Detecting Topics and their Transitions

Victor Mireles, Artem Revenko

SEMANTIC WEB COMPANY
linking data to knowledge

Hybrid Statistical Semantic Understanding and Emerging Semantics Workshop
@ ISWC 2017, Vienna

Introduction

Summary (Humanities)

Can we study quantitatively how discourse is changing with time?

Summary (Mathematics)

If documents are vectors, how does the basis that best describes them evolve?

Summary (Semantic Web)

What sections of the knowledge base are used together, and how do these change?

Contents

Representing Documents

Topic Modeling

Topic Transitions

Experiments and results

Representing Documents

Beyond text

A document

There's global concern about a major nuclear accident in Japan, which could turn a very bad situation into a terrible one, said Shane Oliver , head of investment strategy in Sydney at AMP Capital Investors Ltd., which manages about $98 billion. Paladin Energy Ltd. (PDN) and other Australian uranium producers and explorers slumped for a second day this week on concern demand for nuclear energy will decline.

Beyond text

A document

There's global concern about a major nuclear accident in Japan, which could turn a very bad situation into a terrible one, said Shane Oliver , head of investment strategy in Sydney at AMP Capital Investors Ltd., which manages about $98 billion. Paladin Energy Ltd. (PDN) and other Australian uranium producers and explorers slumped for a second day this week on concern demand for nuclear energy will decline.

Concepts extracted from that document

Accident | Japan | Investment | Strategy | Capital

Energy | Austrialians | Austrialian | Urianium

Demand | Nuclear Energy

Beyond text

A document

There's global concern about a major nuclear accident in Japan, which could turn a very bad situation into a terrible one, said Shane Oliver , head of investment strategy in Sydney at AMP Capital Investors Ltd., which manages about $98 billion. Paladin Energy Ltd. (PDN) and other Australian uranium producers and explorers slumped for a second day this week on concern demand for nuclear energy will decline.

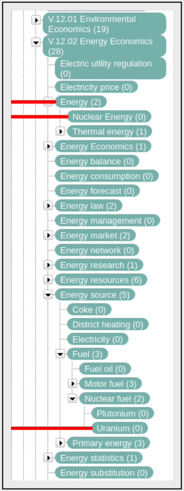
Concepts extracted from that document

Accident | Japan | Investment | Strategy | Capital

Energy | Austrialians | Austrialian | Urianium

Demand | Nuclear Energy

Relationships



- V.12.01 Environmental Economics (19)
- V.12.02 Energy Economics (28)
  - Electric utility regulation (0)
  - Electricity price (0)
  - Energy (2)
    - Nuclear Energy (0)
    - Thermal energy (1)
  - Energy Economics (1)
  - Energy balance (0)
  - Energy consumption (0)
  - Energy forecast (0)
  - Energy law (2)
  - Energy management (0)
  - Energy market (2)
  - Energy network (0)
  - Energy research (1)
  - Energy resources (6)
  - Energy source (5)
    - Coke (0)
    - District heating (0)
    - Electricity (0)
    - Fuel (3)
      - Fuel oil (0)
      - Motor fuel (3)
      - Nuclear fuel (2)
        - Plutonium (0)
        - Uranium (0)
    - Primary energy (3)
  - Energy statistics (1)
  - Energy substitution (0)

Beyond text

Concepts extracted from a document

Accident Japan

Investment Strategy

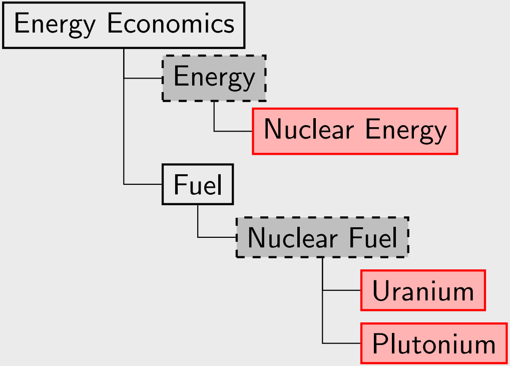
Capital Energy

Austrialians Austrialian

Urianium Demand

Nuclear Energy

A Thesaurus



Energy Economics

Energy

Nuclear Energy

Fuel

Nuclear Fuel

Uranium

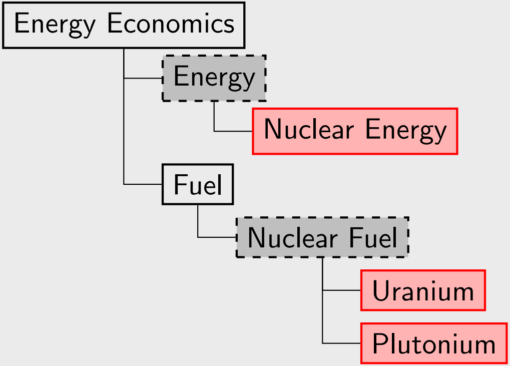
Plutonium

Beyond text

Concepts extracted from a document

Accident Japan
Investment Strategy
Capital Energy
Austrialians Austrialian
Urianium Demand
Nuclear Energy

A Thesaurus



Energy Economics
Energy
Nuclear Energy
Fuel
Nuclear Fuel
Uranium
Plutonium

Leaves

Nuclear Energy	1
Urianium	1
Plutonium	0

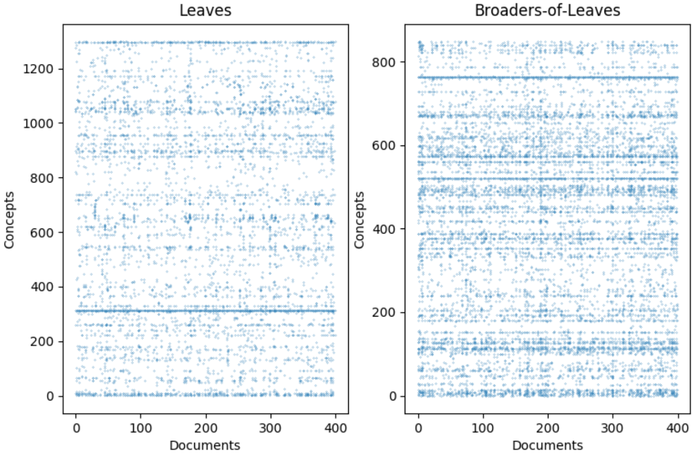
Broaders-of-Leaves

Nuclear Fuel	1
Energy	2

	doc 1
concept 1	0
concept 2	1
concept 3	3
concept 4	1
concept 5	0
concept 6	0
concept 7	0
concept 8	0
concept 9	0
concept 10	2
⋮	
concept N	0

	doc 1	doc 2
concept 1	0	0
concept 2	1	1
concept 3	3	0
concept 4	1	0
concept 5	0	2
concept 6	0	0
concept 7	0	3
concept 8	0	0
concept 9	0	0
concept 10	2	1
$\vdots$	$\vdots$	$\vdots$
concept N	0	4

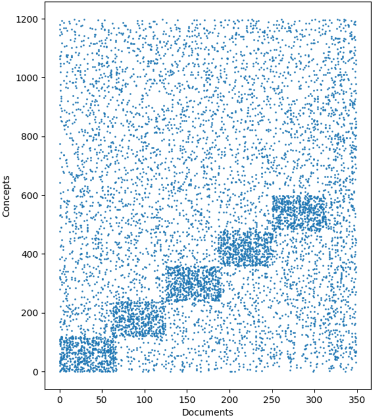
Document-Concept Matrices



Topic Modeling

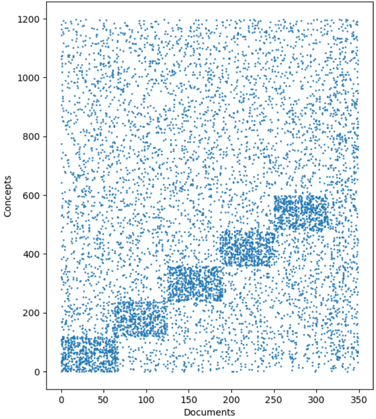
Document-Concept Matrices

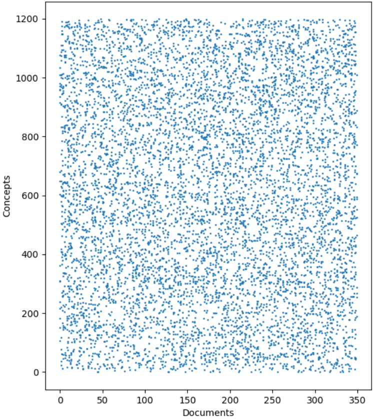
All seems very messy....



Document-Concept Matrices

All seems very messy....





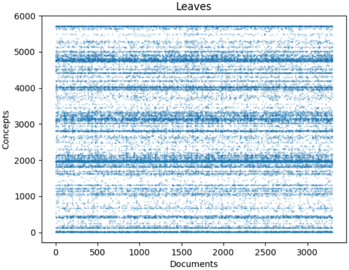
NMF

The setup

- **Given**: A matrix A that encodes documents as vectors of concepts.
- **Output**: Two matrices, B and C such that $A \approx BC$
- B and C are non-negative
- B is the concepts to topics matrix
- C is the topics to documents matrix

Example

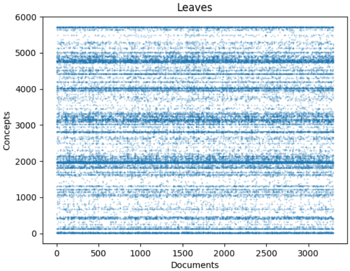
INPUT:



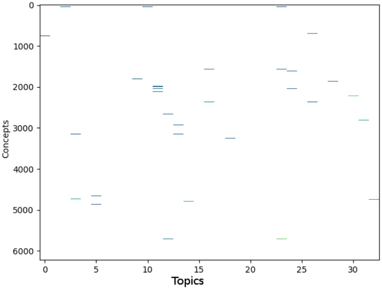
OUTPUT:

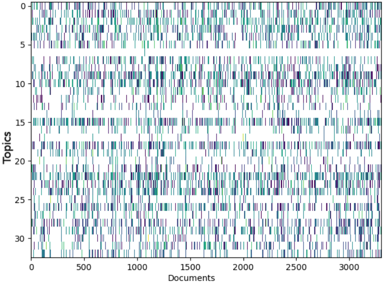
Example

INPUT:



OUTPUT:





Topic Transitions

Topic Transitions

The setup

- Two matrices A_1 and A_2 that encode documents as vectors of concepts.
- Two NMF derived matrices B_1 and B_2 that encode topics as vectors of concepts

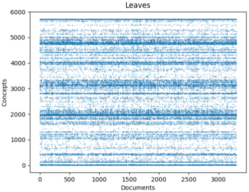
An optimization problem

Find T such that:

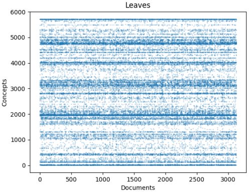
- $B_2 \approx TB_1$
- all entries of T are between 0 and 1

Example

Before

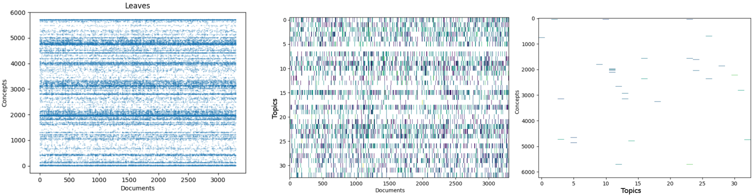


After

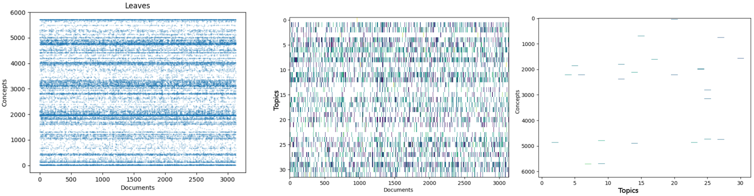


Example

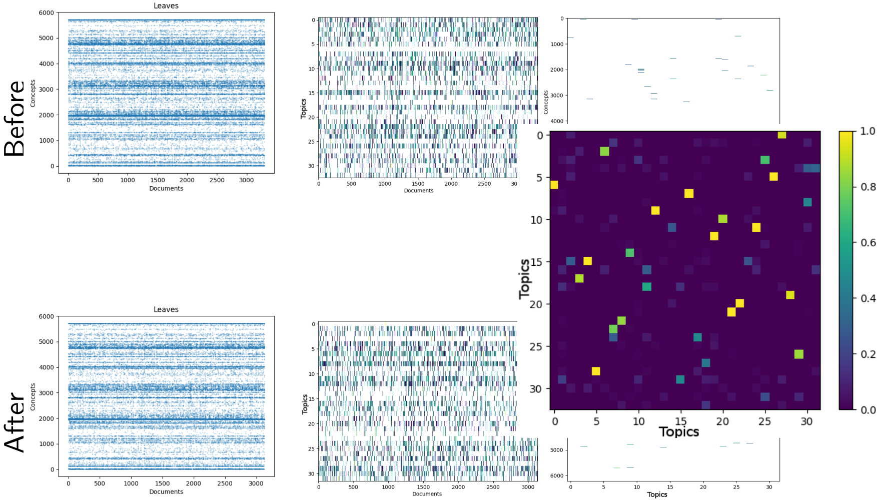
Before



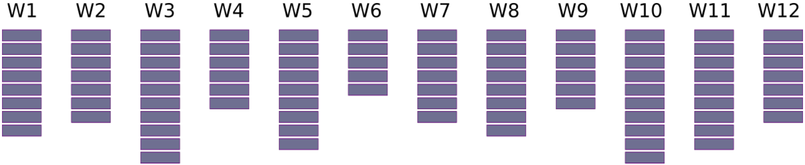
After



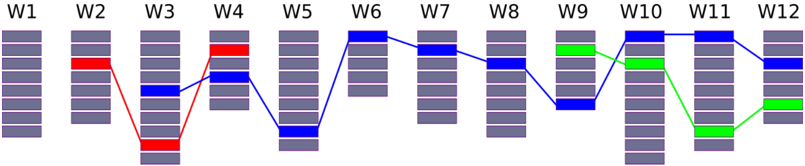
Example


Before
After
Leaves
Concepts
Documents
0
500
1000
1500
2000
2500
3000
4000
5000
6000
Topics
0
5
10
15
20
25
30
0.0
0.2
0.4
0.6
0.8
1.0

Flash and Stable topics



Flash and Stable topics

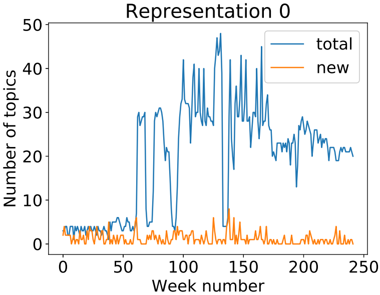


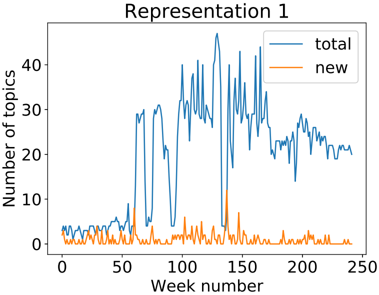
Experiments and results

Setup

1. 400000 Documents
2. Grouped in weeks between 2009 and 2013
3. Between 50 and 4900 documents per week, mean 2580
4. STW, Standard Thesaurus for Economics: 6221 concepts, 4108 Leaves

Topics detected



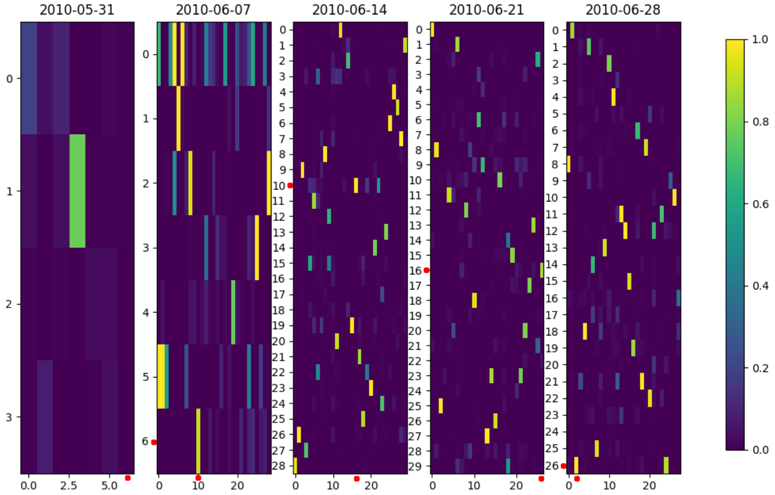


Some Flash Topics

2010 Football World Cup

2010-06-07	2010-06-14	2010-06-21	2010-06-28	2010-07-05
World	World	World	World	World
Sport event	Sport event	Matching	Sport event	Sport event
Football	Football	Sport event	Football	Football
Matching	Matching	Football	Coaching	Spain
South Africa	North Korea	South Africa	Brazil	Netherlands
Slovenia	Brazil	France	South Africa	South Africa
Italy	South Africa	Coaching	Spain	Dutch
Algeria	Coaching	Mexico	Matching	Matching
Australia	Korean	French	Netherlands	Spanish
Paraguay	Koreans	Brazil	Argentina	African
Netherlands	South Korea	American	Ghana	European
Ghana	Portugal	Argentina	Uruguay	Uruguay
Nation	Argentina	North Korea	Portugal	Coaching

Some Flash Topics



Some Flash Topics

Korean Peninsula Artillery Incident

2010-11-15 to 2010-12-13

Koreans | Korean | South Korea | North Korea | South Korean
South Koreans | Officials | Nation | Island | Foreign | World
Chinese | Fire | Government department | International
Sport event | American | India | Department | Brazil | Warship
Football | Australia | News agency | Asian | Office | Peace
Qatar | Beef | Export | Future | Russian | New Zealand
Japanese | Import | Wheat | Mexico | Matching | Climate change
Brazilians | Brazilian | Soldiers | White people | E-mail | Police
Process | Plants | Authority | Western

Some Flash Topics

2011 Drought

2011-02-07 to 2011-03-07

- Wheat
- Crops
- Drought
- Soybean
- World
- Rice
- Food price
- Nation
- Egypt
- Department
- Australia
- International
- Province
- Price
- Western
- Flood
- Sugar
- Future
- Export
- Palms
- Purchase
- Cotton
- Bangladesh
- France
- French
- Palm oil
- Russian
- Renminbi
- Median
- Plants
- Sport event
- Irrigation
- Chinese
- Confidence
- West Asia
- Wheat price

Some Flash Topics

Arab Spring

2011-02-14 to 2011-04-04

- Libya
- International
- Nation
- Foreign
- Arabs
- Arab
- Egypt
- Air
- West Asia
- Tunisia
- Officials
- African
- Industrial action
- Yemen
- Saudi Arabia
- Bahrain
- Humans
- Human rights
- British
- Sanction
- Western
- Head of government
- Fire
- France
- Italy
- French
- Syria
- Civil war
- Qatar
- Iraq
- American
- Export
- Authority
- Government department
- Spa
- Oman
- Air force
- Refugees
- Kuwait
- White people
- European
- Iran
- Venezuela
- World
- Islamic
- London
- River
- Police
- Department
- Society
- Licence
- E-mail
- Oil price
- Intelligence
- Pump
- Newspaper
- Italians
- Italian
- Jordan
- Geneva
- Office
- Malta
- Russian
- Algeria
- Terrorism
- Occupation
- Ferry shipping
- Turkish
- Turkey
- Desert
- Airline
- Petroleum resources

Some Flash Topics

Fukushima Daiichi Nuclear Accident

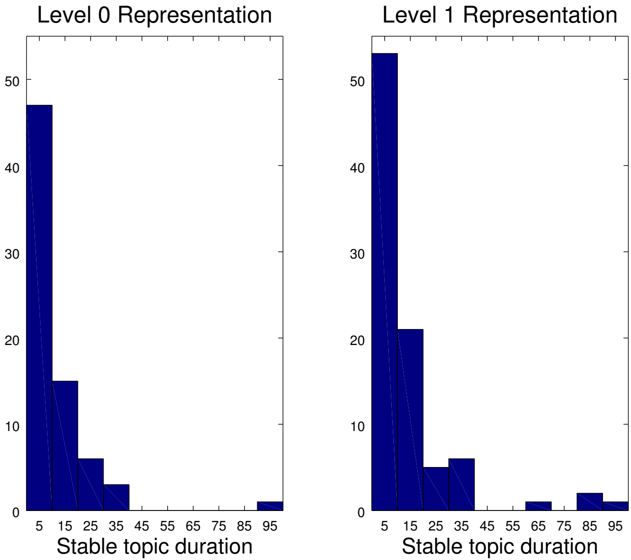
2011-03-07 to 2011-04-11

Plants | Nuclear energy | Manufacturing plant | Electricity | Cooling | Earthquake | Greenhouse gas emissions | Nuclear power plant | Nuclear fuel | Order | Process | Officials | Fire | Japanese | Health | Product | Government department | Nuclear safety | Pump | Core | Engineers | Seed | France | Permit | Electronics | Authors | Taiwan | Light | Iraq | Air | Germans | German | River | Authority | Island | Province | Humans | Laboratory

Some Stable Topics

"Stocks"	"Futures"	"Japan"	"Euro"	"Meat"
Market	Future	Yen	Germans	Beef
Stock market	Soybean	Tokyo	German	Light
Product	Crops	Japanese	European	Bayesian inference
International	Gold	Loss	Greek	Cattle
Purchase	Wheat	Electronics	Greeks	Price
Market value	Department	Newspaper	Greece	Department
Price	Sugar	Sales	Berlin	Plants
Swap	Rubber	Dividend	Nation	Flavour
Benchmarking	Singapore	Services	Bailout	Sales
Hedging	Cocoa	Product	Economy	Product
Loss	Cattle	Semiconductor	Portuguese	Import
Future	Palms	Plants	London	Tokyo

Stable Topics



Thank you!

Questions?

Contact: victor.mireles-chavez@semantic-web.com

SEMANTIC WEB COMPANY
linking data to knowledge

Vistit us: https://research.semantic-web.at