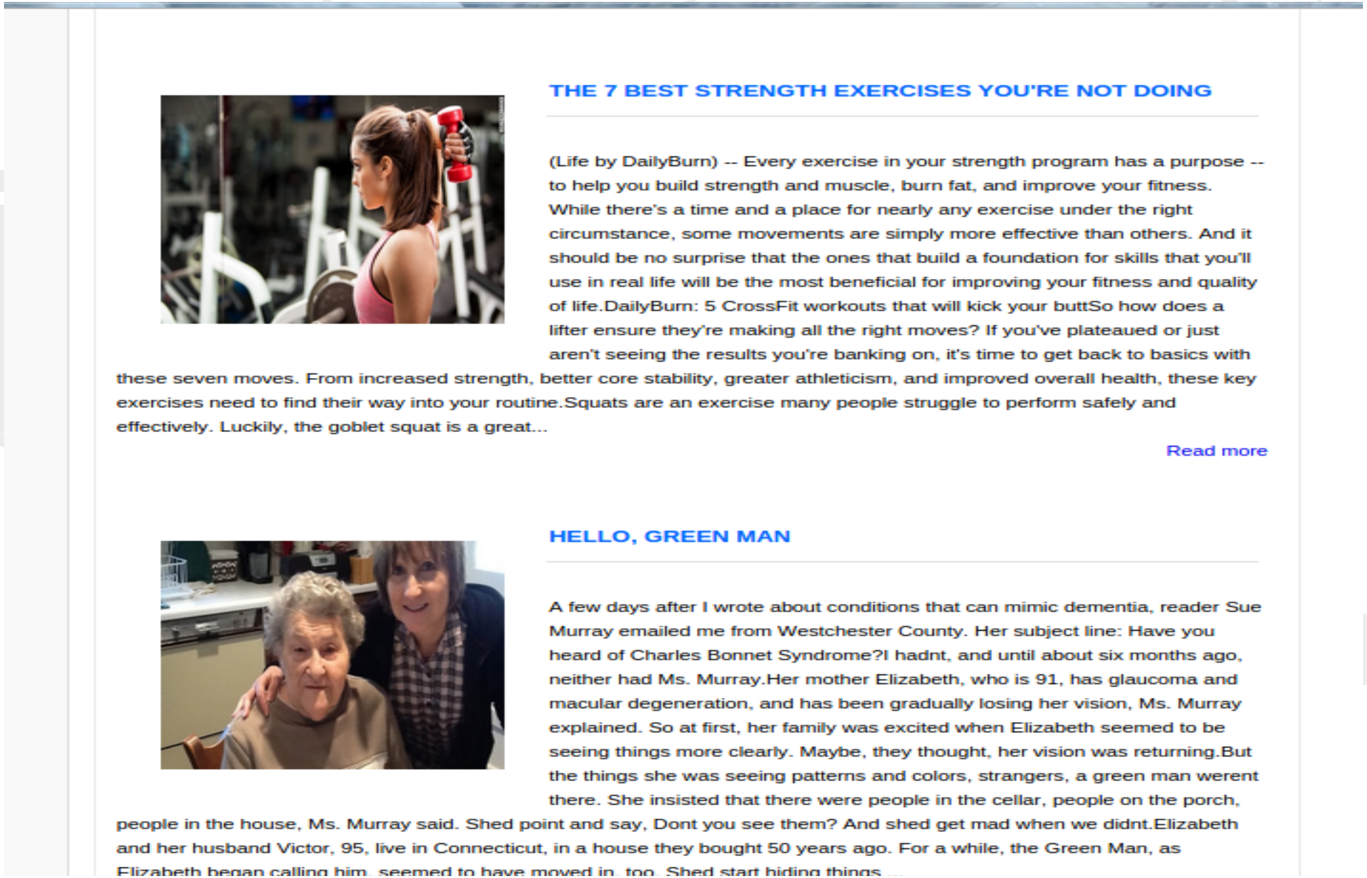
Evaluation (contd)

Topic 12

Health

39/40

Precision ~ 0.98



THE 7 BEST STRENGTH EXERCISES YOU'RE NOT DOING

(Life by DailyBurn) -- Every exercise in your strength program has a purpose -- to help you build strength and muscle, burn fat, and improve your fitness. While there's a time and a place for nearly any exercise under the right circumstance, some movements are simply more effective than others. And it should be no surprise that the ones that build a foundation for skills that you'll use in real life will be the most beneficial for improving your fitness and quality of life.DailyBurn: 5 CrossFit workouts that will kick your buttSo how does a lifter ensure they're making all the right moves? If you've plateaued or just aren't seeing the results you're banking on, it's time to get back to basics with these seven moves. From increased strength, better core stability, greater athleticism, and improved overall health, these key exercises need to find their way into your routine.Squats are an exercise many people struggle to perform safely and effectively. Luckily, the goblet squat is a great...

Read more

HELLO, GREEN MAN

A few days after I wrote about conditions that can mimic dementia, reader Sue Murray emailed me from Westchester County. Her subject line: Have you heard of Charles Bonnet Syndrome?I hadnt, and until about six months ago, neither had Ms. Murray.Her mother Elizabeth, who is 91, has glaucoma and macular degeneration, and has been gradually losing her vision, Ms. Murray explained. So at first, her family was excited when Elizabeth seemed to be seeing things more clearly. Maybe, they thought, her vision was returning.But the things she was seeing patterns and colors, strangers, a green man werent there. She insisted that there were people in the cellar, people on the porch, people in the house, Ms. Murray said. Shed point and say, Dont you see them? And shed get mad when we didnt.Elizabeth and her husband Victor, 95, live in Connecticut, in a house they bought 50 years ago. For a while, the Green Man, as Elizabeth began calling him, seemed to have moved in, too. Shed start hiding things ...

Future Work

- Developing a workflow for
  - Assignment of new articles to topics
  - Reindexing articles
- Exploring relationships between topics
- Live data

Bibliography

- D. Blei , A. Ng, M. Jordan, **Latent dirichlet allocation** , *Journal of machine Learning research* , 2003
- H. Wallach, **Topic modeling: beyond bag-of-words** , *International conference on Machine learning* , 2006
- Hayes, P., Knecht, L., & Cellio, M. (1988). **A news story categorization system**. *Proceedings of the Second Conference on Applied Natural Language Processing (ANLC ‘88), 9-17.* http://dx.doi.org/10.3115/974235.974238
- OuYang, L. (n.d.). Newspaper: Article scraping & curation. Retrieved from Lucasao website: http://newspaper.readthedocs.org/en/latest/

Thank you !