
# Time Series Analysis of Textual Data

Dipartimento di Ingegneria dell'Informazione
Università degli Studi di Siena
Dottorato di Ricerca in Ingegneria dell'Informazione
XVIII ciclo

Candidate:
Marco Turchi

Advisor:
Prof. Marco Maggini

# Analysis of patterns using textual information and temporal relations

Avalanche of data are available on the Web and huge amount of techniques have been applied on static data.

This work was born from the following questions:

- What can we do introducing the variable "**time**" in the analysis of these data?
- Can different patterns be extracted?
- Can more information be retrieved?

# Analysis of patterns using textual information and temporal relations

We study two different interactions of these factors:

1. Given a sequence of written documents in different languages, we want to find temporal relations among languages.

2. Given a sequence of documents in time, we want to find patterns which merge textual information and temporal relations.

# Outlines Language Evolution

- Neutral Theory of Evolution.
- Language Evolution as a Random Walk.
- Statistical Language Signature.
- Results.
- Conclusions.

Italian, English, Latvian and Spanish

Which are the most similar languages?

Are we able to compare languages?

- If we could compare languages, we would:
  - compute the distance among languages
  - construct phylogenetic tree of languages
  - analyze the drift of languages from a common ancestor

# The answer is … YES

1. We develop a ”statistical signature” of a language (SLS) .
2. We show its stability within languages and its discriminative power among languages.
3. We study the language evolution:
   - we reconstruct a phylogenetic tree of IndoEuropean (IE) languages

# *Neutral Theory of Evolution*

- Kimura (1970): biological evolution as a random walk in sequence space.
- Most mutations are selectively neutral, yet they reach fixation (due to chance).
- They can be used as markers to reconstruct phylogenetic relations.

- Genetic drift leads to differentiation

# *Language Evolution as a Random Walk*

- Neutral mutations are accumulated.
- Some mutations can become fixed in the population, over time
  - language mutations in one speaker: can be lost
  - language mutations in most speakers: can be fixed.
- A random walk on languages can be created using these mutations.

# *Statistical Language Signature*

- We are looking for features that evidence those mutations.
- We observe that the frequency with which a language uses n-grams (sequence of n adjacent letters) is a highly conserved feature of the language
  - all documents in a language have similar statistical signature
  - all languages have their characteristic SLS, and that they can be reliably identified by it.

# *Statistical Language Signature*

- For n=2, the SLS is a matrix 27 x 27, and each entry is represented by:
  - di-gram Frequency.
  - Odds-ratio.
- When n > 2, we are not able to represent the SLS, so we introduce two different kernels:
  - p-Spectrum kernel.
  - Mismatch kernel.

# *Statistical Language Signature*

- di-gram Frequency
  - Frobenius Distance:

$$D_F(X,Y) = \left\| X - Y \right\|_F =$$

$$\left\langle \left(x_{i,j} - y_{i,j}\right), \left(x_{i,j} - y_{i,j}\right) \right\rangle = \sqrt{\sum_{i=1}^{27}\sum_{j=1}^{27} \left| x_{i,j} - y_{i,j} \right|^2}$$

- Odds Ratio
  - Karlin (1-norm) Distance:

$$D_{L1}(X,Y) = \frac{1}{(27)^2} \sum_{i=1}^{27}\sum_{J=1}^{27} \left| x_{i,j} - y_{i,j} \right|$$

# *Statistical Language Signature*

- Kernel string:
  - Distance:

$$D(X,Y)^2 = k(X,X) + k(Y,Y) - 2k(X,Y)$$

  - Distance Matrix:

    The distance matrix KD is the matrices whose entries represent the distance between all pairs of sequences. Each sequence is a string that represents a language.

# *Statistical Language Signature*

Test the stability of SLS:

- We use 50 **written** documents, 10 each for English, German, Spanish, Italian and French.

- We compute the average pairwise distance for documents in the same language (Intra*A*) and for documents in different languages (Extra*A*).

# *Statistical Language Signature*

Test the stability of SLS:

- We compare their ratio with the same quantity measured for randomly create sets of 10 documents.

$$r = \frac{IntraA}{ExtraA}$$

- We repeat this 10,000 times.
- Each time the resulting ratio is larger: with p-value < 0.0001.

# *Statistical Language Signature*

These features:

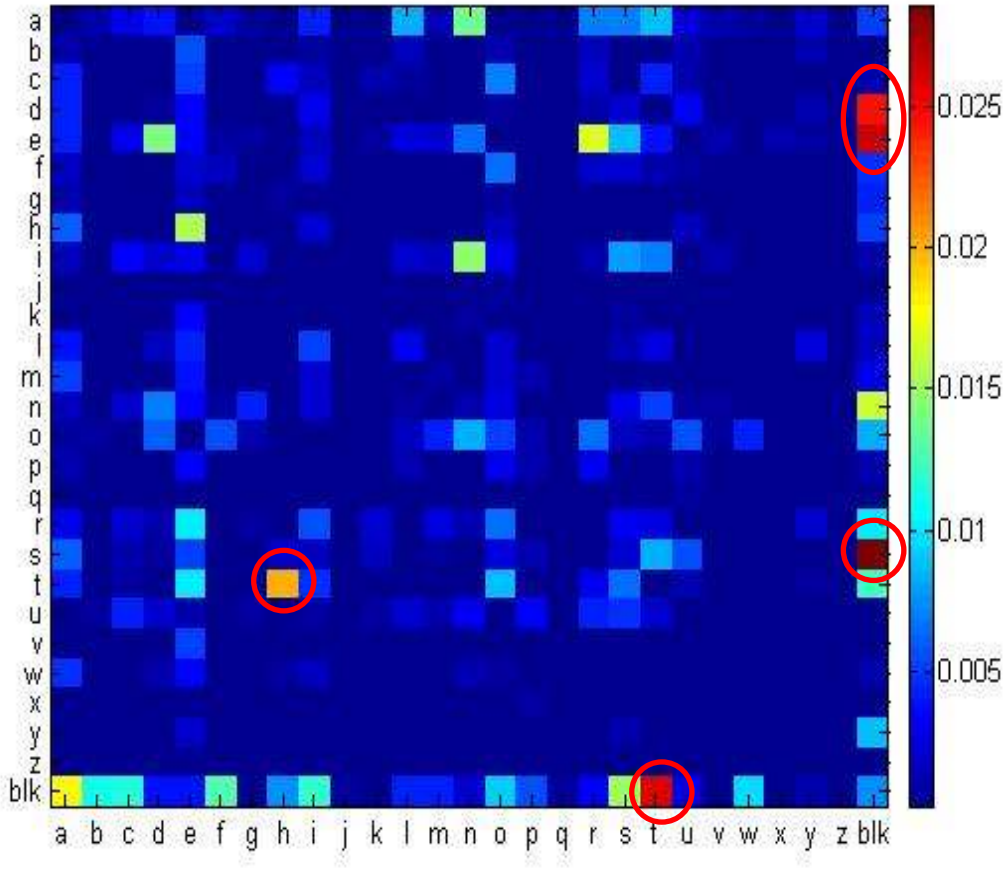
1. are stable
2. are properties of the language
3. are not properties of the given document
4. they do not depend on topics and authors.

# *Statistical Language Signature*

## All languages have their characteristic SLS

**Stores face fines in toxic scandal**

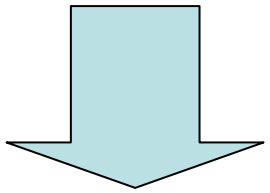
FSA officials are furious that the additive has been allowed to enter the food chain. They believe food companies and supermarkets should have conducted more rigorous tests after an alert over the dye in 2003. Sources at the FSA confirmed yesterday that charges could be brought against companies and their directors for "selling food injurious to health". Under food safety laws, the companies involved could face unlimited fines.
One said the organisation had been warned by its experts more than a year ago to step up its policing of Sudan 1 but had failed to do so.
Chris Grayling, Conservative health spokesman, said: "I am genuinely quite worried that the FSA seems to have acted very slowly.. . .
(The Sunday Times – Britain)



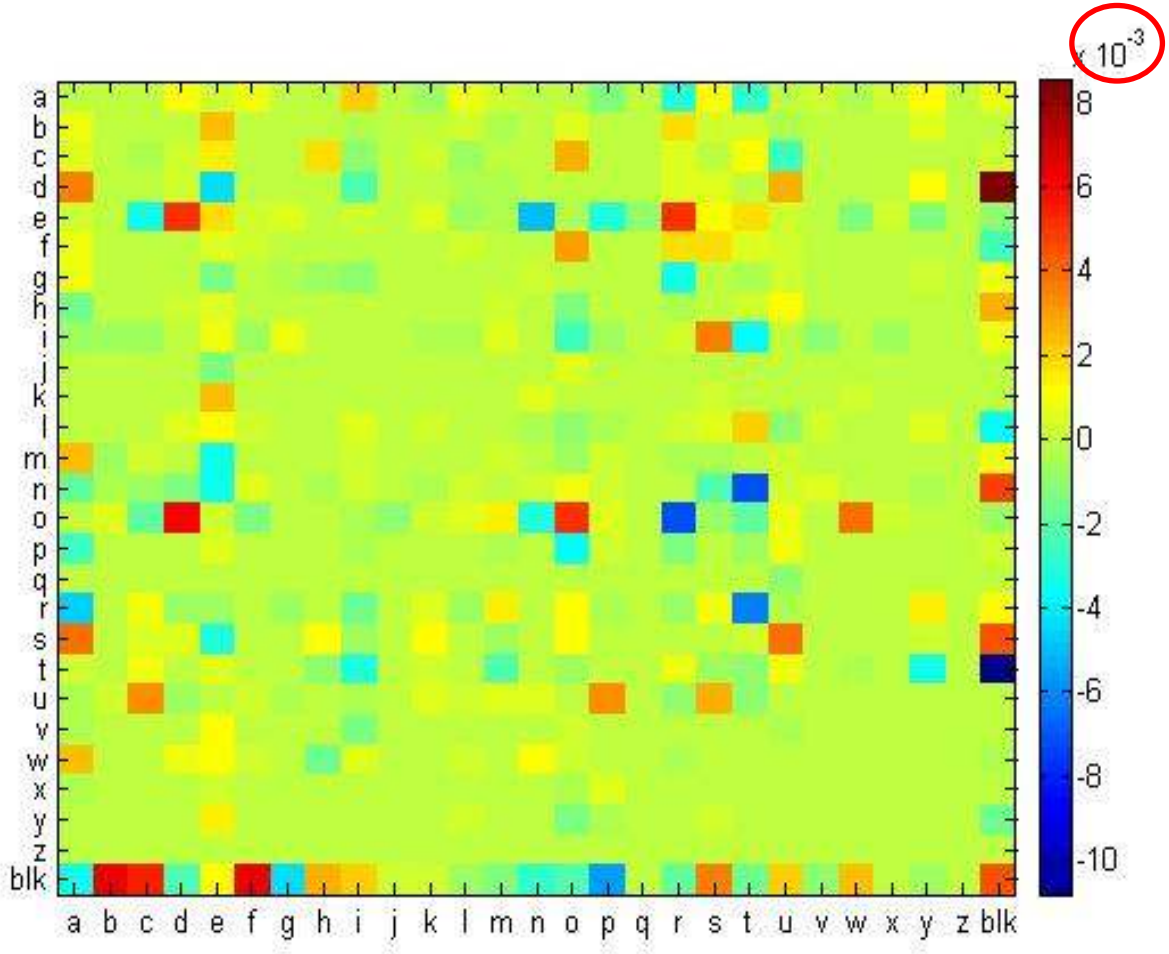
# *Statistical Language Signature*

all documents in a language have similar statistical signature

Difference matrix of two English documents.



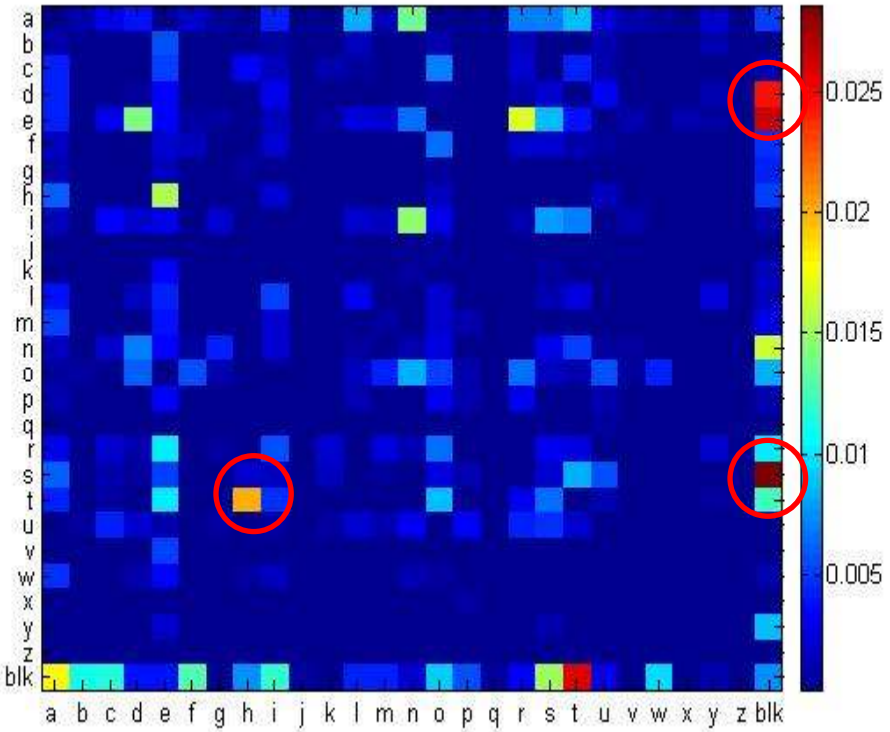
Each entry has low value



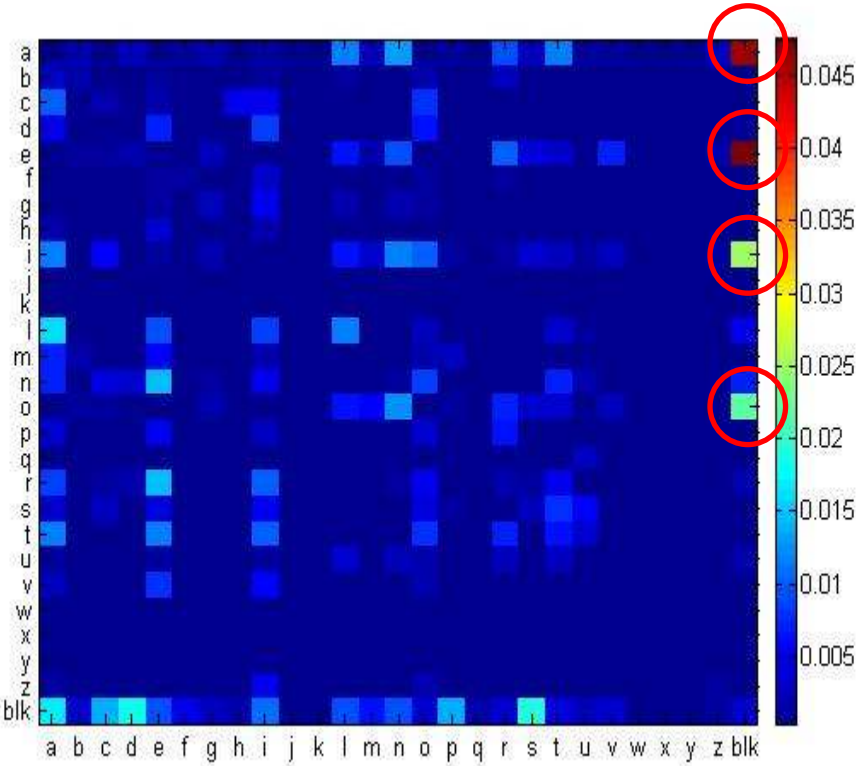
# *Statistical Language Signature*

Documents in different languages have different SLS

English

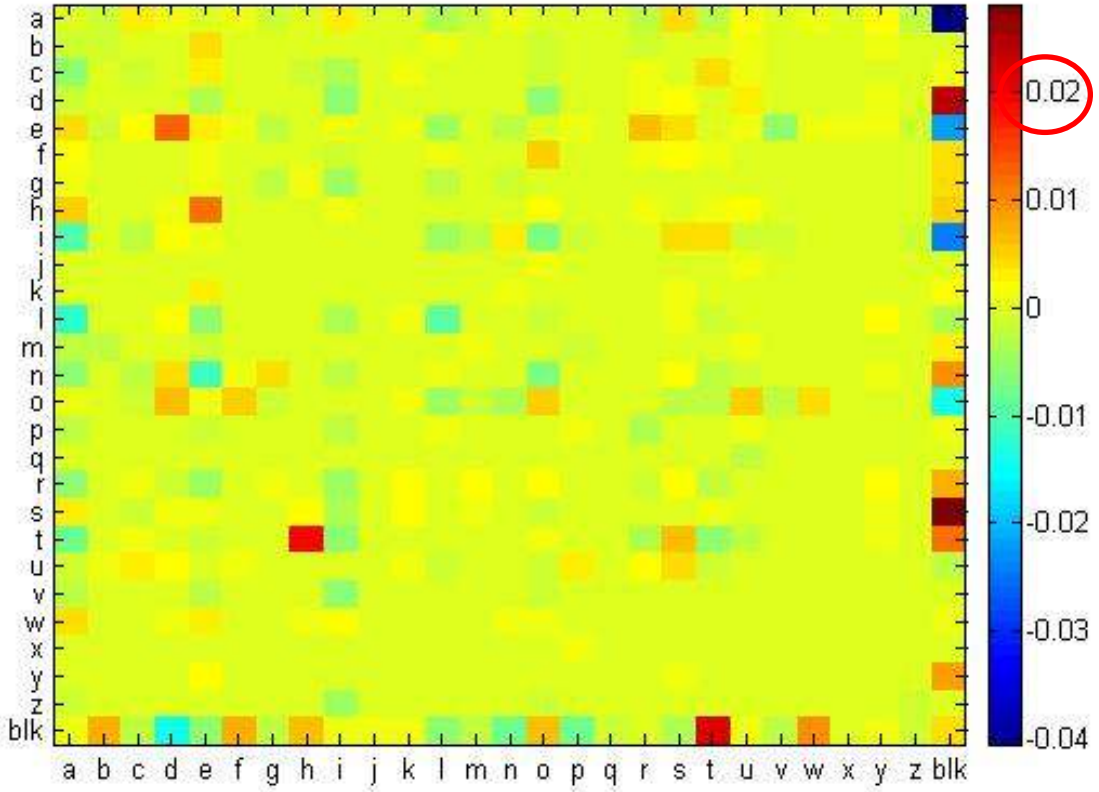



Italian



# *Statistical Language Signature*

Documents in different languages have different SLS



# Dataset

- We use two datasets.
  - First is made of 50 **written** documents: 10 each for English, German, Spanish, Italian and French and each document is a news story.
  - Second is made of 42 written documents: each document is the translation of the "Universal Declaration of Human Rights" in a different language.
- For each document:
  - We substitute:
    - the upper case letters with lower case one
    - we remove diacritical marks
  - We remove each character that does not belong to our set (all the 26 letters and the space)

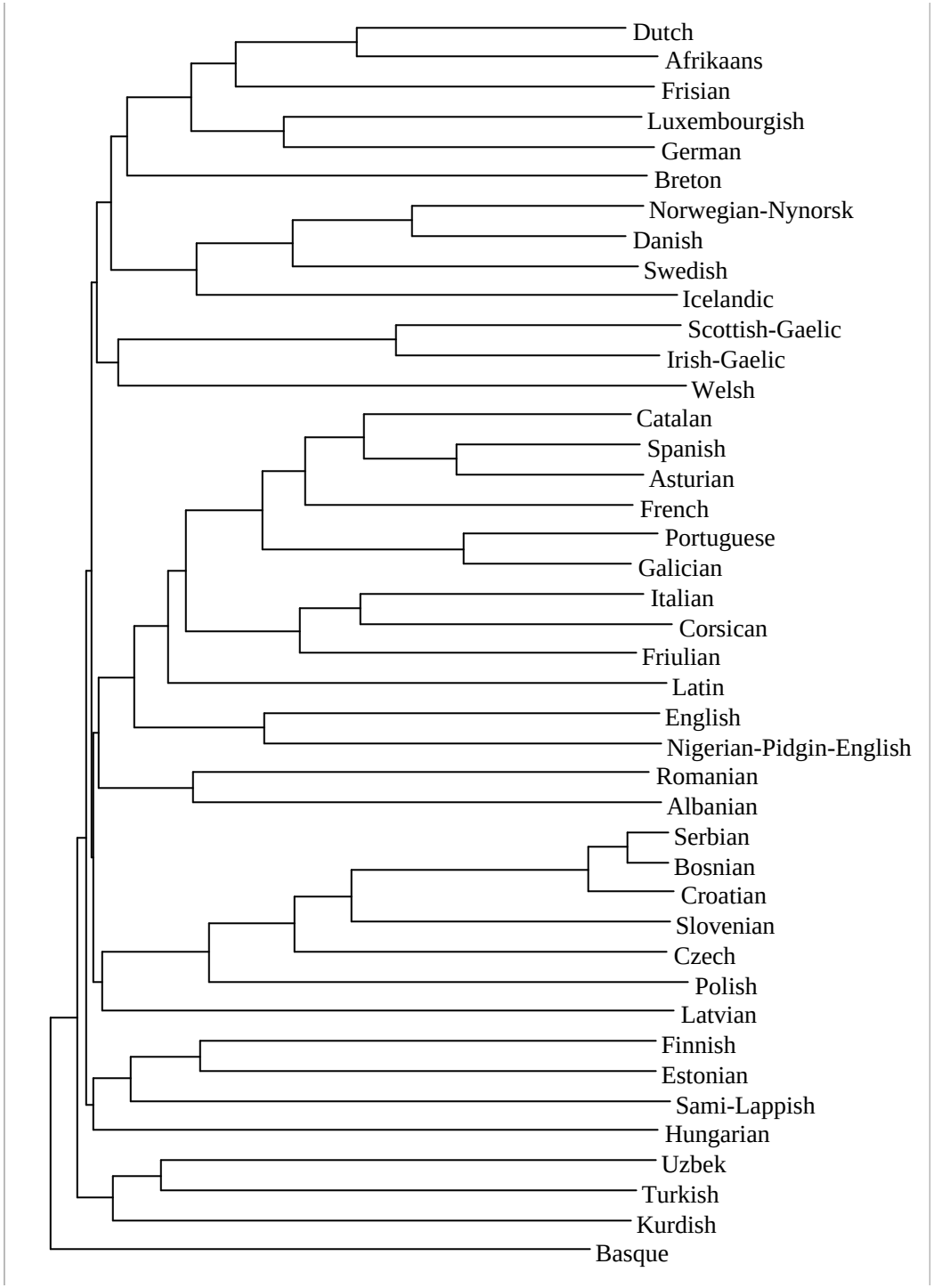
# Conversion Table

| *Stressed Letters* | *Substituted Letter* |
|---|---|
| ć, č, ç | c |
| š, ś | s |
| ž, ż, ź | z |
| d, ð | d |
| à, ă, â, å, á, ã, ä, ą | a |
| ý | y |
| ù, ú, û, u̇, ű, ü | u |
| í, î, ì, ï, i | i |
| è, é, ê, ë, ę, ě | e |
| ř | r |
| ô, ó, ò, õ, ő, ö, ø | o |
| ł | l |
| ń, ñ, ň | n |
| æ | ae |
| $\theta$ | t |
| ġ | g |
| ß | ss |
| Þ | th |

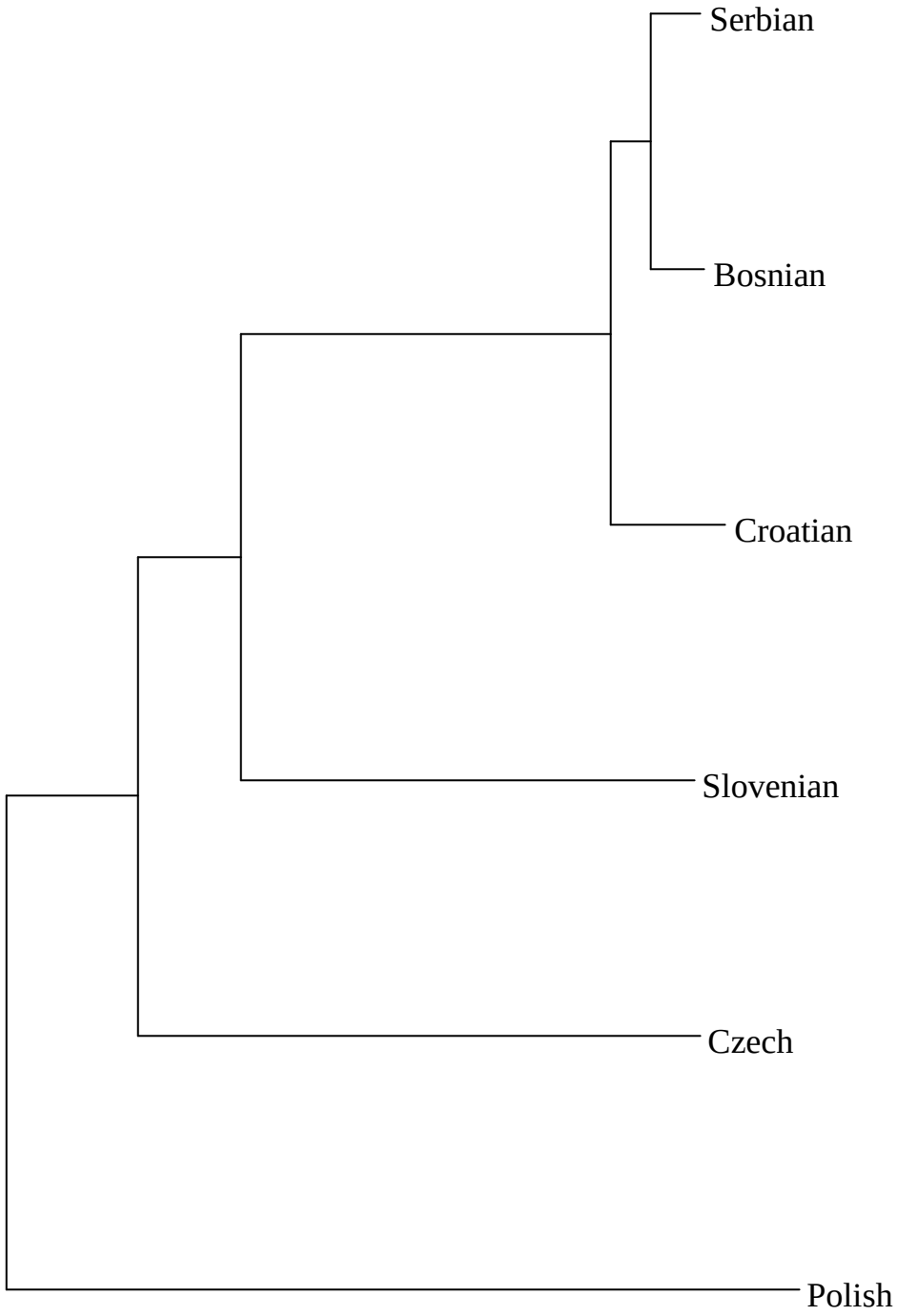
# Language Evolution

- We compute the pairwise distance matrix obtained with the expressions above on the second dataset.
- We use the SLS representations:
  - di-gram and odds ratio with n=2
  - P-spectrum kernels with p=4
  - Mismatch kernel with p=4 and m=1.
- We use the standard algorithm Neighbor Joining to reconstruct a phylogenetic trees.

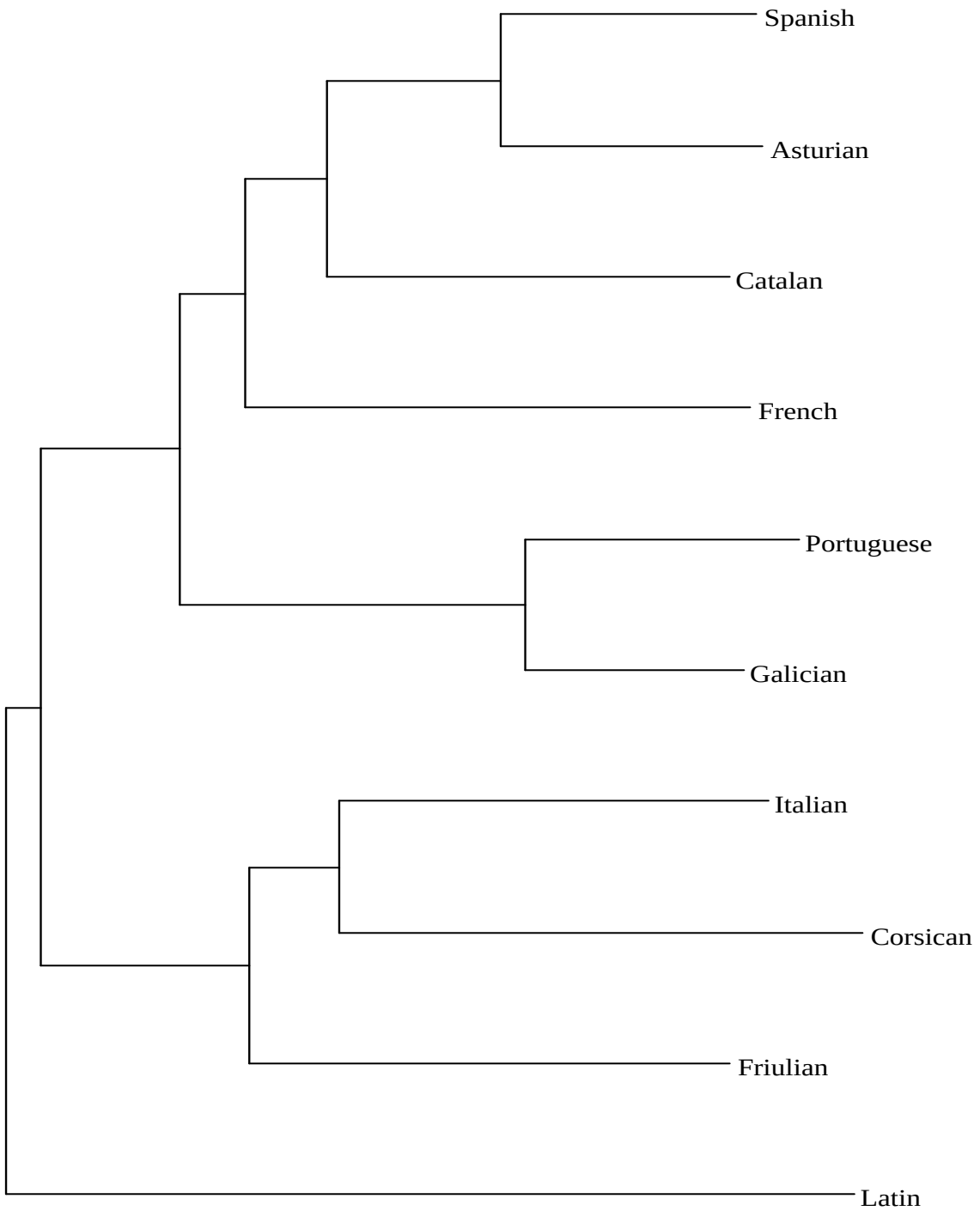
# Language Evolution



Slavic Branch

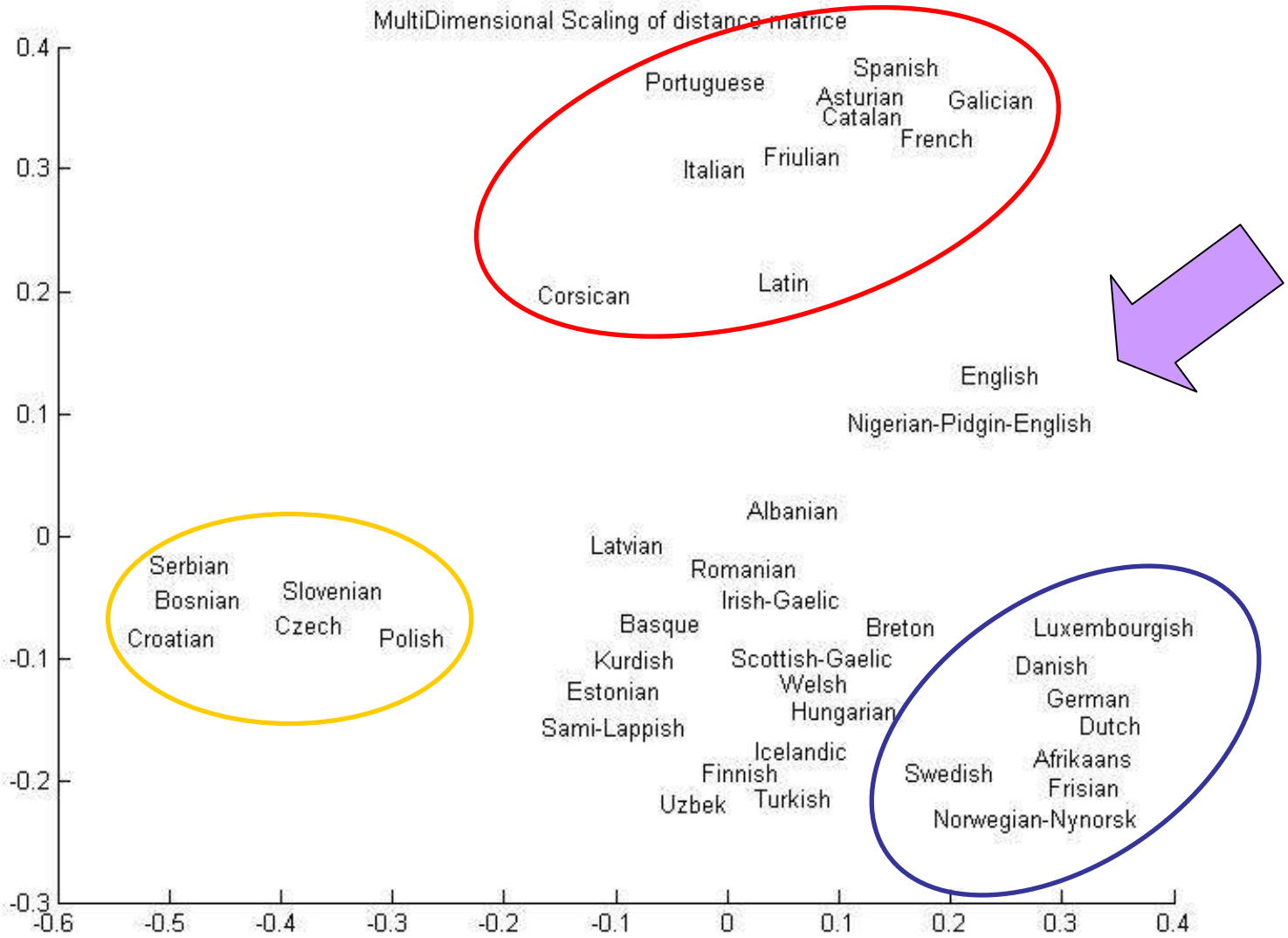


Romance Branch



# Language Evolution

# Multi Dimensional Scaling



# Conclusions

- The SLS approach is independent of any subjective choice and is repeatable automatically.

- The resulting tree is generally in agreement with the commonly accepted view of the IE family, although some exceptions.

# Conclusions

- Some aspects of historical linguistics can be investigated by using statistical tools, but:
  - this approach is robust enough to ignore the effect of alternative spelling conventions, but it produces a mistake (Breton);
  - we see the effect of borrowings (English and Romanian).
- The assumption that the evolutionary history of languages can be represented by a tree is not justified: "phylogenetic networks".

# Analysis of patterns using textual information and temporal relations

We study two different interactions of these factors:

1. Given a sequence of written documents in different languages, we want to find temporal relations among languages.

2. Given a sequence of documents in time, we want to find patterns which merge textual information and temporal relations.

# Event Detection

- Analysis of patterns using textual information and temporal relations.
- What is an event?
- KP extractor.
- KP selector.
- The Algorithms
  - HMM.
  - Event Detection.
- Results.
- Conclusions.

| Fri Apr 1 2005 | Sat Apr 2 2005 | Sun Apr 3 2005 | Mon Apr 4 2005 | Tue Apr 5 2005 | Wed Apr 6 2005 | Thu Apr 7 2005 |
|---|---|---|---|---|---|---|
| U S | John Paul | John Paul | John Paul | St Peter | U S | U S |
| United States | St Peter | U S | U S | Prime Minister | United States | John Paul |
| the United States | U S | John Paul II | John Paul II | John Paul II | the United States | St Peter |
| John Paul | St Peter s | the United States | St Peter | U N | the U S | United States |
| the U S | The pope | The pope | St Peter s | New York | John Paul II | the U S |
| North Korea | John Paul II | Pope John | United States | John Paul | Prime Minister | St Peter s |
| Michael Schiavo | United States | St Peter s | the U S | the United States | Hong Kong | the United States |
| General Re | Peter s Square | Pope John Paul | Catholic Church | United States | Navarro Valls | John Paul II |
| Khmer Rouge | St Peter s Square | U N | iron ore | the U S | New York | Editor s Note |
| Paul II | Navarro Valls | the U S | the United States | U S | John Paul | Prime Minister |

| Fri Apr 8 2005 | Sat Apr 9 2005 | Sun Apr 10 2005 | Mon Apr 11 2005 | Tue Apr 12 2005 | Wed Apr 13 2005 | Thu Apr 14 2005 |
|---|---|---|---|---|---|---|
| U S | U S | U S | Prime Minister | the United States | long term | the United States |
| John Paul | the U S | United States | the U N | the U S | Parkinson s disease | the U S |
| John Paul II | U N | the United States | the West Bank | New York | All comments | O REILLY |
| St Peter | Associated Press | the U S | road map | human rights | chief executive | New York |
| United States | United States | Prime Minister | John Paul II | Prime Minister | the New York | South Korean |
| St Peter s | Saturday April <N | John Paul | New York | European Union | Prime Minister | Security Council |
| the United States | the United States | years ago | the United Nations | John Paul II | the U S | Web site |
| the U S | tribunal president | Al Aqsa | the United States | Hugo Boss | the United States | European Union |
| Pope John | Al Aqsa | Charles and Camilla | the U S | United States and | State Department | oil for food program |
| Pope John Paul | Parker Bowles | Hong Kong | India and China | security forces | Schr der | Al Jazeera |

| Fri Apr 15 2005 | Sat Apr 16 2005 | Sun Apr 17 2005 | Mon Apr 18 2005 | Tue Apr 19 2005 | Wed Apr 20 2005 | Thu Apr 21 2005 |
|---|---|---|---|---|---|---|
| the United States | White House | White House | the United States | the United States | John Paul II | the U S |
| the U S | Prime Minister | security forces | North Korea | the U S | the U S | the United States |
| credit card | the United States | Security Council | the U S | Prime Minister | New York | John Paul |
| John Paul | Saturday April | around the world | New York | Pope Benedict XVI | Catholic Church | human rights |
| cord blood | U N | S military | O REILLY | Pope John Paul II | North Korea | FOX News |
| Prime Minister | the U S | West Bank | Associated Press | Catholic Church | President Bush | Prime Minister |
| Prime Minister T | Associated Press | John Paul II | John Paul II | a year earlier | the United Nations | New York |
| New York | told CNN | Prime Minister | Prime Minister | St Peter | Pope Benedict XVI | a year earlier |
| O REILLY | mass graves | the United States | a year earlier | New York | Joseph Ratzinger | South Korea |
| Lori Hacking | Saddam Hussein | the U S | St Peter | around the world | cents per share | U N |

| Fri Apr 22 2005 | Sun Apr 24 2005 | Mon Apr 25 2005 | Tue Apr 26 2005 | Wed Apr 27 2005 | Thu Apr 28 2005 | Fri Apr 29 2005 | Sat Apr 30 2005 |
|---|---|---|---|---|---|---|---|
| the United States | the United States | North Korea | the United States | the United States | the United States | the Middle East | the U S |
| the U S | John Paul | the United States | the U S | the U S | the U S | the D A | North Korea |
| Abu Ghraib | the U S | the U S | North Korea | al Qaeda | South Korea | the S P | the United States |
| North Korea | al Qaeda | New York | al Zarqawi | al Jaafari | silicone breast imp | the West Bank | Saturday April |
| Al Jazeera | Associated Press | Prime Minister | New York | New York | Social Security | S military | S military |
| New York | North Korea | the Czech Republic | FOX News | President Bush | White House | blood sugar | Prime Minister |
| White House | Abu Ghraib | Associated Press | Social Security | Associated Press | news conference | Social Security | Los Angeles |
| FOX News | Prime Minister | South Korea | Bush administration | sex offenders | human rights | the United States | Wake Island |
| al Qaeda | St Peter | the White House | secretary general | the State Departmen | Prime Minister | the U S | On Friday |
| H usser | death penalty | Bush administration | oil for food | North Korea | New York | U N | a U S |

# Textual Time Series

- Given a time series of documents we want to:

**find automatically**

**the most “relevant” events**

What do we need?

1. *Informative features.*
2. *Techniques that merge temporal relations with textual information.*

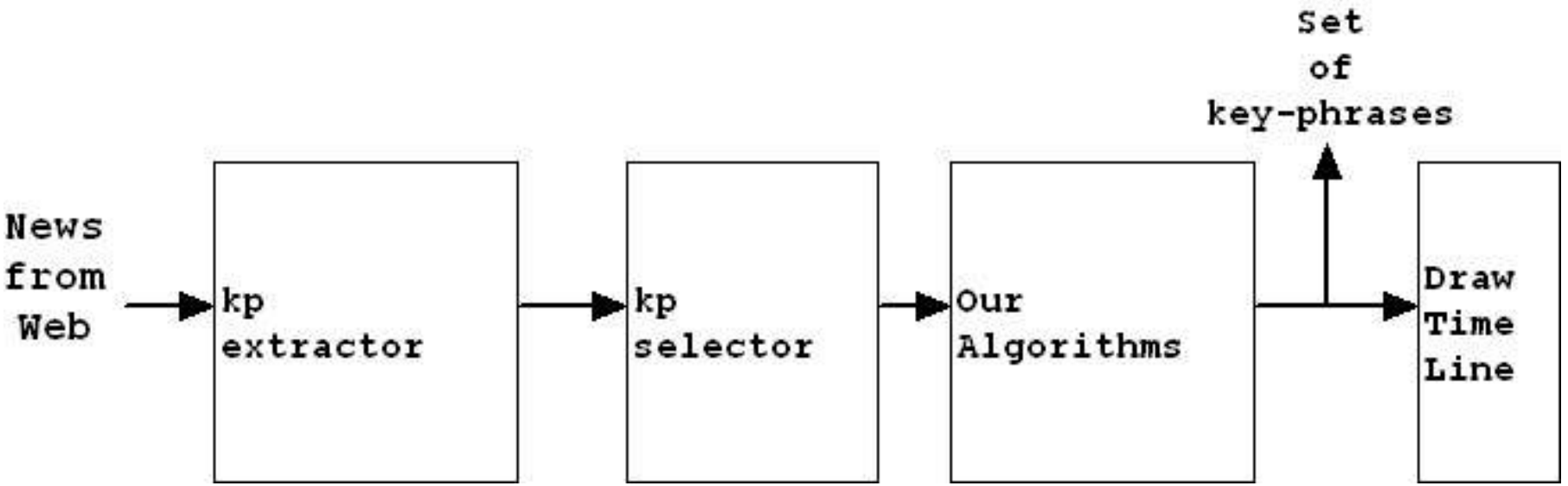
# What is an event?

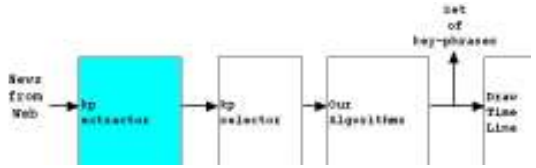
- We consider an event:
  - one or more than one terms that appear continuously in time.
- An event is relevant if:
  - the sequence of instants of times where these terms appear has a length include between *t* and *q* days;
  - the terms that identify the event are meaningful and represent clearly a concept.

# Type of Events

1. ***Long event***: a term appears slowly, but with a constant increase. These events have long tails before they appear and are relatively long.
2. ***Middle long event***: on a given day a term focused on the same topic appears. This entrance is abrupt and swift, but in general the event which causes it is not long, but it continues for some days..
3. ***Short event***: a term appears few times and the event persists for few days.
4. ***Hidden event***: the term appears when there are not other important events, and disappear when something happens. It is latent and emerges only in particular conditions. This kind of event in a lot of cases can be considered noise.

# Work Pipeline






# KP Extractor

- For each day extracts the top ten key-phrases.
- A key-phrases is a sequence of continuous words, where in a corpus

$$P(xyz)>P(x)P(y)P(z)$$

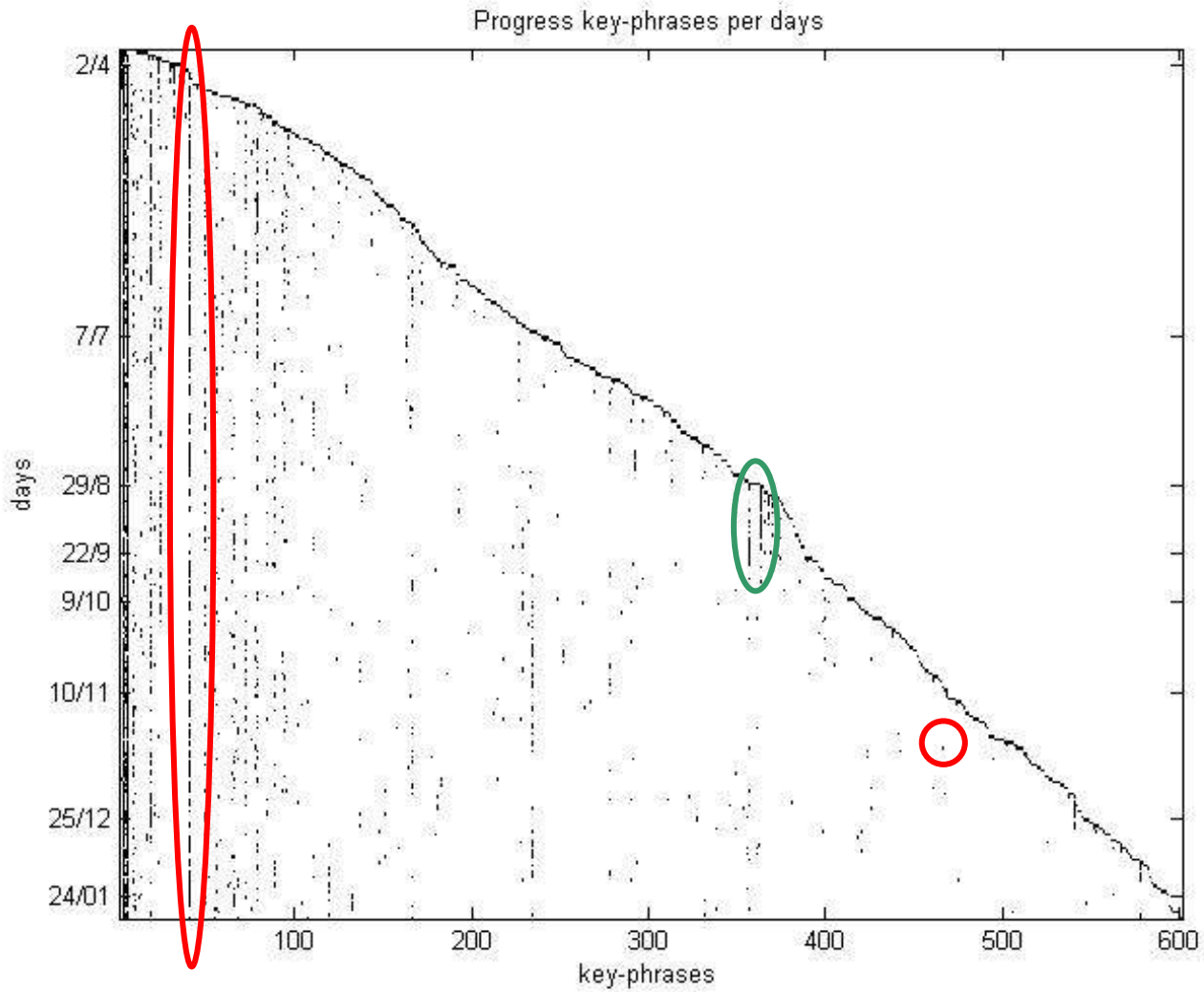
- It runs a slide windows on documents and counts the number of time that each possible key-phrases occur.
- Ranks the key-phrases using the following criteria:
  - KP frequency
  - penalizes KPs which contain a lot of stopwords.

# KP Extractor

- For each day it produces:

| *KP* | *Frequency* | *Id* |
|---|---|---|
| the U S | 161 | 115 |
| the United States | 95 | 113 |
| John Paul II | 91 | 142 |
| the United Nations | 68 | 123 |
| the West Bank | 45 | 120 |
| the U N | 40 | 169 |
| New York | 27 | 129 |
| road map | 22 | 168 |
| Prime Minister | 16 | 158 |
| India and China | 15 | 170 |

# Progress KPs per days

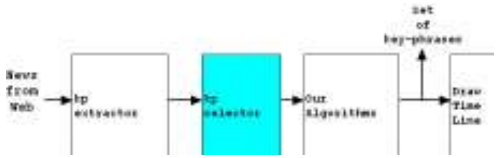


| Fri Apr 1 2005 | Sat Apr 2 2005 | Sun Apr 3 2005 | Mon Apr 4 2005 | Tue Apr 5 2005 | Wed Apr 6 2005 | Thu Apr 7 2005 |
|---|---|---|---|---|---|---|
| U S | John Paul | John Paul | John Paul | St Peter | U S | U S |
| United States | St Peter | U S | U S | Prime Minister | United States | John Paul |
| the United States | U S | John Paul II | John Paul II | John Paul II | the United States | St Peter |
| John Paul | St Peter s | the United States | St Peter | U N | the U S | United States |
| the U S | The pope | The pope | St Peter s | New York | John Paul II | the U S |
| North Korea | John Paul II | Pope John | United States | John Paul | Prime Minister | St Peter s |
| Michael Schiavo | United States | St Peter s | the U S | the United States | Hong Kong | the United States |
| General Re | Peter s Square | Pope John Paul | Catholic Church | United States | Navarro Valls | John Paul II |
| Khmer Rouge | St Peter s Square | U N | iron ore | the U S | New York | Editor s Note |
| Paul II | Navarro Valls | the U S | the United States | U S | John Paul | Prime Minister |

| Fri Apr 8 2005 | Sat Apr 9 2005 | Sun Apr 10 2005 | Mon Apr 11 2005 | Tue Apr 12 2005 | Wed Apr 13 2005 | Thu Apr 14 2005 |
|---|---|---|---|---|---|---|
| U S | U S | U S | Prime Minister | the United States | long term | the United States |
| John Paul | the U S | United States | the U N | the U S | Parkinson s disease | the U S |
| John Paul II | U N | the United States | the West Bank | New York | All comments | O REILLY |
| St Peter | Associated Press | the U S | road map | human rights | chief executive | New York |
| United States | United States | Prime Minister | John Paul II | Prime Minister | the New York | South Korean |
| St Peter s | Saturday April <N | John Paul | New York | European Union | Prime Minister | Security Council |
| the United States | the United States | years ago | the United Nations | John Paul II | the U S | Web site |
| the U S | tribunal president | Al Aqsa | the United States | Hugo Boss | the United States | European Union |
| Pope John | Al Aqsa | Charles and Camilla | the U S | United States and | State Department | oil for food program |
| Pope John Paul | Parker Bowles | Hong Kong | India and China | security forces | Schr der | Al Jazeera |

| Fri Apr 15 2005 | Sat Apr 16 2005 | Sun Apr 17 2005 | Mon Apr 18 2005 | Tue Apr 19 2005 | Wed Apr 20 2005 | Thu Apr 21 2005 |
|---|---|---|---|---|---|---|
| the United States | White House | White House | the United States | the United States | John Paul II | the U S |
| the U S | Prime Minister | security forces | North Korea | the U S | the U S | the United States |
| credit card | the United States | Security Council | the U S | Prime Minister | New York | John Paul |
| John Paul | Saturday April | around the world | New York | Pope Benedict XVI | Catholic Church | human rights |
| cord blood | U N | S military | O REILLY | Pope John Paul II | North Korea | FOX News |
| Prime Minister | the U S | West Bank | Associated Press | Catholic Church | President Bush | Prime Minister |
| Prime Minister T | Associated Press | John Paul II | John Paul II | a year earlier | the United Nations | New York |
| New York | told CNN | Prime Minister | Prime Minister | St Peter | Pope Benedict XVI | a year earlier |
| O REILLY | mass graves | the United States | a year earlier | New York | Joseph Ratzinger | South Korea |
| Lori Hacking | Saddam Hussein | the U S | St Peter | around the world | cents per share | U N |

| Fri Apr 22 2005 | Sun Apr 24 2005 | Mon Apr 25 2005 | Tue Apr 26 2005 | Wed Apr 27 2005 | Thu Apr 28 2005 | Fri Apr 29 2005 | Sat Apr 30 2005 |
|---|---|---|---|---|---|---|---|
| the United States | the United States | North Korea | the United States | the United States | the United States | the Middle East | the U S |
| the U S | John Paul | the United States | the U S | the U S | the U S | the D A | North Korea |
| Abu Ghraib | the U S | the U S | North Korea | al Qaeda | South Korea | the S P | the United States |
| North Korea | al Qaeda | New York | al Zarqawi | al Jaafari | silicone breast imp | the West Bank | Saturday April |
| Al Jazeera | Associated Press | Prime Minister | New York | New York | Social Security | S military | S military |
| New York | North Korea | the Czech Republic | FOX News | President Bush | White House | blood sugar | Prime Minister |
| White House | Abu Ghraib | Associated Press | Social Security | Associated Press | news conference | Social Security | Los Angeles |
| FOX News | Prime Minister | South Korea | Bush administration | sex offenders | human rights | the United States | Wake Island |
| al Qaeda | St Peter | the White House | secretary general | the State Departmen | Prime Minister | the U S | On Friday |
| H usser | death penalty | Bush administration | oil for food | North Korea | New York | U N | a U S |




# KP selector

1. Computes the Bag of Words kernel on all the key phrases. It produces a kernel matrix, where each entry is the number of words that two key-phrases share.

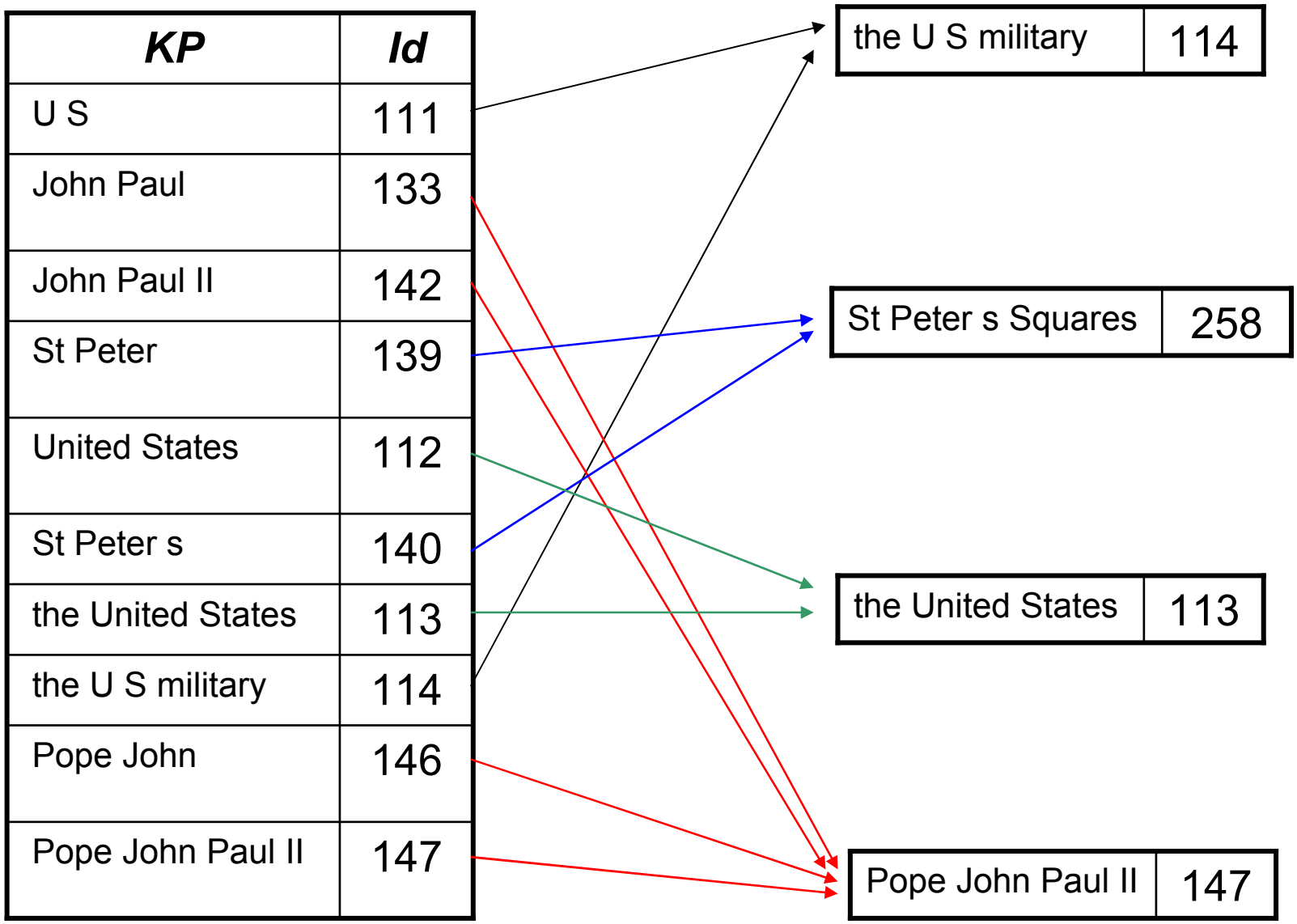
2. Analyzing the kernel matrix, it checks if a KP is completely embedded in another. If it is, they are clustered.

3. For each cluster, the common longest sequence has been found, and it has been used to labeled the cluster.

4. Data have been modified using the cluster information substituting  all the KPs of the cluster with its label.

# KP selector

| *KP* | *Id* |
|---|---|
| U S | 111 |
| John Paul | 133 |
| John Paul II | 142 |
| St Peter | 139 |
| United States | 112 |
| St Peter s | 140 |
| the United States | 113 |
| the U S military | 114 |
| Pope John | 146 |
| Pope John Paul II | 147 |



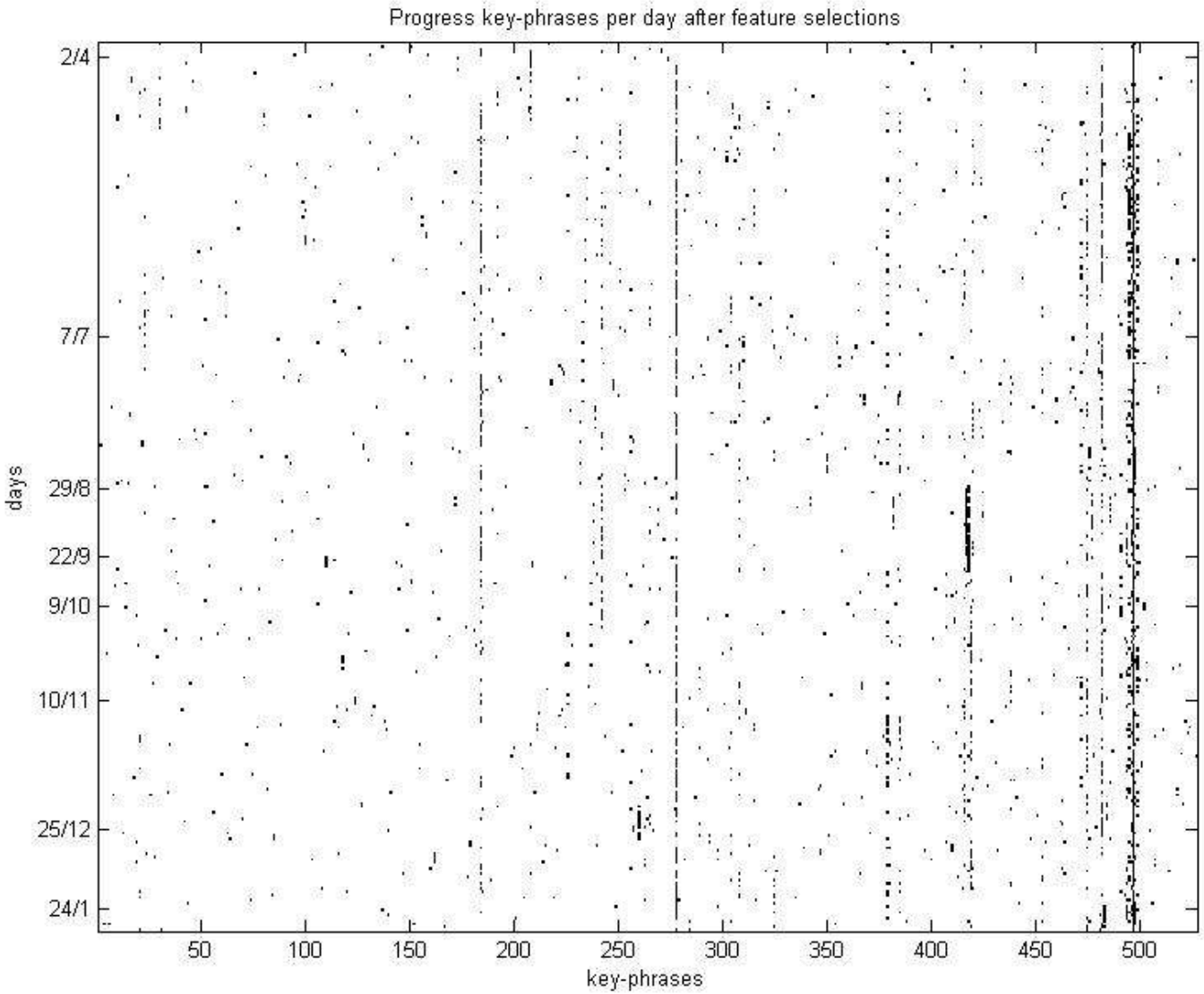
| Fri Apr 1 2005 | Sat Apr 2 2005 | Sun Apr 3 2005 | Mon Apr 4 2005 | Tue Apr 5 2005 | Wed Apr 6 2005 | Thu Apr 7 2005 |
|---|---|---|---|---|---|---|
| Pope John Paul II | Pope John Paul II | Pope John Paul II | Pope John Paul II | Pope John Paul II | Pope John Paul II | Pope John Paul II |
| the United States | St Peter s Square | The U S military | the U S military | Prime Minister Tony Blair | Prime Minister Tony Blair | the U S military |
| the U S military | the U S military | the United States | Catholic Church | St Peter s Square | the U S military | St Peter s Square |
| North Korea s | The pope | The pope | St Peter s Square | the United States | the United States | the United States |
| Michael Schiavo | The United States | St Peter s Square | iron ore | the New York | of Hong Kong | Editor s Note |
| General Re | Navarro Valls | the U N Security Council | the United States | the U N Security Council | the New York | Prime Minister Tony Blair |
| Khmer Rouge | | | | the U S military | Navarro Valls | |
| st Peter s Square | | | | | | |

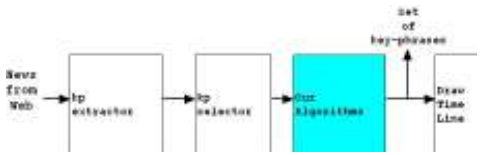
| Fri Apr 8 2005 | Sat Apr 9 2005 | Sun Apr 10 2005 | Mon Apr 11 2005 | Tue Apr 12 2005 | Wed Apr 13 2005 | Thu Apr 14 2005 |
|---|---|---|---|---|---|---|
| Pope John Paul II | the U S military | Pope John Paul II | Pope John Paul II | Pope John Paul II | long term | the United States |
| The U S military | tribunal president | the United States | the West Bank | the U S military | Parkinson s disease | the U S military |
| St Peter s Square | Associated Press | Prime Minister Tony Blair | the U N Security Council | the New York | All comments | O REILLY |
| the United States | Saturday April <N | The U S military | road map | human rights | chief executive | the New York |
| | the United States | years ago | the New York | Prime Minister Tony Blair | the New York | South Korean |
| | the U N Security Council | Al Aqsa | Prime Minister Tony Blair | the European Union | Prime Minister Tony Blair | the U N Security Council |
| | Al Aqsa | Charles and Camilla | the United Nations | the United States | the U S military | Web site |
| | Parker Bowles | of Hong Kong | the United States | Hugo Boss | the United States | the European Union |
| | | | the U S military | the United States | the State Department | oil for food program |
| | | | India and China | security forces | Schr der | Al Jazeera |

| Fri Apr 15 2005 | Sat Apr 16 2005 | Sun Apr 17 2005 | Mon Apr 18 2005 | Tue Apr 19 2005 | Wed Apr 20 2005 | Thu Apr 21 2005 |
|---|---|---|---|---|---|---|
| Pope John Paul I | the White House | Pope John Paul II | Pope John Paul II | Pope John Paul II | Pope John Paul II | the U S military |
| the U S military | Prime Minister Tony Blair | Iraqui security forces | North Korea s | the U S military | the U S military | the United States |
| credit card | the United States | the U N Security Council | the U S military | Prime Minister Tony Blair | the New York | Pope John Paul II |
| the United States | Saturday April | around the world | the New York | Pope Benedict XVI | Catholic Church | human rights |
| Prime Minister Tony Blair | the U N Security Council | West Bank | O REILLY | the United States | North Korea s | FOX News |
| cord blood | the U S military | the White House | Associated Press | Catholic Church | President George W Bush | Prime Minister Tony Blair |
| the New York | Associated Press | Prime Minister Tony Blair | the United States | a year earlier | the United Nations | the New York |
| O REILLY | told CNN | the U S military | Prime Minister Tony Blair | St Peter s Square | Pope Benedict XVI | a year earlier |
| Lori Hacking | mass graves | the United States | a year earlier | the New York | Joseph Ratzinger | in South Korea |
| | Saddam Hussein | | St Peter s Square | around the world | cents per share | the U N Security Council |

| Fri Apr 22 2005 | Sun Apr 24 2005 | Mon Apr 25 2005 | Tue Apr 26 2005 | Wed Apr 27 2005 | Thu Apr 28 2005 | Fri Apr 29 2005 | Sat Apr 30 2005 |
|---|---|---|---|---|---|---|---|
| the United States | Pope John Paul II | North Korea s | the United States | the United States | the United States | the Middle East | the U S military |
| the U S military | the United States | the United States | the U S military | the U S military | the U S military | the D A | North Korea s |
| Abu Ghraib | the U S military | the U S military | North Korea s | al Qaeda | South Korea | the S P | the United States |
| North Korea s | al Qaeda | the New York | al Zarqawi | al Jaafari | silicone breast implants | the West Bank | Saturday April |
| Al Jazeera | Associated Press | Prime Minister Tony Blair | the New York | the New York | Social Security beneficts | blood sugar | the U S military |
| the New York | North Korea s | the Czech Republic | FOX News | President George W Bush | the White House | Social Security benefict | Prime Minister Tony Blair |
| the White House | Abu Ghraib | Associated Press | Social Security benefict | Associated Press | news conference | the United States | Los Angeles Times |
| FOX News | Prime Minister Tony Blair | South Korea | the Bush administration | sex offenders | human rights | the U S military | Wake Island |
| al Qaeda | St Peter s Square | the White House | secretary general | the State Department | Prime Minister Tony Blair | the U N Security Council | On Friday |
| H usser | the death penalty | the Bush administration | oil for food program | North Korea s | the New York | | a U S |

# KP selector


Progress key-phrases per day after feature selections
2/4
7/7
29/8
22/9
9/10
10/11
25/12
24/1
days
50
100
150
200
250
300
350
400
450
500
key-phrases




# The Algorithms

- Hidden Markov Model.
- Event Detection.

# Hidden Markov Model

- We suppose that all the possible KP sources are activated every day.
- We associate the same HMM two states for each source and each source can emit or not only one KP.
- For each KP, the HMM produces the hidden sequence.
- The KPs have been ranked using the total probability for each hidden sequence.

# Hidden Markov Model

- The states of the HMM are:
  - $s_i^t$ = the source *i* has emitted at the time *t*
  - *not* $s_i^t$ = the source *i* has not emitted at the time *t.*

- The *hidden states are :*
  - $o_i^t$ = the key-phrases *i* is present at the time *t*
  - *not* $o_i^t$ = the key-phrases *i* is not present at the time *t*.

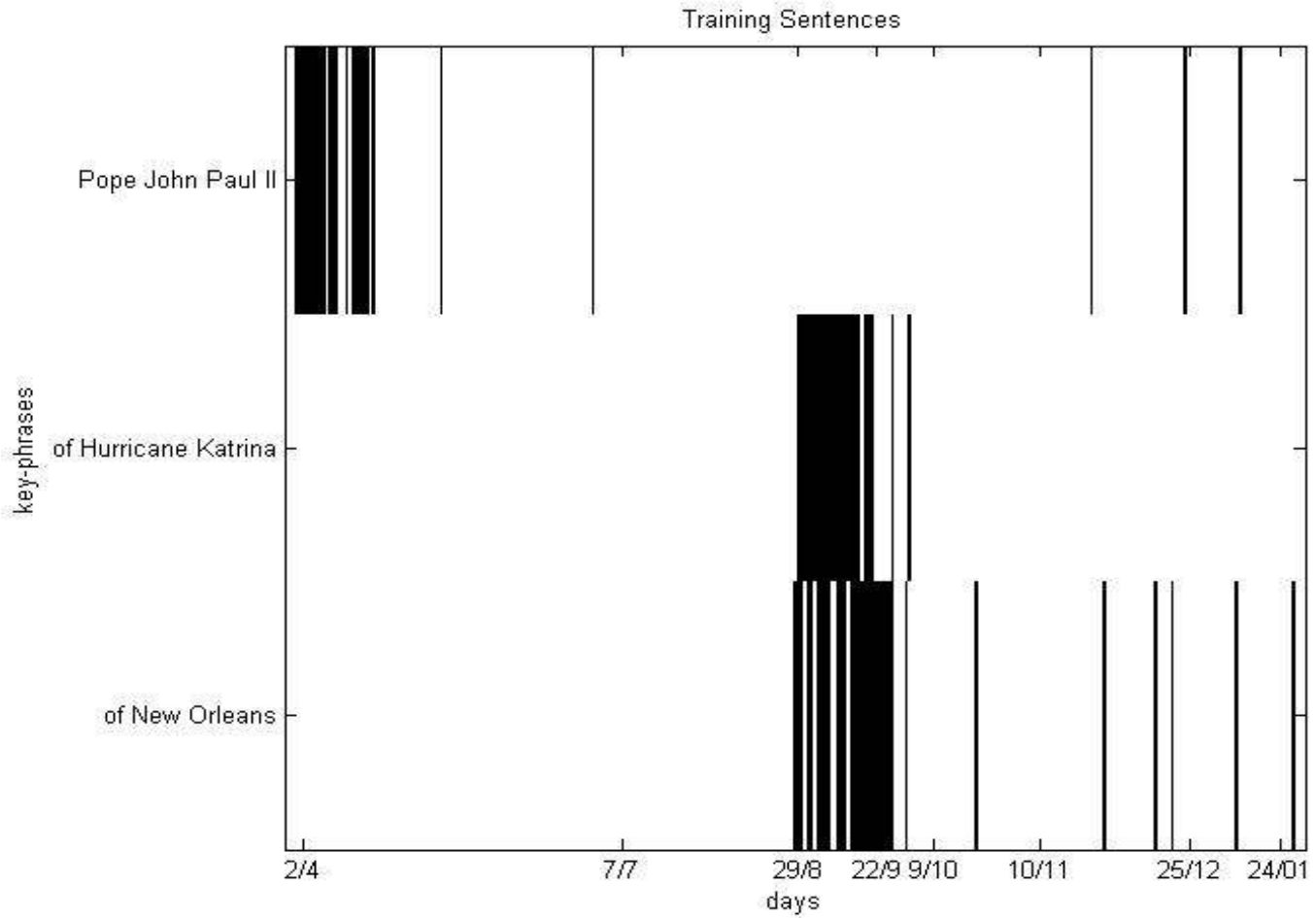
- The output alphabet is: 0,1.

# Hidden Markov Model

- We train just one time the model on 3 selected KPs.
- Complexity cost: $\#iteration_{EM}$ * O((N^2)T), where N is the number of states, T is the length of the sequence.
- We run the model on all the KPs producing the hidden sequences.
- Complexity cost: m * O((N^2)T) where m is the number of KPs

# Hidden Markov Model

- We train the HMM using the following sequences:



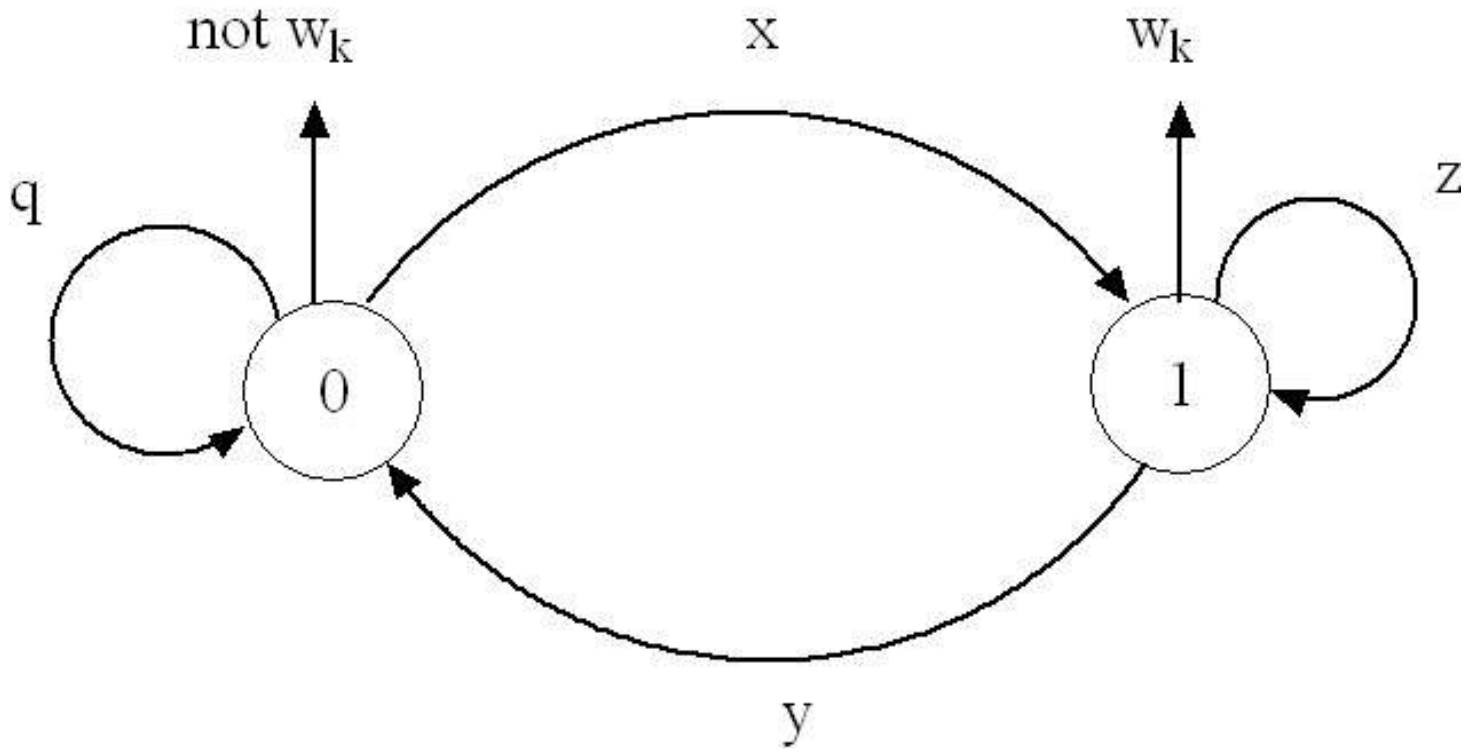
# Hidden Markov Model

- We obtain the following transition matrix T and emission matrix E:

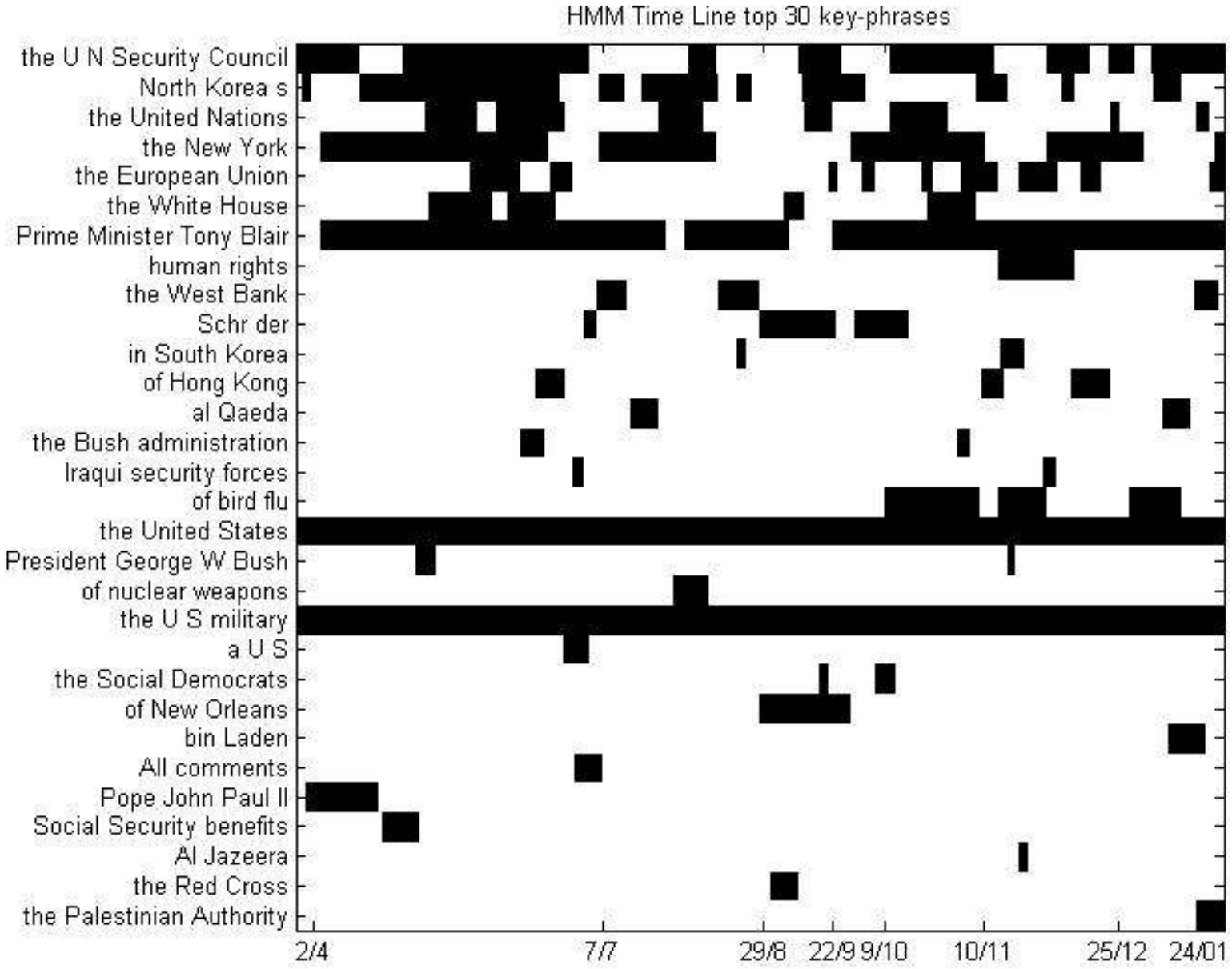
$$T = \begin{pmatrix} 0,9961 & 0,0038 \\ 0,0408 & 0,9592 \end{pmatrix}$$

$$E = \begin{pmatrix} 0,9832 & 0,0167 \\ 0,1588 & 0,8411 \end{pmatrix}$$

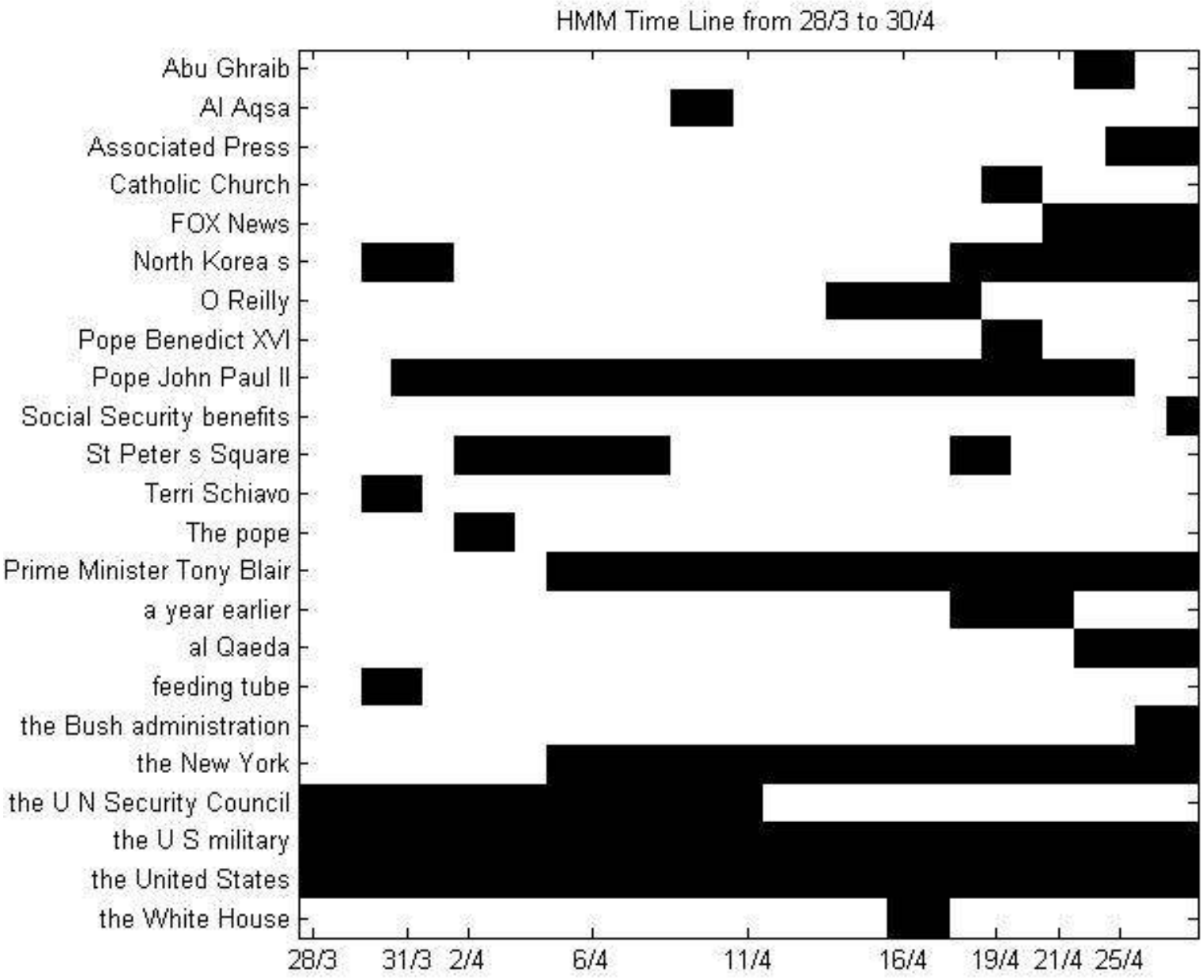
# Hidden Markov Model



# Hidden Markov Model



# Hidden Markov Model



# Hidden Markov Model

- Problem:
  - selects some KPs which do not respect the concept of “relevant” event.

  Like:

  Prime Minister Tony Blair, the United States, the U S military, the U N Security Council, North Korea s and the New York

# Event Detection

- We use a *random walk* to scan each time series obtained for a key-phrase and to discover the trend on each of them.

- The key-phrases which have a particular shape and persist for defined instants of time are selected.

- The algorithm  automatically provides the most relevant KPs.

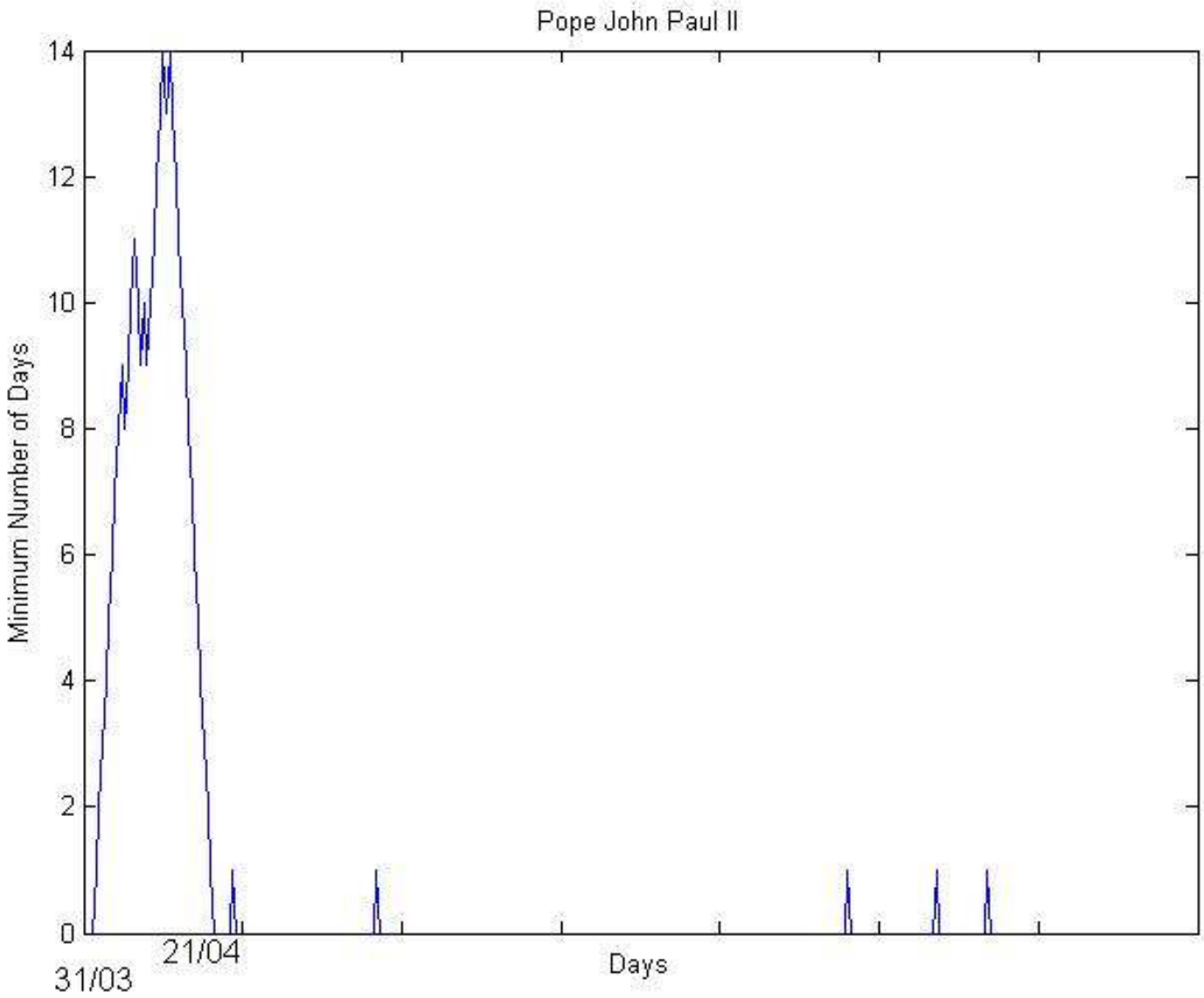
# Event Detection

- A random walk is a formalization of the intuitive idea of taking successive steps, each in a random direction.

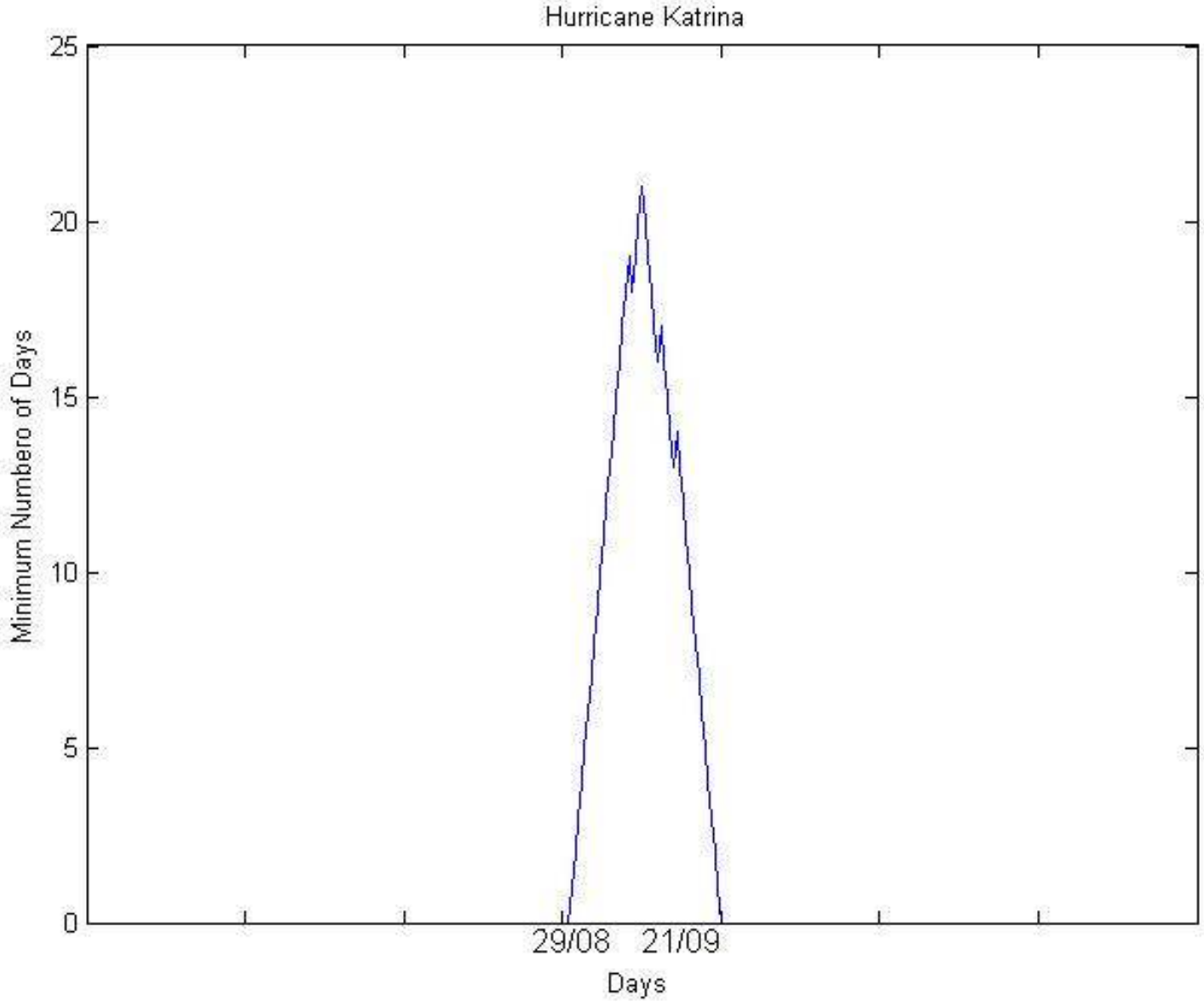
- We allow the random walk to advance in either direction by a fix step, -1 when $m_{i,j} = 0$, +1 when $m_{i,j} = 1$.

- We limit the random walk on the left. When it finds a new 0, and the sum from the beginning to now is equal to 0, the sum can not go below 0.

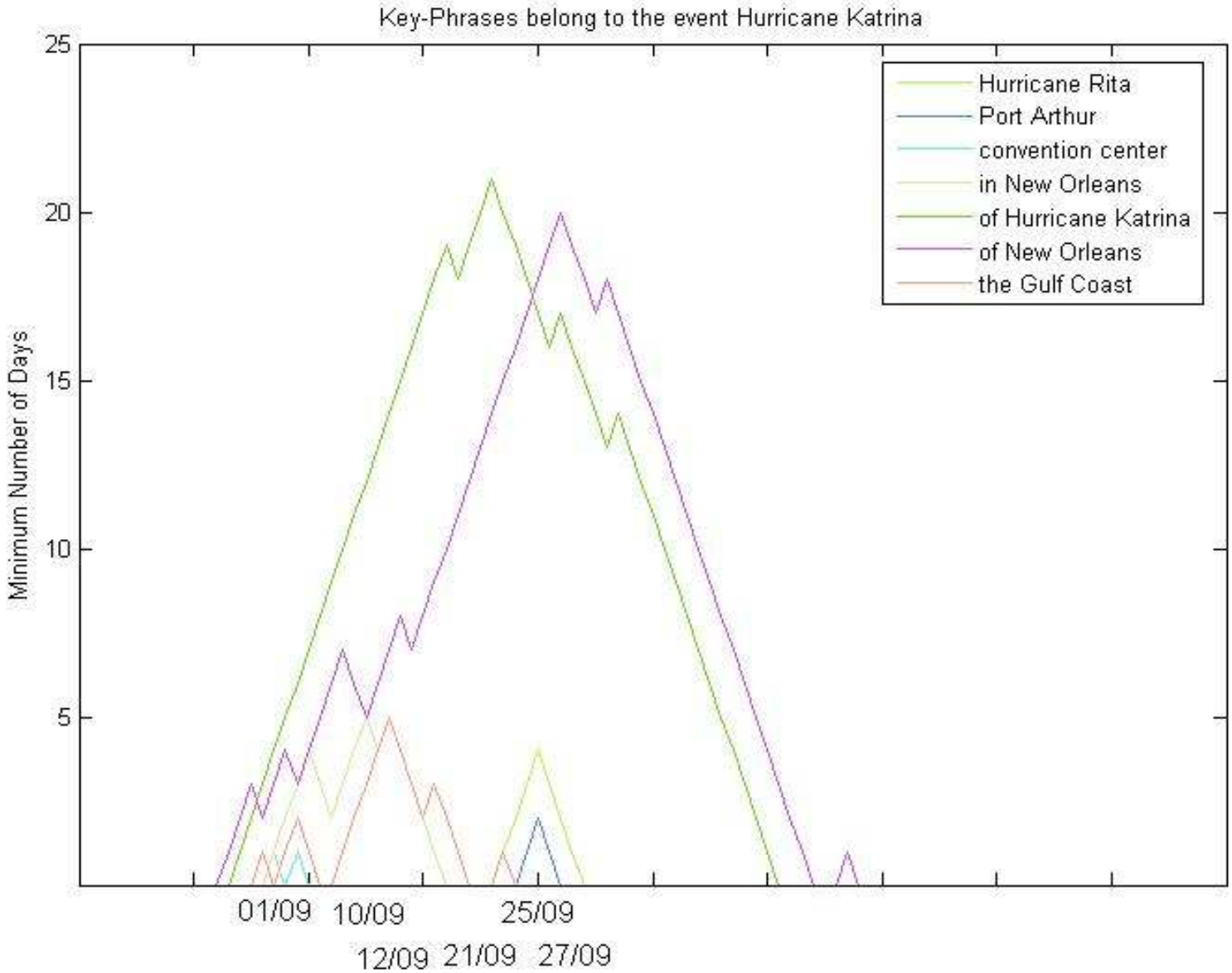
# Event Detection



# Event Detection



# Event Detection



# Event Detection

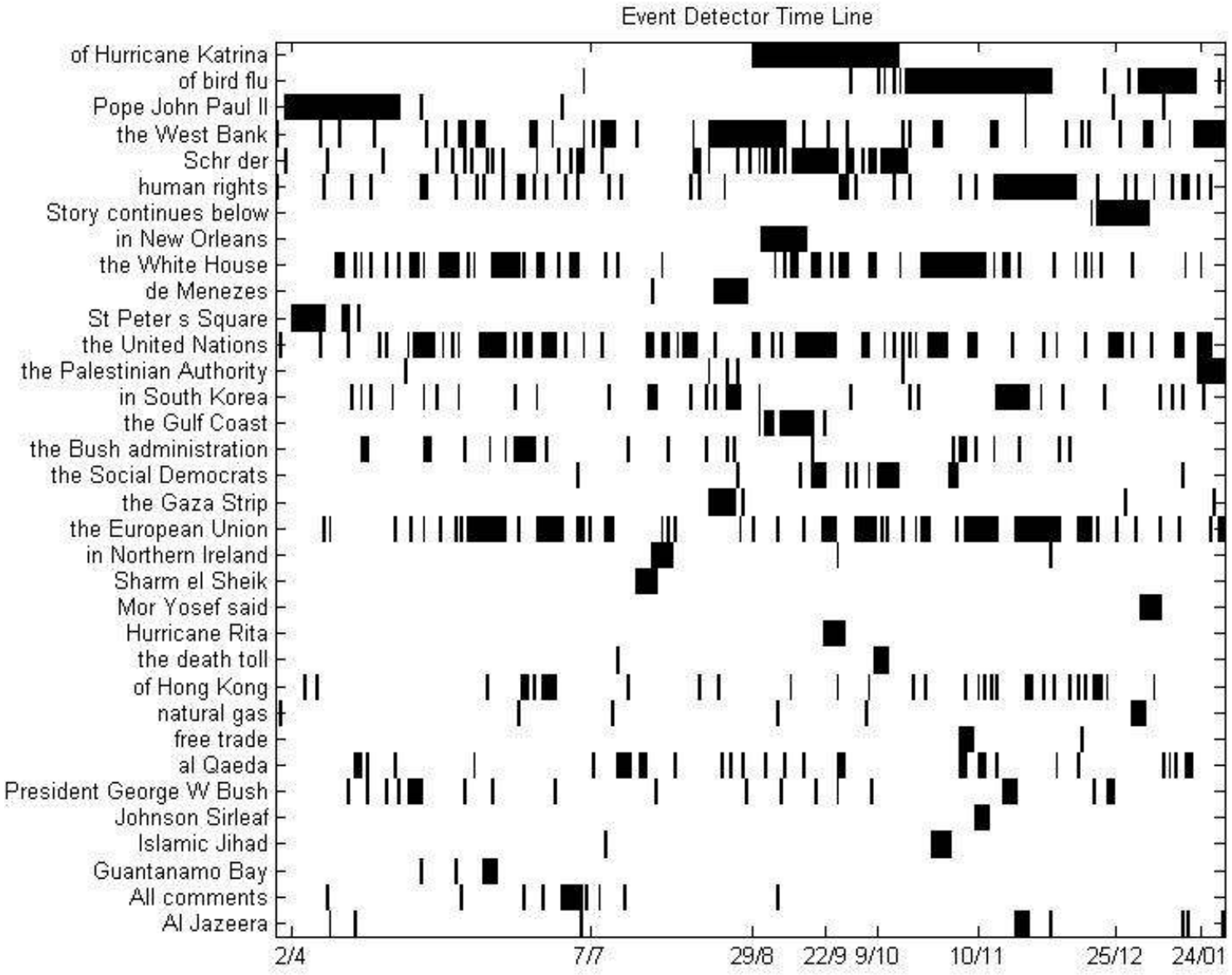
- For each KP, we store:
  - the maximum value that the random walk reaches during its walk and its position;
  - the initial and the final position of the event where the maximum was placed.

- We consider as length of the event the distance between the position where the maximum was placed and the position where the event began.

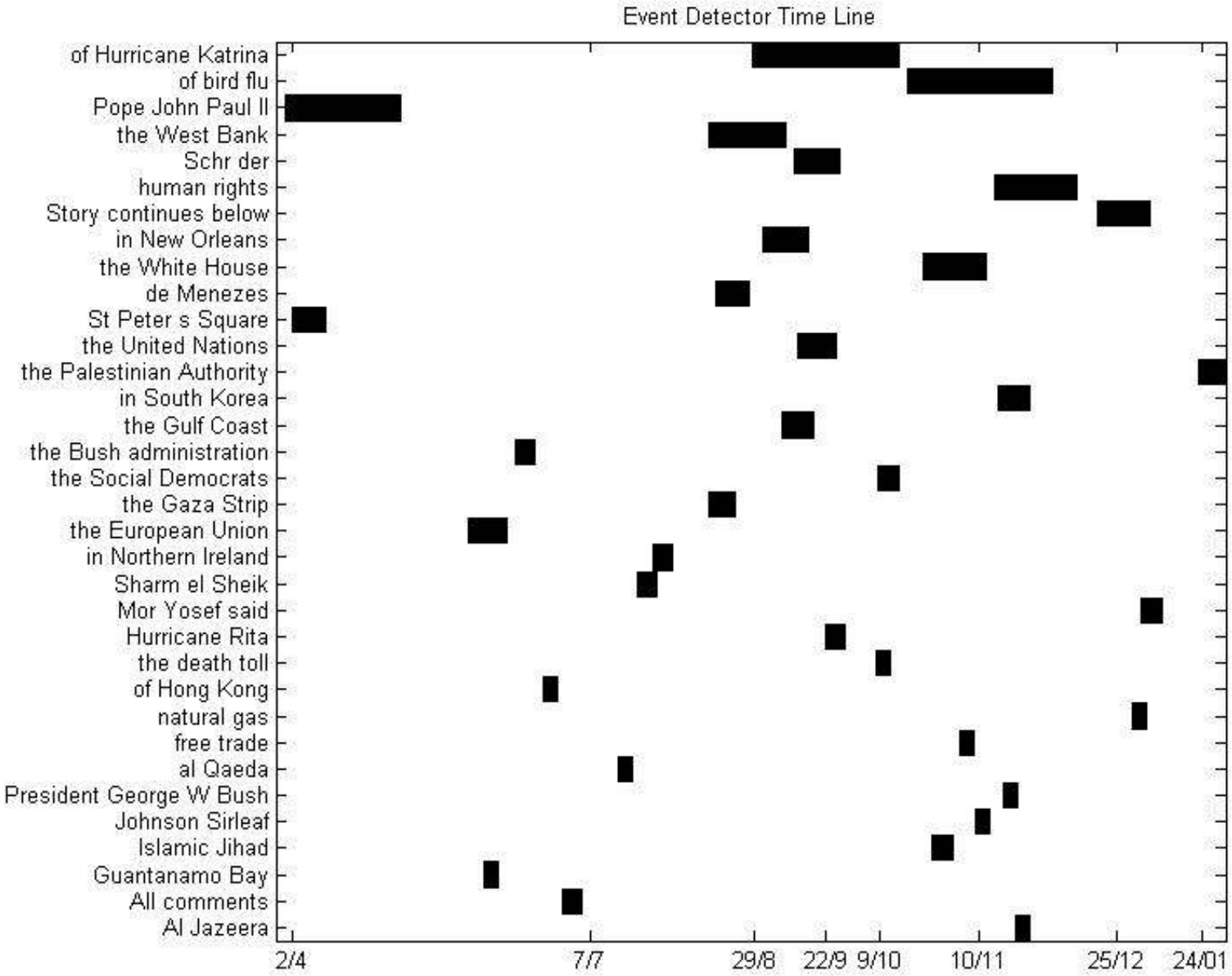
# Event Detection

- For each automatically selected KP, we can plot:
  - *all the events* considering an event from its beginning day to its ending;
  - *only the main event* considering this event from its *beginning* to its *ending* day;
  - only the *main event* considering this event from its *beginning* day to the day where the *maximum* has been placed;
  - it has cost O(mn), where m is the number of KPs, and n is the number of days.

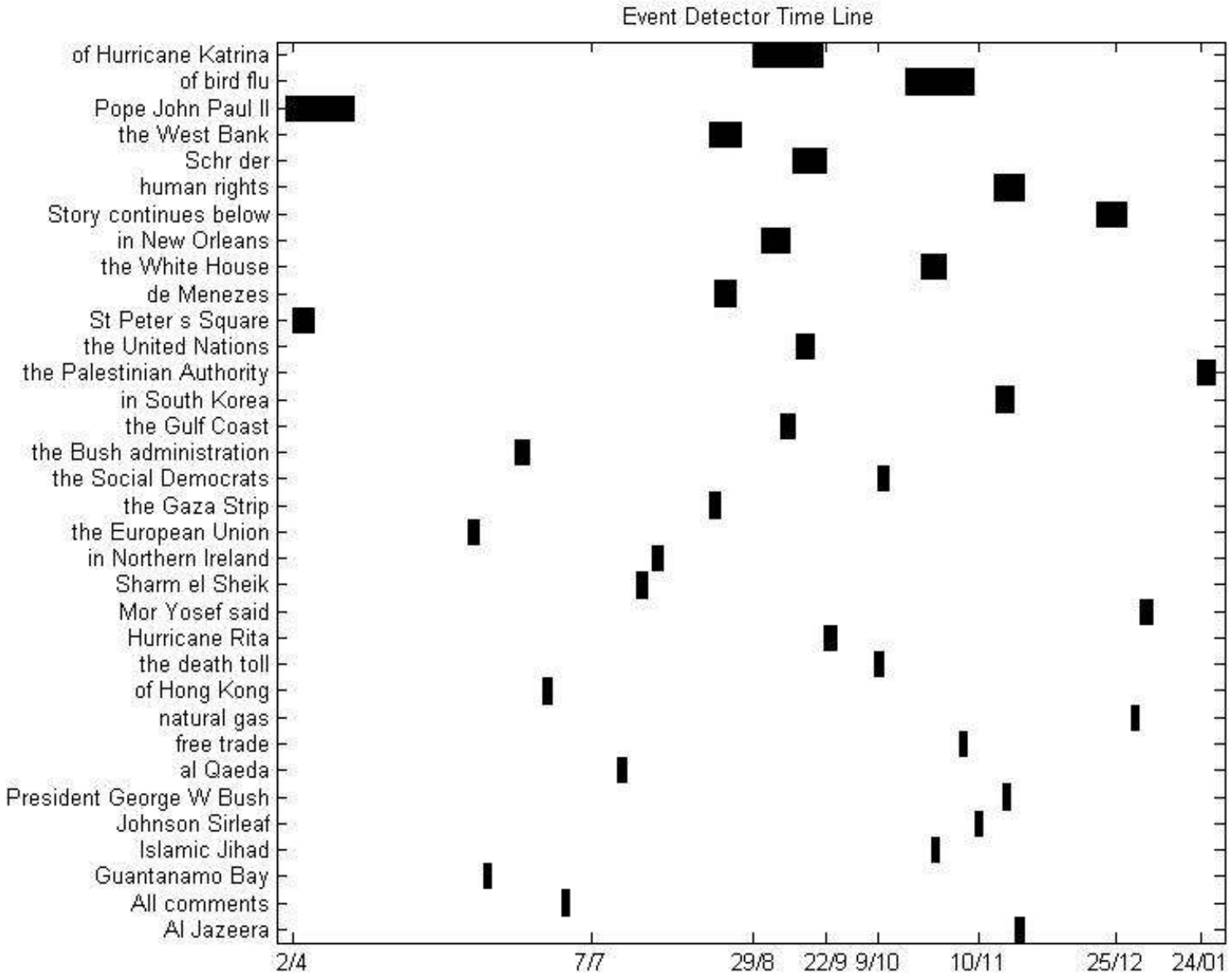
# Event Detection



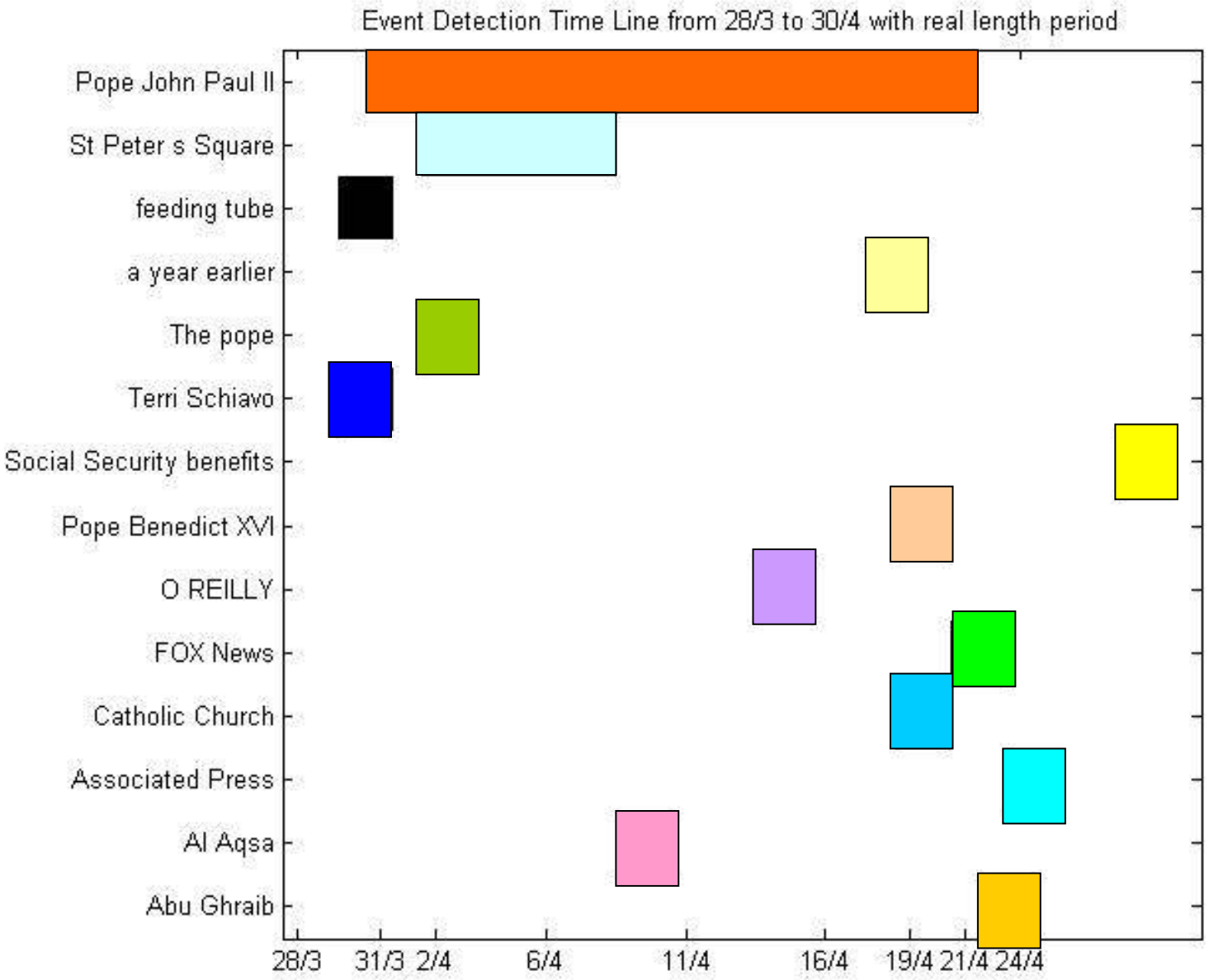
# Event Detection



# Event Detection



# Event Detection



| Fri Apr 1 2005 | Sat Apr 2 2005 | Sun Apr 3 2005 | Mon Apr 4 2005 | Tue Apr 5 2005 | Wed Apr 6 2005 | Thu Apr 7 2005 |
|---|---|---|---|---|---|---|
| Pope John Paul II | Pope John Paul II | Pope John Paul II | Pope John Paul II | Pope John Paul II | Pope John Paul II | Pope John Paul II |
| the United States | St Peter s Square | The U S military | the U S military | Prime Minister Tony Blair | Prime Minister Tony Blair | the U S military |
| the U S military | the U S military | the United States | Catholic Church | St Peter s Square | the U S military | St Peter s Square |
| North Korea s | The pope | The pope | St Peter s Square | the United States | the United States | the United States |
| Michael Schiavo | The United States | St Peter s Square | iron ore | the New York | of Hong Kong | Editor s Note |
| General Re | Navarro Valls | the U N Security Council | the United States | the U N Security Council | the New York | Prime Minister Tony Blair |
| Khmer Rouge | | | | the U S military | Navarro Valls | |
| St Peter s Square | | | | | | |

| Fri Apr 8 2005 | Sat Apr 9 2005 | Sun Apr 10 2005 | Mon Apr 11 2005 | Tue Apr 12 2005 | Wed Apr 13 2005 | Thu Apr 14 2005 |
|---|---|---|---|---|---|---|
| Pope John Paul II | the U S military | Pope John Paul II | Pope John Paul II | Pope John Paul II | long term | the United States |
| The U S military | tribunal president | the United States | the West Bank | the U S military | Parkinson s disease | the U S military |
| St Peter s Square | Associated Press | Prime Minister Tony Blair | the U N Security Council | the New York | All comments | O REILLY |
| the United States | Saturday April <N | The U S military | road map | human rights | chief executive | the New York |
| | the United States | years ago | the New York | Prime Minister Tony Blair | the New York | South Korean |
| | the U N Security Council | Al Aqsa | Prime Minister Tony Blair | the European Unior | Prime Minister Tony Blair | the U N Security Council |
| | Al Aqsa | Charles and Camilla | the United Nations | the United States | the U S military | Web site |
| | Parker Bowles | of Hong Kong | the United States | Hugo Boss | the United States | the European Union |
| | | | the U S military | the United States | the State Department | oil for food program |
| | | | India and China | security forces | Schr der | Al Jazeera |

| Fri Apr 15 2005 | Sat Apr 16 2005 | Sun Apr 17 2005 | Mon Apr 18 2005 | Tue Apr 19 2005 | Wed Apr 20 2005 | Thu Apr 21 2005 |
|---|---|---|---|---|---|---|
| Pope John Paul I | the White House | Pope John Paul II | Pope John Paul II | Pope John Paul II | Pope John Paul II | the U S military |
| the U S military | Prime Minister Tony Blair | Iraqui security forces | North Korea s | the U S military | the U S military | the United States |
| credit card | the United States | the U N Security Council | the U S military | Prime Minister Tony Blair | the New York | Pope John Paul II |
| the United States | Saturday April | around the world | the New York | Pope Benedict XVI | Catholic Church | human rights |
| Prime Minister Tony Blair | the U N Security Council | West Bank | O REILLY | the United States | North Korea s | FOX News |
| cord blood | the U S military | the White House | Associated Press | Catholic Church | President George W Bush | Prime Minister Tony Blair |
| the New York | Associated Press | Prime Minister Tony Blair | the United States | a year earlier | the United Nations | the New York |
| O REILLY | told CNN | the U S military | Prime Minister Tony Blair | St Peter s Square | Pope Benedict XVI | a year earlier |
| Lori Hacking | mass graves | the United States | a year earlier | the New York | Joseph Ratzinger | in South Korea |
| | Saddam Hussein | | St Peter s Square | around the world | cents per share | the U N Security Council |

| Fri Apr 22 2005 | Sun Apr 24 2005 | Mon Apr 25 2005 | Tue Apr 26 2005 | Wed Apr 27 2005 | Thu Apr 28 2005 | Fri Apr 29 2005 | Sat Apr 30 2005 |
|---|---|---|---|---|---|---|---|
| the United States | Pope John Paul II | North Korea s | the United States | the United States | the United States | the Middle East | the U S military |
| the U S military | the United States | the United States | the U S military | the U S military | the U S military | the D A | North Korea s |
| Abu Ghraib | the U S military | the U S military | North Korea s | al Qaeda | South Korea | the S P | the United States |
| North Korea s | al Qaeda | the New York | al Zarqawi | al Jaafari | silicone breast implants | the West Bank | Saturday April |
| Al Jazeera | Associated Press | Prime Minister Tony Blair | the New York | the New York | Social Security beneficts | blood sugar | the U S military |
| the New York | North Korea s | the Czech Republic | FOX News | President George W Bush | the White House | Social Security benefict | Prime Minister Tony Blair |
| the White House | Abu Ghraib | Associated Press | Social Security benefict | Associated Press | news conference | the United States | Los Angeles Times |
| FOX News | Prime Minister Tony Blair | South Korea | the Bush administration | sex offenders | human rights | the U S military | Wake Island |
| al Qaeda | St Peter s Square | the White House | secretary general | the State Department | Prime Minister Tony Blair | the U N Security Council | On Friday |
| H usser | the death penalty | the Bush administration | oil for food program | North Korea s | the New York | | a U S |

# Conclusions

- Given a textual time series we:
  - extract automatically the most meaningful key-phrases from new stories;
  - extract automatically the relevant events;
  - merge temporal relations with textual information;
  - monitor the world events using data mining approaches in automatic way.

Grazie per la vostra attenzione!!!!

Thanks for your attention!!!

Merci pour vôtre attention!!!