Analysis of Wikileaks Cables Using NLP Techniques

CS671:Natural Language Processing
A Project Proposal
October 21, 2013

Sugam Anand *sugambh@iitk.ac.in* 10730
Arpit Jain *arpitj@iitk.ac.in* 10145

Introduction And Motivation

Wikileaks embassy cables revelations covered a huge dataset of official documents counting around 251,287 , from more than 250 worldwide US embassies and consulates. It's a unique picture of US diplomatic language. The cables show the extent of US spying on its allies and the UN; turning a blind eye to corruption and human rights abuse in "client states"; backroom deals with supposedly neutral countries; lobbying for US corporations; and the measures US diplomats take to advance those who have access to them.Such a huge,rich and structured dataset has not been analyzed with natural language and Information retrieval techniques.

Objective

Knowledge of the problem domain intuitively suggests that the US embassies and Secretary of State in Washington communicated regarding some topic of their interests referencing certain people ,places and organization.These communication have been from 28th December 1966 to 28th February 2010.We wish to follow an unsupervised or semi-supervised approach for extracting these latent topics and entities and mapping these to timeline[1].We wish to provide visualization that could provide some insights into these cables.Our Second goal is to do extract semantic relations and sentiments from these cable.For this project we want to do this analysis for New Delhi, Secretary of State Washington DC , Baghdad , Islamabad ,Paris and Moscow.

Methodolgy

Currently we have dataset containing 227630 cables from 1966 to 2009 containing important diplomatic information.Each cable is well structured and comprise of:

Source : Embassy which sent the cable:
Destination : Target Embassies
Date : Sending date
Body : Containing the raw text
Tags : Containing meta information regarding cable like classified,unclassified or secret etc.

Extracting Hidden and Semantic Analysis:
Documents can be represented as numeric vectors in the space of words.The order of words is lost but the co-occurrences of words may still provide useful insights about the topical content of a collection of documents.PLSA [2] is an unsupervised method based on this idea.We seek to use it to find out what topics are there in a collection of documents.It is also a good basis for information retrieval systems.
Also , a Named Entity recognition can provide the possible places,persons and organisations being talked about.

A POS Tagger can give the list of adjectives in the cables , further sentiment analysis on that would provide good insight on how sentiments are varying on different topics over the period of time.
As goal of the project we want to find topics and the sentiments associated with them for the US and map that to the timeline.

Dataset

1.6 GB dataset released from wikileaks

Tools

- Gensim
- NLTK
- Stanford CoreNLP

Reference

[1] O'Connor, Brendan, Brandon M. Stewart, and Noah A. Smith. "Learning to Extract International Relations from Political Context." *Association of Computational Linguistics*. 2013.
[2] Hofmann, Thomas. "Probabilistic latent semantic indexing." *Proceedings of the 22nd annual international ACM 3SIGIR conference on Research and development in information retrieval*. ACM, 1999.