Analysing Wikileaks Documents with SOMs

Rudolf Mayer, Andreas Rauber
Vienna University of Technology
Information & Software Engineering Group
Outline

- Motivation
- Document Collection
- SOM Methods applied
- Cabledgate Map
- Conclusions and future work
Motivation

- Wikileaks cablegate collection caused a lot of frenzy
- Collection seemed interesting
  - Previously unpublished articles
  - Many different authors
  - Many diverse topics

- Many different websites offering access to collection
  - Wikileaks itself, newspapers such as The
Data Collection

- Cables (messages) from US Embassies to US state department
  - Covering political / economical relevant aspects
  - Summaries of happenings in the respective countries
  - Personal comments on people / situations

- Total of 251,287 documents
  - From 1960s to 2010; very few before 2000
  - 100,000 labelled confidential, 15,000 secret

- First 220 documents released on Nov 28\textsuperscript{th} 2010
  - Plan to publish around 80 per day
<table>
<thead>
<tr>
<th>Date</th>
<th># Docs</th>
<th># new / day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nov 28, 2010</td>
<td>220</td>
<td></td>
</tr>
<tr>
<td>Jan 8, 2011</td>
<td>2,017</td>
<td>44</td>
</tr>
<tr>
<td>Jan 17, 2011</td>
<td>2,497</td>
<td>53</td>
</tr>
<tr>
<td>Jan 21, 2011</td>
<td>2,628</td>
<td>33</td>
</tr>
<tr>
<td>Jan 23, 2011</td>
<td>2,658</td>
<td>15</td>
</tr>
<tr>
<td>Jan 27, 2011</td>
<td>2,722</td>
<td>16</td>
</tr>
<tr>
<td>Feb 5, 2011</td>
<td>3,373</td>
<td>72</td>
</tr>
<tr>
<td>Jun 13, 2011</td>
<td>14,717</td>
<td>89</td>
</tr>
</tbody>
</table>
S E C R E T  SECTION 01 OF 03 KABUL 001767

SIPDIS

R.O. 12958: DECIL: 07/03/2019
TAGS: PGOV PREL AF
SUBJECT: KARZAI ON THE STATE OF US-AFGHAN RELATIONS

Classified By: Ambassador Karl W. Eikenberry for reasons 1.4 (b) and (d).

%1. (S/NF) SUMMARY. In a series of meetings this week, I discussed with President Karzai a number of issues, including Karzai's concerns with the future of US policy on Afghanistan, and his opinions on where US policy here had failed. The meetings were cordial, but raised certain concerns for me about the status of the US-Afghan relationship, both in the lead up to the election and over the long term, should Karzai win reelection. Karzai's comments reinforce my belief that our relationship must be a two-way partnership of obligations and responsibilities. We must also convince Karzai to put his backing behind democratic institutions and professionalized security forces are better equipped to lead Afghanistan into the future, rather than Karzai's preference for tribal structures and informal power networks. This cable summarizes the content of these four recent meetings, as well as a subsequent conversation with Interior Minister Atmar. End Summary.

%2. (S/NF) I met with Karzai on four occasions this week: on June 30 with COMISAF Gen. McChrystal and most of Karzai's cabinet, on July 1 with the first vice president and the National Assembly, on July 2 with the second vice president, and on July 3 with the Kalonji I. (Cable 09KABUL1767#1 for context.) The president and his team were in a good mood despite their defeat in the recent elections. Karzai and his team are preparing to hand over power to a new government and have begun “redistributing the deck.”

The president and his team are acutely aware of the challenges ahead. However, he seemed to have a positive outlook that the next government will work with him toward national unity and stability.

I met with the president on July 3 to discuss our conversation on July 1. He repeated his comments that the US can help阿富汗 to become a normal country. He said that he and his team are preparing the country for the transition to the next government, and they did not expect the current government to remain in power.

I let the president know that the US continues to support his government and that it is committed to a long-term relationship. The president appreciated our support and said that he looks forward to continuing our partnership.

We discussed the recent elections and the challenges facing the next government. The president acknowledged the challenges and said that he will focus on building a strong and capable team. He also expressed his confidence in the new government's ability to handle the challenges.

I expressed my support for the president and his team and assured them of our commitment to a long-term partnership. The president appreciated our support and said that he looks forward to continuing our dialogue.

I also discussed with the president the importance of security forces and the need for them to be professionalized. The president agreed with me and said that he will continue to work with us on this issue.

In conclusion, I believe that our relationship with Afghanistan is strong and that we can continue to work together to achieve our common goals. The president expressed his appreciation for our support and assured me of his commitment to our partnership. I look forward to continuing our dialogue and working towards a brighter future for Afghanistan.

(End S/m)
Atmar on Karzai's paranoia

9. (S/NF) I told Interior Minister Atmar at our July 2 meeting that with his conspiratorial behavior, Karzai would run the risk of leaving USG interlocutors with the impression that we have accomplished very little here and that the Afghan government believed most of the failures lay with us. This is not a dialogue that will lead to an effective partnership. The US has been clear in its past shortcomings in Afghanistan, but we have yet to see Karzai admit to the serious shortcomings in his administration.

10. (S/NF) Atmar agreed with the importance of Karzai presenting himself as a better international partner. He also detailed three existing paranoia's that affected Karzai's worldview: (1) Karzai did not understand US policy in the region and suspected ulterior motives in our relationship with neighboring countries; (2) he suspected the US was contemplating a short-term strategy in Afghanistan that would result in our disengagement within the next two years; and (3) the US was intent on dumping Karzai and supporting another candidate. A fourth emerging paranoia was of a US plot to divide and weaken Pashtuns on both sides of the Afghan-Pakistani border. Despite these paranoia's, Atmar assured me that Karzai had confidence in his personal relationship with me and recognized the need to meet more frequently to work through his misperceptions of US policy and intentions.

Our Relationship with Karzai

11. (S/NF) In these meetings and other recent encounters with Karzai, two contrasting portraits emerge. The first is of a paranoid and weak individual unfamiliar with the basics of nation building and overly self-conscious that his time in the spotlight of glowing reviews from the international community has passed. The other is that of an ever-shrewd politician who sees himself as a nationalist here who can...
4. C) In luncheon discussions, Berlusconi unashamedly stated that he prized Italy’s relationship with the U.S. and that his government stood ready to help us, whatever the issue. He noted that Italy owed the United States a debt of gratitude for salvation during World War II, and for protection throughout the Cold War. On Afghanistan, bashing issues, and other tough problems, Italy was committed to helping the U.S. get to the right solution. He noted that he hoped that the Italian increase in Afghanistan would help President Obama and address the situation on the ground.

Berlusconi on European Leaders
---------------------------------

5. C) The Prime Minister and Letta also gave their run down of current European political leadership. Berlusconi assessed that Sarkozy’s star was clearly waning in European circles and that the French President did not command the same influence he did a year ago. Letta was less convinced, noting to the Ambassador that both Berlusconi and Sarkozy were “big dogs angling for the same spotlight.” Berlusconi recounted how he had played an instrumental role in the Spring in persuading a reluctant Erdogan to accept Danish PM Rasmussen as the new NATO Secretary General, overcoming the Turkish President’s profound irritation at Rasmussen for not silencing Kurdish language radio stations despite numerous requests. On Russia, Berlusconi felt that Putin and Medvedev got along well, respected each other, and had an effective relationship. In fact, farewells after lunch were unexpectedly cut short by an incoming call from Putin.

Iran: Appalled by violence, Will try to rope in ENI
---------------------------------

6. C) On Iran, Berlusconi noted that he was appalled by the Iranian crackdown. He offered the capabilities of the Italian intelligence services and assured full sharing of information gathered on the internal dynamics in Tehran. Letta
opposition, and experience and intuition told xxxxx that his study would fall on deaf ears, a prediction already supported by the empty auditorium at his press conference hours before our meeting. Russian Communist Party (KPRF) leader Gennadiy Zuganov publicly and privately praised xxxxx for his research as a boost to KPRF calls for greater press freedom. Medvedev’s sagging standing in the polls, despite lavish T.V. coverage, is one indication of his difficulty in coming out from Putin's shadow. End Summary.

Medvedev Main Headliner, but Putin Top Dog

2. (C) On August 4, xxxxx told PolOff that while Medvedev received twice as much press coverage as Putin during the testing period, most of his meetings were largely ceremonial, while Putin’s on-air performances continued to give the impression that he held the reins of power. xxxxx unabashedly called Medvedev a "no one" in Russian politics, claiming he was nothing more than a "pen in one's pocket" or "a glass on the table" as he raised a pen and glass from his desk. Medvedev continued to attract interest with his anti-corruption campaign, but Putin still called the shots that mattered.

3. (C) An August 7 poll from Levada Center indicated only 9% of Russians believe Medvedev controlled Russia's direction, while 36% believed Putin remained in sole power and almost 50% saw a tandem presidency. A related Levada Center poll from July 18-22 showed that 82% of Russians believed, to some degree, that Medvedev merely continued Putin's policies.
Data Collection

- Retrievable as full package via torrents
  - December/January: 2-3 snapshots per day
    - retrievable in a few minutes
  - Currently: latest snapshot that can be found dates from mid-march
    - torrents not seeding

- Frenzy has ended ...
  - Both in media and on the net
Data Collection

- Documents in HTML format; pre-processing to
  - Remove HTML markup
  - Remove metadata, such as date, origin, classified state, author, … from documents
  - Store metadata for later analysis of categories on the map
Feature Extraction

- Basic bag-of-words indexing approach
  - Each word is a token
- Vector space model
  - Each document is a vector
- Vectors computed by TFxIDF weighting
  - Each token = element of the vector
  - Vector values: frequency in document (TF), weighted by frequency in whole collection (DF)
Feature Extraction

- Total of 65,000 tokens
- Not each token is relevant
  - Also computational aspects
- Stopword removal
- Feature selection: frequency thresholding
  - Remove tokens that are
    - Too frequent: cannot distinguish between documents
    - Too infrequent: not relevant
- Finally 5,500 tokens remaining
SOM Methods

- **Java SOMToolbox**
  - Different visualisations, clustering, labels and interaction

- **LabelSOM**
  - Automatically compute labels for nodes/clusters
  - Based on weight vector values of each unit
  - Select those terms (components) that have a high value, and low in the other vectors (a bit like TFxIDF)
SOM Methods

- Class visualisation to visualise distribution of category labels (e.g. origin)

- Map clustering to obtain clear cluster boundaries & cluster labels (Ward’s linkage)
First Results

Colours indicate origin label
First Results
Feature Extraction

- Map separates documents too well
- Reason: too many country-specific terms
  - i.e. cables from a specific country mention the capital and country name too often
- Map not much different than category-based browsing ... !
- Manual processing of document-specific stop words
  - remove country-specific terms & specific abbreviations
Cablegate Map – Cable Origin
Online Version

- Online version with limited functionality at www.ifs.tuwien.ac.at/dm/wikileaks
  - Currently with around 12,000 documents
    - (In)frequently updated
Challenges & Future Work

- Legal aspects
  - Wikileaks created huge havoc in the US

- Online Version only links to documents on Wikileaks.org, doesn’t redistribute documents
Challenges & Future Work

- Data set is continuously growing
  - Approximately 2,500 new documents per month

- Need to update map frequently
  - Easy to map a few new documents
  - When is there a need to retrain the map?
  - When to redo the feature extraction?

- How to not change the map too much for the user?
Thanks for your attention!

Questions?