

CS 6604 Final Presentation

# Computational Linguistics PJ

-Event Extraction from  
Newswires and Twitter

*Client: Mohamed Magdy*

*by Tianyu Geng, Wei Huang, Ji Wang, and Xuan Zhang*

May, 1st, 2014  
Blacksburg, VA

# Table of Contents

*1. Introduction*

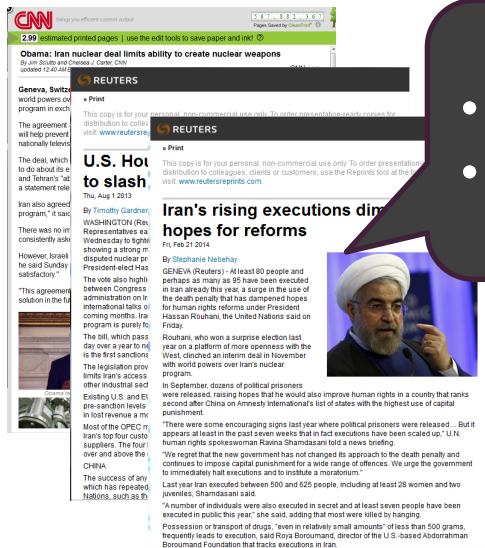
*2. Approach & Results*

*3. Discussion*

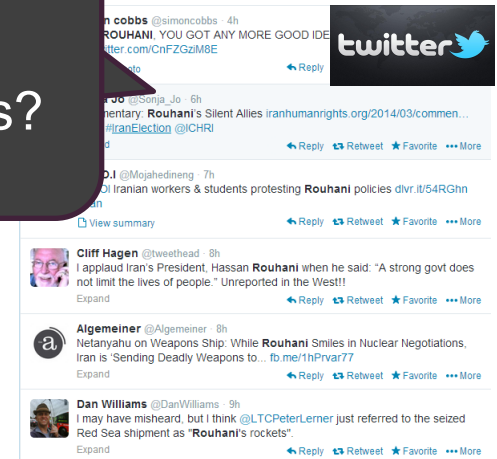
*4. Conclusion*

# 1. Problem to Solve

**Motivation:** Big data era, much news, much tweets, but...



- Key info in news & tweets?
- Relation between news & tweets?



Mainstream News

Tweets

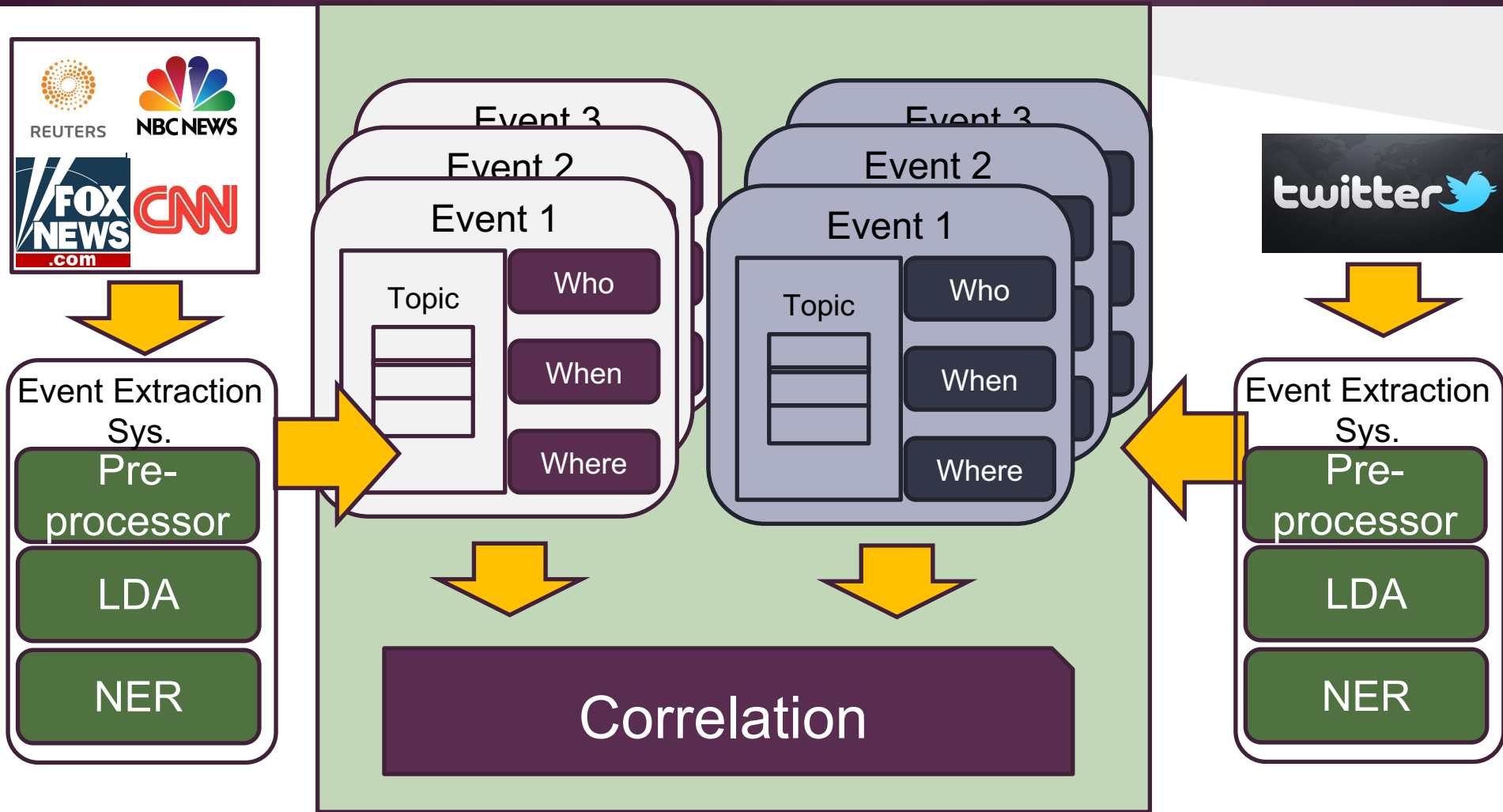
- Objects:
1. Summarize key events in news & tweets
  2. Explore correlation between news & tweets

## 2. Approach Overview

### Main Process

1. Fetch text from news & tweets respectively
2. Preprocess texts: stemming, stop-word...
3. Extract events from news and tweets  
*Event: [Topic, Named entities(who, where, when)]*
4. Link Twitter Events to News events

## 2.1 Overall Architecture



## 2.2 Data Set

### Datasets (Feb, 18th ~ Apr, 18th):

1) **4084** news about “Ukraine Crisis” from Reuters.



**4084**

N/A

N/A

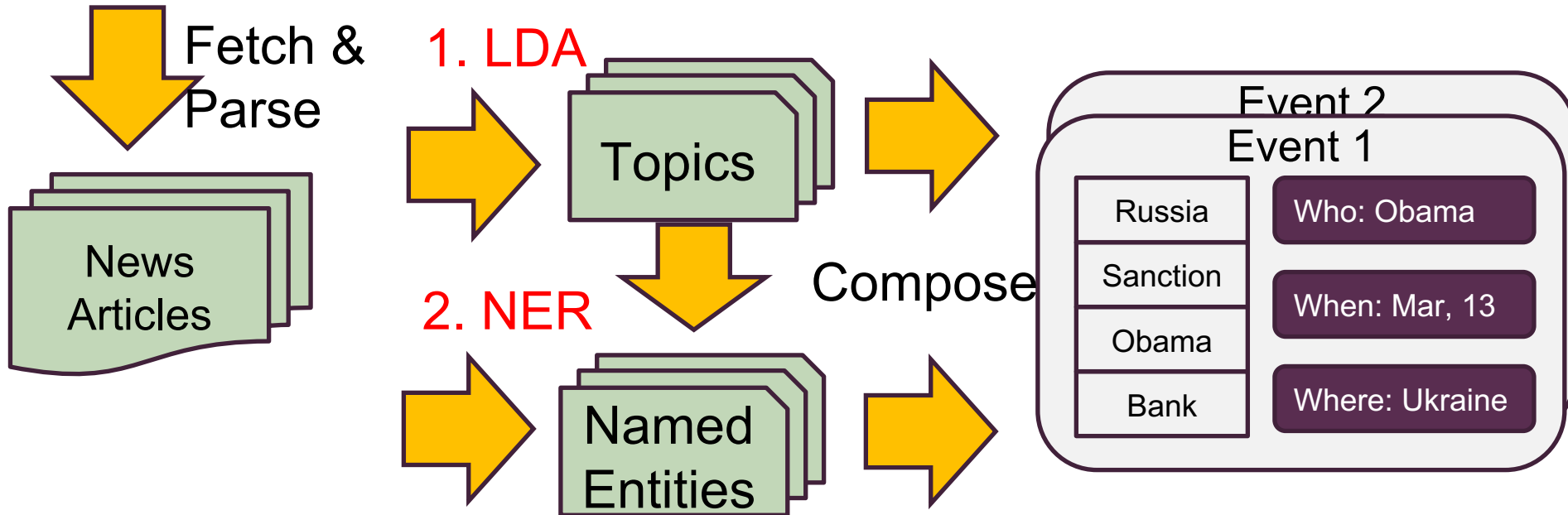
N/A

2) About **130,000** tweets about “Ukraine Crisis” from Twitter.

## 2.3.1 News Analysis Pipeline



**Event:** [**Topic**, **Named entities**(who, where, when)]



**LDA:** Latent Dirichlet Allocation

**NER:** Named Entity Recognition

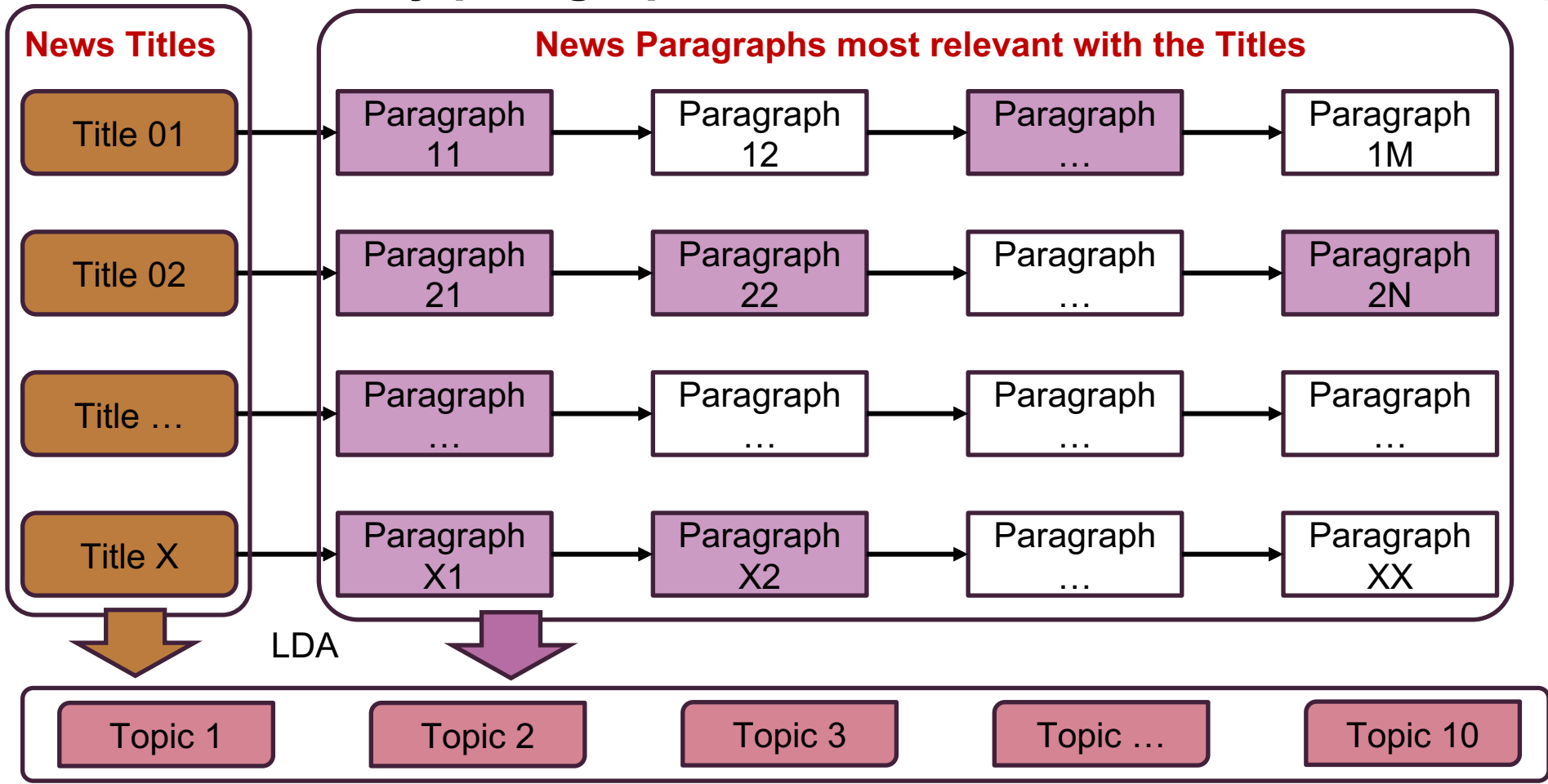
## 2.3.2 Topic Extraction (LDA)

**Source:**

- ☐ News titles
- ☐ Key paragraphs

**Tools:**

- ☐ Mallet
- ☐ JGibbLDA





# Result of Topic Modeling

## ❑ Extract topics from **news titles**

- **Strength:** titles are good summaries of the news
- **Weakness:** small data set

## ❑ Extract topics from **key paragraphs**

- **Strength:** large data set
- **Weakness:** more noise

## Our experience:

- **For news with homogeneous topics** (e.g. **Ukraine crisis news**):  
Titles are better choice
- **For news with heterogeneous topics** (e.g. Monthly news of Apple Inc.):  
Key paragraphs are better
- **Problems:**  
**Overlap** & **Noise** in about 25% topics

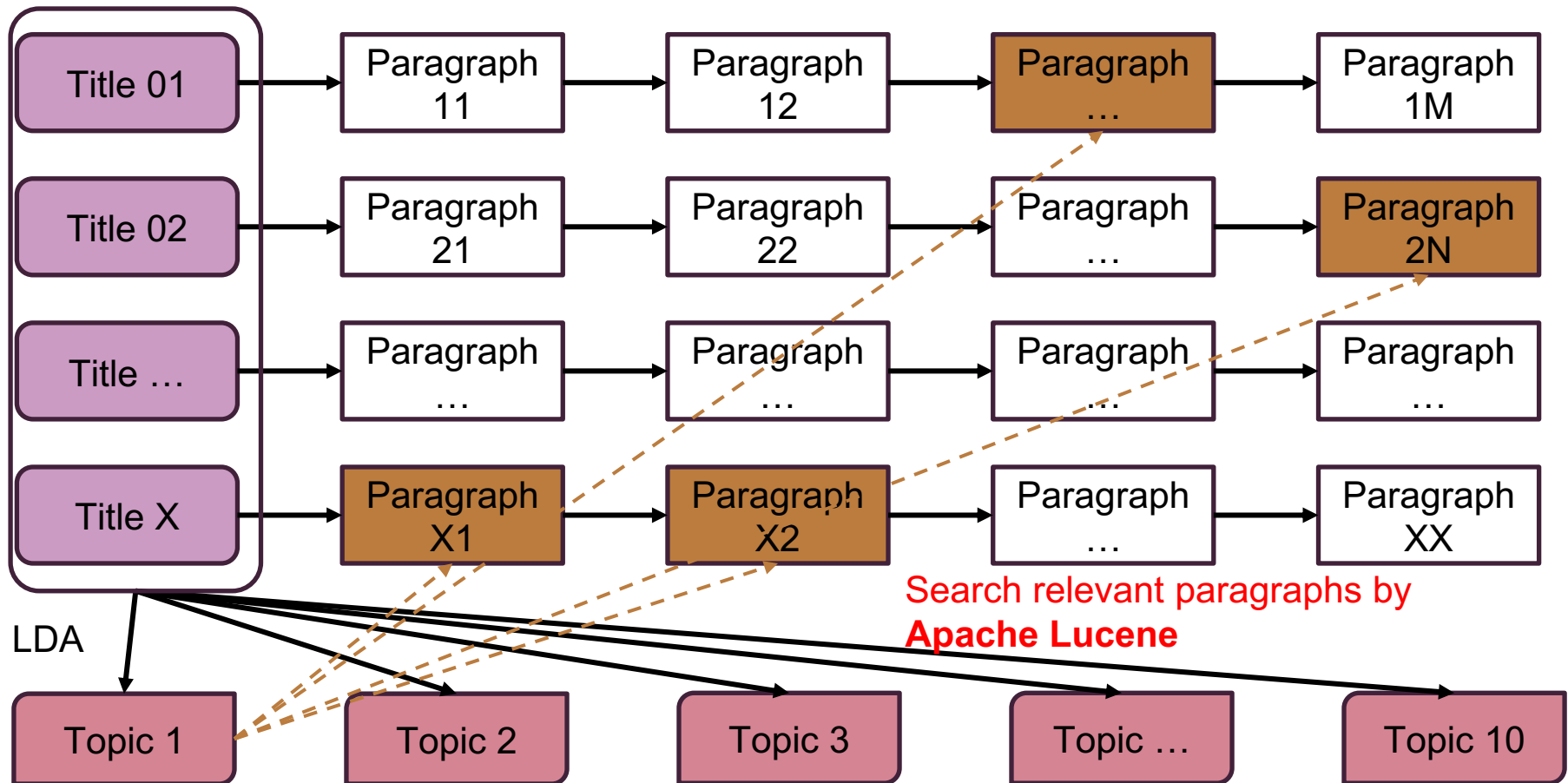
# Efforts to Improve Topic Modeling

- ❑ **Cluster the news by their titles before LDA**
  - No good result. News titles are short texts, thus the title vectors are too sparse to be clustered accurately.
- ❑ **Cluster the news by entire article before LDA**
  - No good result. There is too much noise in the news body, which deteriorates the clustering result.
- ❑ **Split the entire dataset into datasets with shorter periods**
  - Topics with finer-granularity obtained.
  - However, more noise emerged compared to topics from the entire data set.

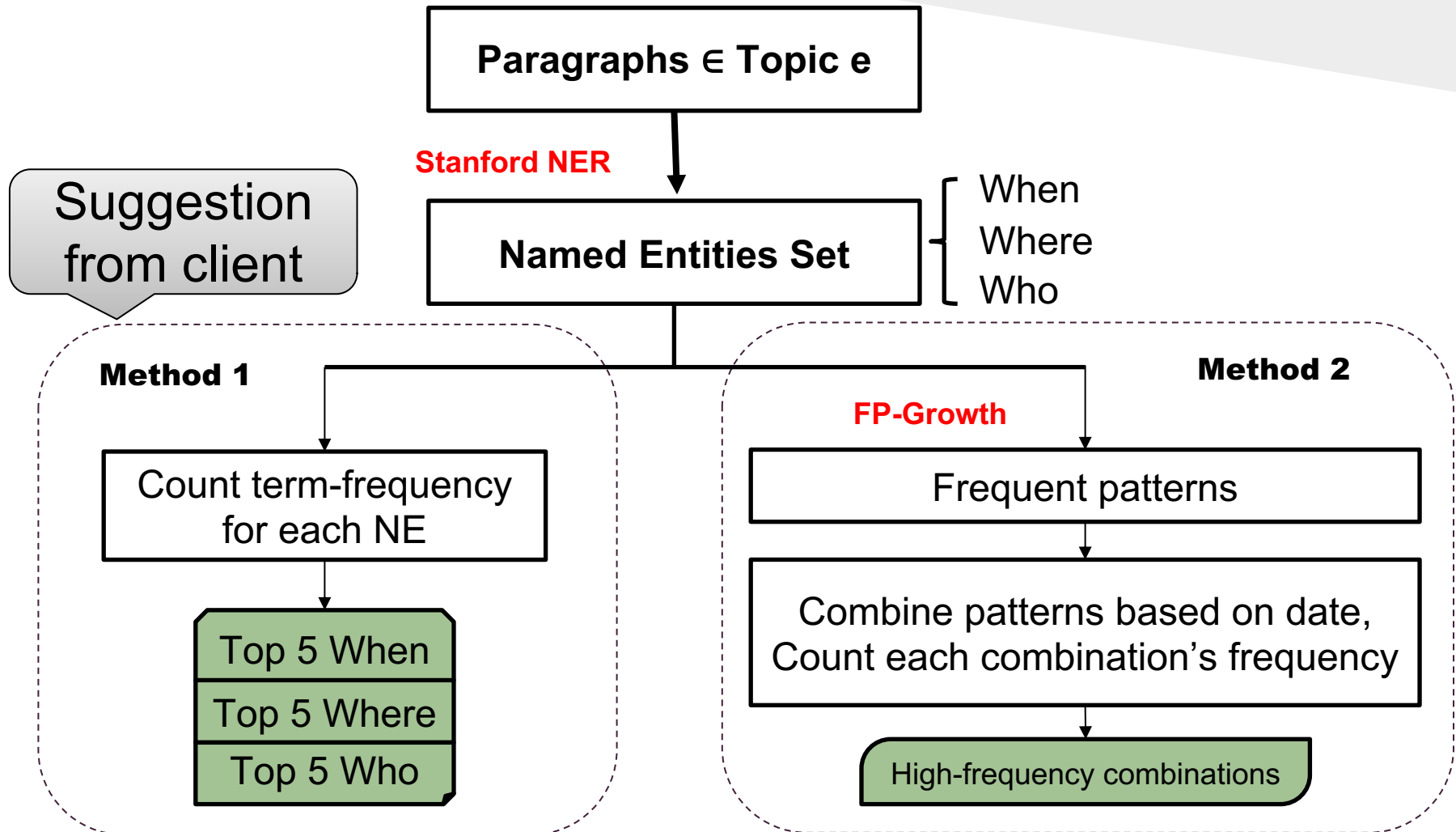
Suggestions from Dr.Fox

## 2.3.3 Named Entity Extraction

- ❑ Hard to find Named Entities (WHO,WHERE,WHEN) in topics.
- ❑ We need to search NEs in relevant news paragraphs



# Extract Named Entities - Methods



# Extract Named Entities - Results

*Topic: [crimea, ukraine, russian, troops, border]*

Method 1: High-frequency 3W named entities

WHO: NATO; Oleksander Turchinov; Kerry; Lavrov; Vladimir Putin;

WHEN: Mar 15 , 2014; Thursday; Apr 16 , 2014; Mar 3 , 2014; Mar 24 , 2014;

WHERE: Ukraine; Crimea; Russia; U.S.; Kiev;

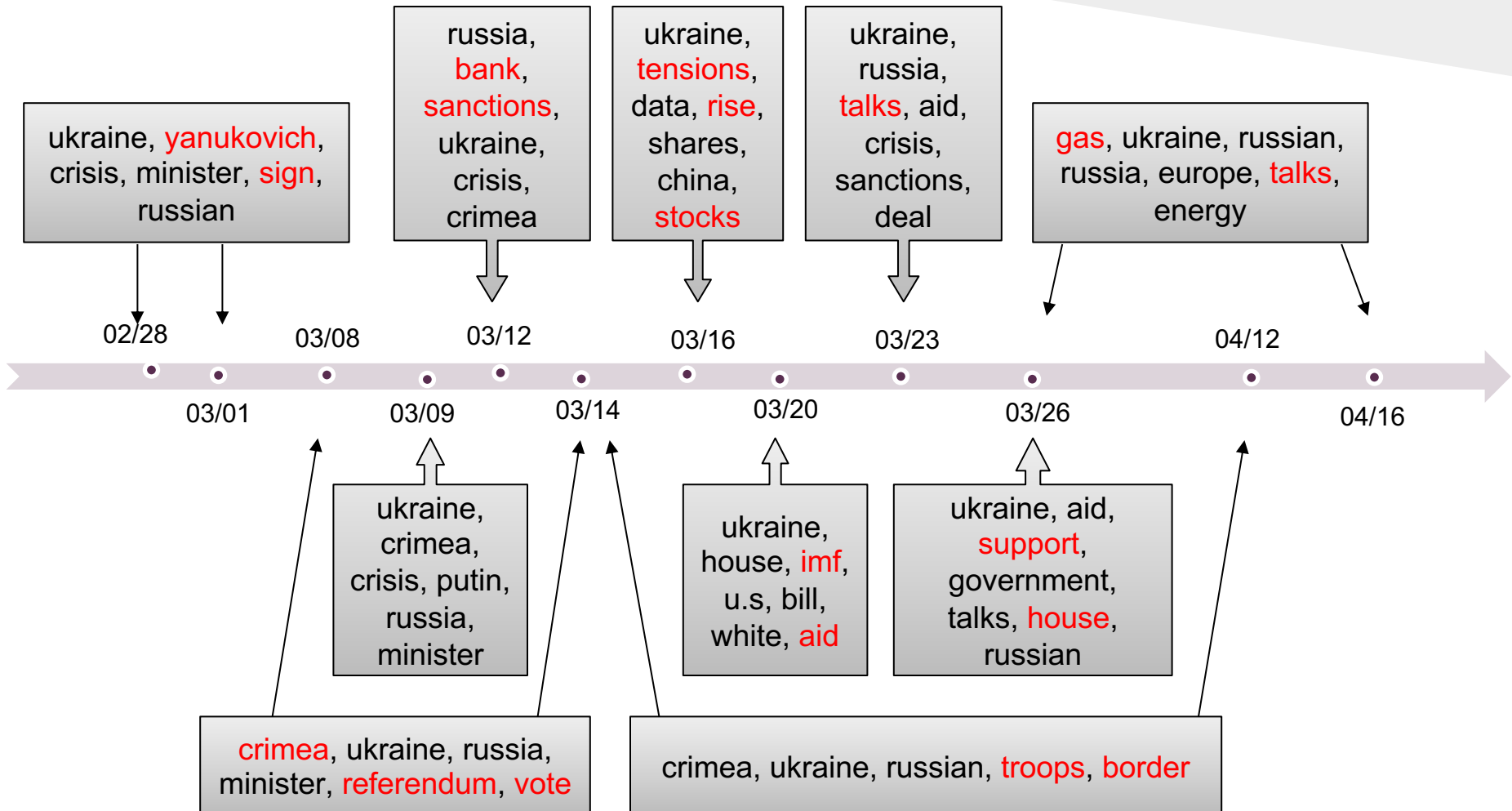
Method 2: High-frequency named entities combinations

[Mar 15 , 2014; Ukraine; Crimea; Donetsk; Kharkiv; Arbatskaya Strelka; Oleksander Turchinov]

[Mar 29 , 2014; Russia; Ukraine; Crimea; Lavrov; Vladimir Putin]

[Apr 12 , 2014; Russia; Moscow; Ukraine; Crimea; NATO]

# Extracted Events on a Time Line



# Extracted Events Sample

**2014/03/08 - 2014/03/14;**

**Topic:** [crimea, ukraine, russia, minister, referendum, ukrainian, vote]

**WHO:** U.N. Security Council; Arseny Yatseniuk; Vladimir Kirichenko; Obama; Putin;

**WHERE:** Russia; Crimea; Ukraine; Kiev; China;

**Combination:** [Mar 9 , 2014; Russia; Ukraine; Crimea; United States; Vladimir Kirichenko; Obama]

**Combination:** [Mar 14 , 2014; London; Russia; Ukraine; West; Crimea; Moscow; Arseny Yatseniuk; Kerry; Russian Federation]

**2014/03/20 - 2014/03/21;**

**Topic:** [ukraine, house, imf, u.s, bill, white, aid]

**WHO:** IMF; Senate; White House; House of Representatives

**WHERE:** Ukraine; WASHINGTON; Kiev; United States;

**Combination:** [Mar 21 , 2014; Ukraine; U.S.; New York; Senate; Royce; House Foreign Affairs Democrat; Nita Lowey; Eliot Engel; House Appropriations Committee; IMF]

## 2.4.1 Approaches

- Tweets are from IDEAL collection.
- Assign topics to each tweet by LDA.
- Apply Method 2 (FP-Growth) in the NER step of tweets analysis.



## 2.4.2 Results

Who; Where

News Facts:

Feb 18: The initial riots began

Feb 20: Ukraine Government Snipers Shooting protesters in Kiev

Topic 1

live, snipers, protests, control, here, watch, video

European Union; Ukraine

European Union; Ukraine

Hotel Ukraina; Kiev

Peter Brookes; Kiev

EU; Rome

02/18

02/20

02/21

02/22

03/06

Yanukovich; Kiev

Topic 2

today, president, storm, backed, threaten, forces, shooting

News Facts:

Feb 20/21: President Yanukovich signed a compromise deal with opposition leaders. Then, he left Ukraine.

# Tweets Events Samples

- Topic 1
  - Keywords: live, snipers, protests, control, here, watch, video
  - Events:
    - {Feb 18, 2014; European Union; Ukraine}
    - {Feb 20, 2014; EU; Ukraine}
    - {Feb 20, 2014; Hotel Ukraina; Kiev}
    - {Feb 22, 2014; Peter Brookes; Kiev}
    - {Mar 06, 2014; EU; Rome}
- Topic 2
  - Keywords: today, president, storm, backed, threaten, forces, shooting
  - Events:
    - {Feb 20, 2014; Yanukovich; Kiev}

# 2.5 Correlation between Twitter Events & News Events

## News Event

Term 1	Term 2	Term 3	...	Term 7	WHO	WHERE	WHEN
--------	--------	--------	-----	--------	-----	-------	------

## Twitter Event

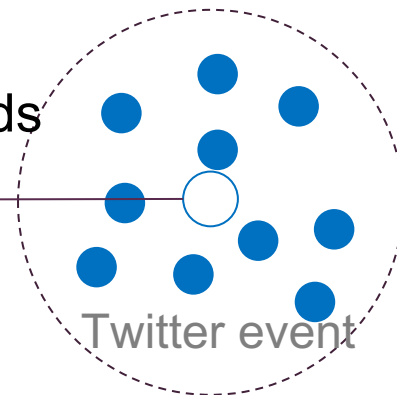
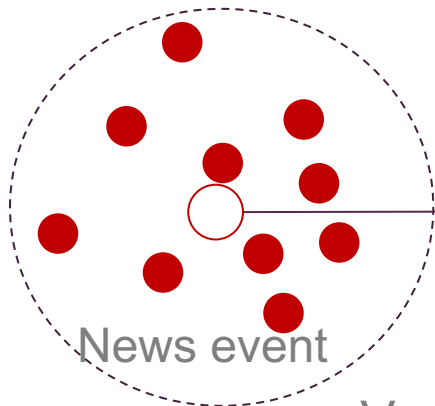
Term 1	Term 2	Term 3	...	Term 7	WHO	WHERE	WHEN
--------	--------	--------	-----	--------	-----	-------	------



Word2Vec (by Google)

### Event Similarity:

Distance between centroids



Vectors in 200-Dimension hyper space

# 3.1 Issues & Lessons

## --- News Analysis

### **Open Issues:**

1. Overlap and noise in the extracted topics
2. Noise in the extracted named entities
3. Similarity model to link the Twitter events to news events

### **Lessons:**

1. Collect more data from other news websites
2. Remove overlap in topics by splitting the data set
3. Remove duplicates from the frequent NE combinations

## 3.2 Issues, & Lessons

### --- Tweets Analysis

1. There are very big noise in tweets themselves.
2. It's not easy to extract Named Entities from tweets.

#### **Lessons:**

1. topic model tweets via LDA (Python Gensim)
2. extract name entities from tweets. (PyNER)
3. use FP-Growth algorithm to pick the most high frequency keywords combination in event description.

## 3.3 Potential Usages

1. A tool for event extraction and news summarization
2. A tool for the “Computational Linguistic” course
3. A component for the IDEAL project
4. A tool to extract pure text from archived web pages

# 4.1 Conclusion

- Developed an effective tool for web page event extraction
- Explored various methods regarding every step of the event extraction
- More efforts are needed to link the tweets events to news events

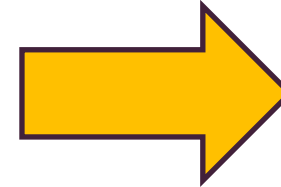
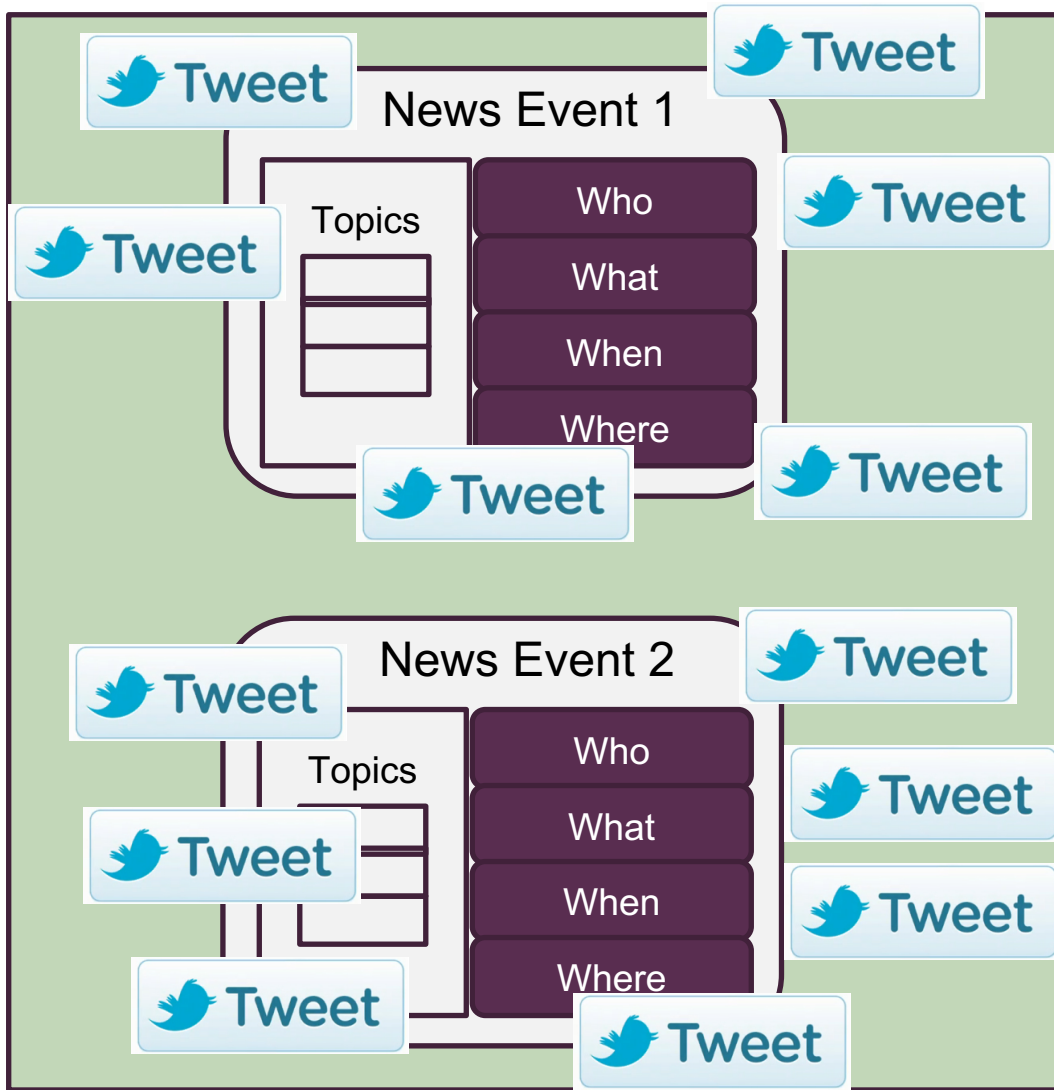
## 4.2 Structure of Deliverables

1. Presentation slides
2. A project report (including tool manual)
3. Source code
  - News event extractor (in Java)
  - Twitter event extractor (in Python)
4. Data
  - Text version of news dataset
  - Output results

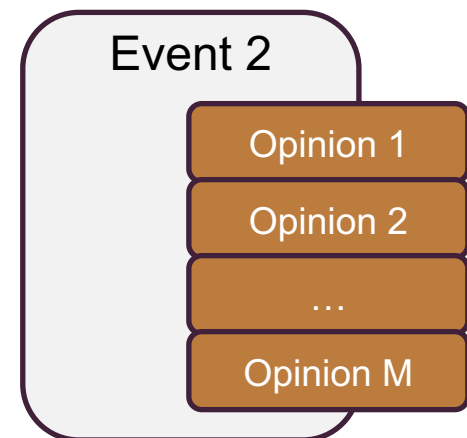
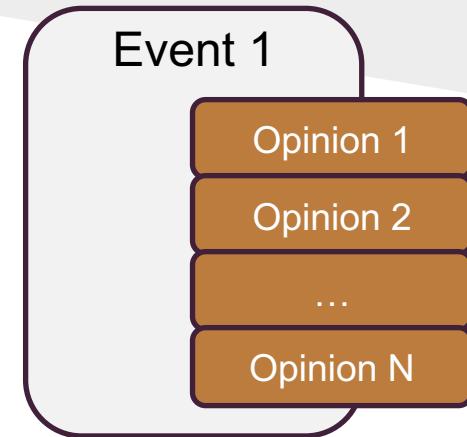
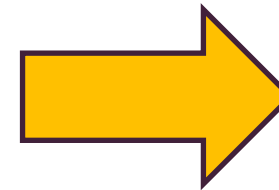


# *Appendix*

## 4.3 Future Work: Opinion Mining



Opinion  
Mining



# Topics from Clustered News Titles

## **Cluster 1:**

Ukraine Russia say eastern update Russian hit  
Ukraine Putin say leader WRAPUP Crimea send  
Ukraine update crisis say tension police military  
Crimea russian from order after lawmaker official

## **Cluster 2:**

U.S. IMF gas aid bill reform Kerry  
Ukraine EU help ukrainian gas crisis export

## **Cluster 3:**

update Russia may Germany trade Merkel government  
Ukraine talk House bank White german aid  
sanction russian EU against Obama over energy  
Russia EU Ukraine sanction Obama war agree

# Topics from Clustered News Bodies

## **Cluster 1:**

Russia Ukraine EU warn Moscow Crimea sanction  
Russia Ukraine take call say Putin separatist  
Ukraine after force U.N. putin vote fear

## **Cluster 2:**

Ukraine after eye over rouble import gas  
Crimea emerge bond Bank bank market more  
russian say Ukraine gas hit may ukrainian

## **Cluster 3:**

Putin Ukraine Obama call discuss Merkel White  
Ukraine update against urge aid after leader  
Ukraine minister Yanukovich gas Poland president polish  
Ukraine force NATO pm security seize gas

# Topics from Key Paragraphs

Topic [0]: [ukraine, reuters, Ukraine, russia, crisis, foreign, president]

Topic [1]: [reuters, kiev, yanukovich, president, viktor, ukrainian, ukraine's]

Topic [2]: [military, nato, reuters, ukraine, u.s, russia, crimea]

Topic [3]: [sanctions, crimea, russia, moscow, reuters, russian, referendum]

Topic [4]: [russian, reuters, crimea, ukraine, ukrainian, forces, military]

Topic [5]: [gas, ukraine, reuters, russia, russian, energy, moscow]

Topic [6]: [percent, ukraine, reuters, march, Ukraine, tensions, u.s]

Topic [7]: [ukraine, reuters, aid, billion, Ukraine, washington, international]

Topic [8]: [russia, putin, ukraine, russian, vladimir, war, president]

Topic [9]: [rating, fitch, ratings, bank, banks, currency, ukraine]

# Topics from Splitted Data Sets

## **Data set 1:**

Ukraine, Kiev, protesters, police, team, Games, square  
Yanukovich, Ukraine, opposition, crisis, talks, deal, Ukraine's  
Hryvnia, bank, assets, record, low, foreign, gains

## **Data set 2:**

Ukraine, U.S, Russia's, war, discuss, crisis, Merkel  
Crimea, Ukraine, Russia, Putin, force, back, troops  
Ukraine, tensions, China, rise, stocks, tension, ease

## **Data set 3:**

Ukraine, Russia, IMF, aid, crisis, talks, deal  
Russia, Crimea, military, Crimea's, vote, Moscow, U.N  
Ukraine, Russia, IMF, aid, crisis, talks, deal

## **Data set 3:**

Russian, Ukraine, Ukraine's, military, embassy, agency, suspected  
gas, Ukraine, Russia, talks, Europe, supply, debt  
Ukraine, Putin, Russia, U.S, data, call, House

# Progress: Text Extraction

