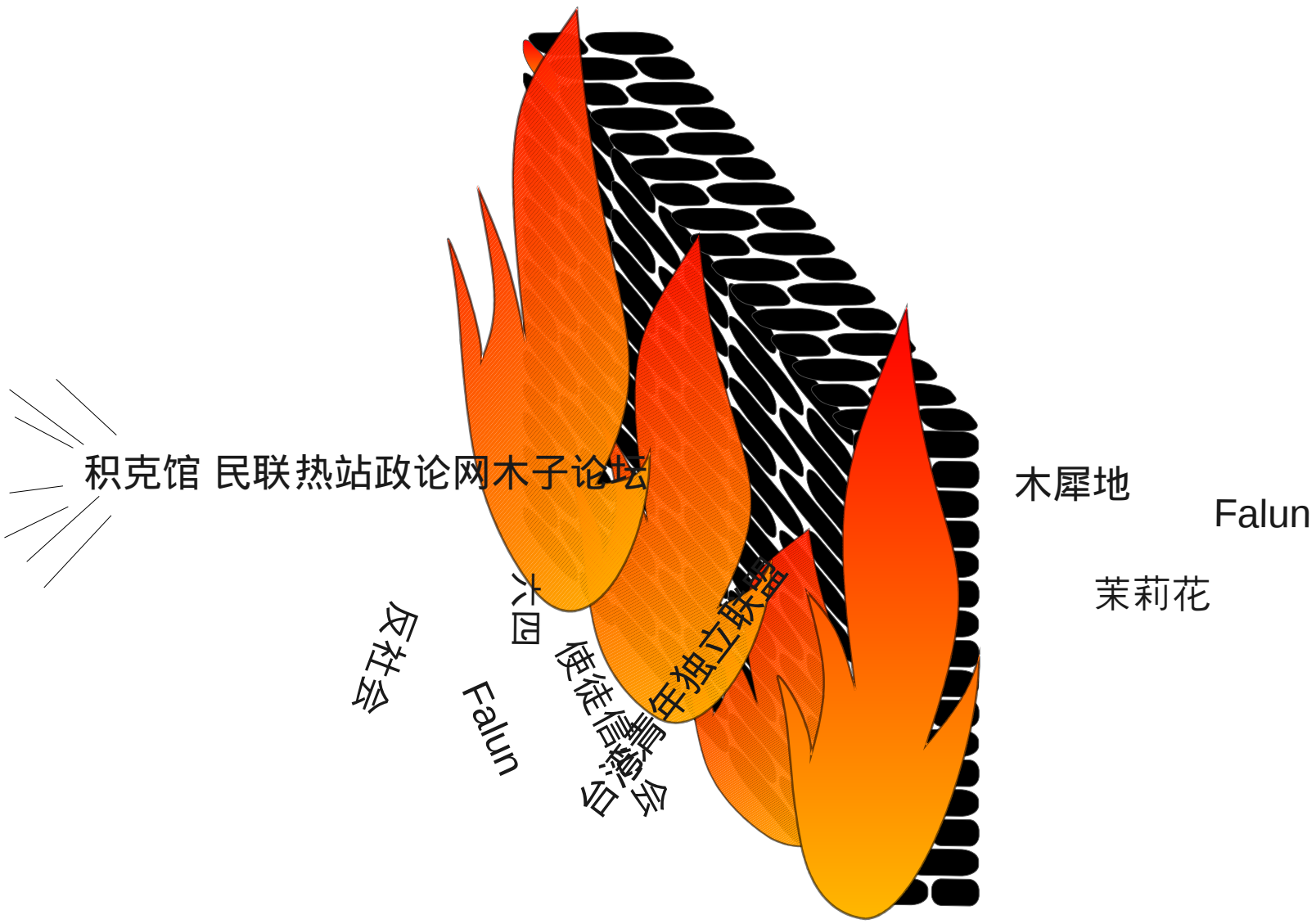
Automated Named Entity Extraction for Tracking Censorship of Current Events

Antonio M. Espinoza & Jedidiah R. Crandall
Computer Science Department
University of New Mexico

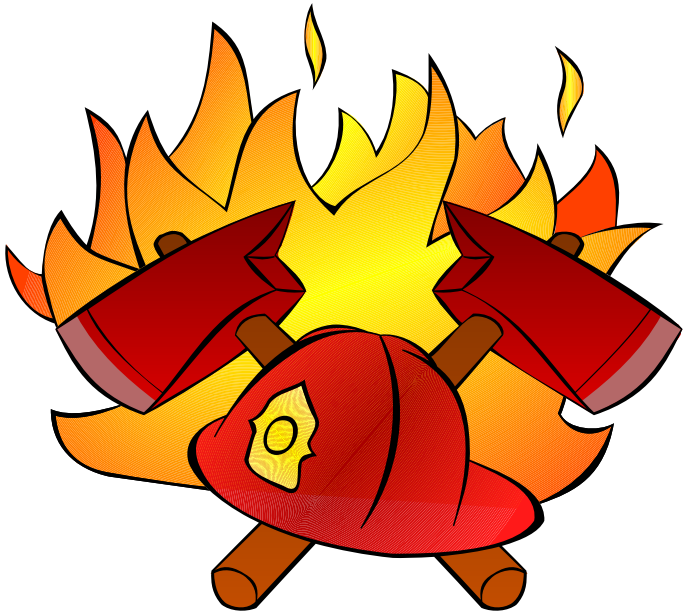
{amajest,crandall}@cs.unm.edu

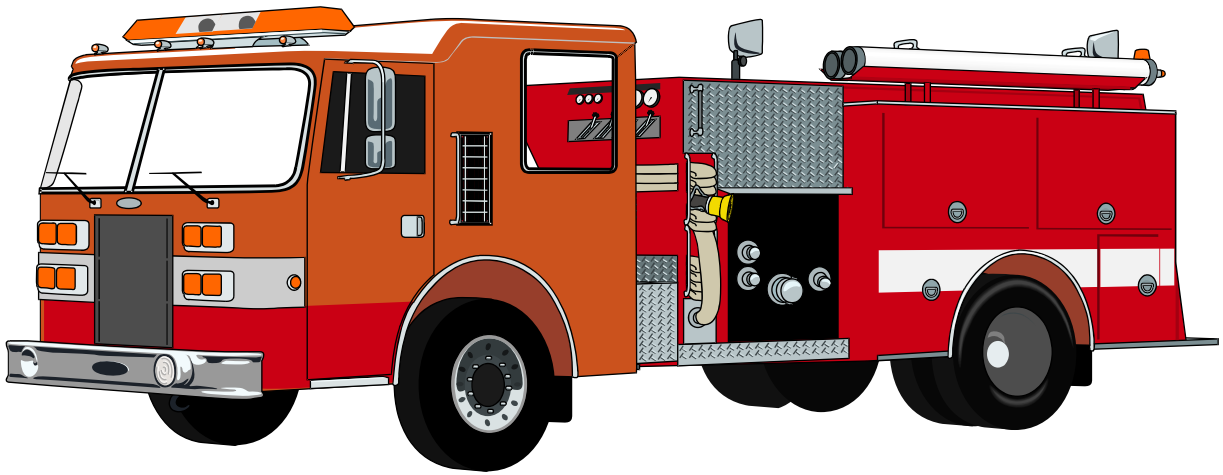
THE UNIVERSITY of NEW MEXICO

Firewall???



Fire Department





What's Censored?

Text	Translation
dajiyuan	Pinyin for Epoch Times
茉莉花革命	Jasmine Revolution
罢课	Strike
freedom	Freedom
请愿书	Petition
华夏论坛	China Forum
学潮	Student protests

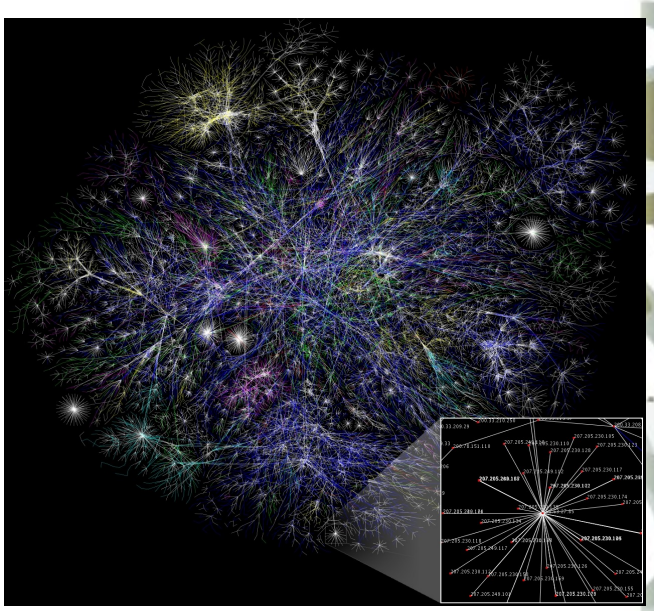
April 22nd, 1989

Stages

- Express grievances
- Build a social network
- Create organization/movement
- Plan protest
- Establish demands
- Advertisement
- Action

Stages

- Express grievances
- Build a social network
- Create organization/movement
- Plan protest
- Establish demands
- Advertisement
- Action



新浪微博

myspace

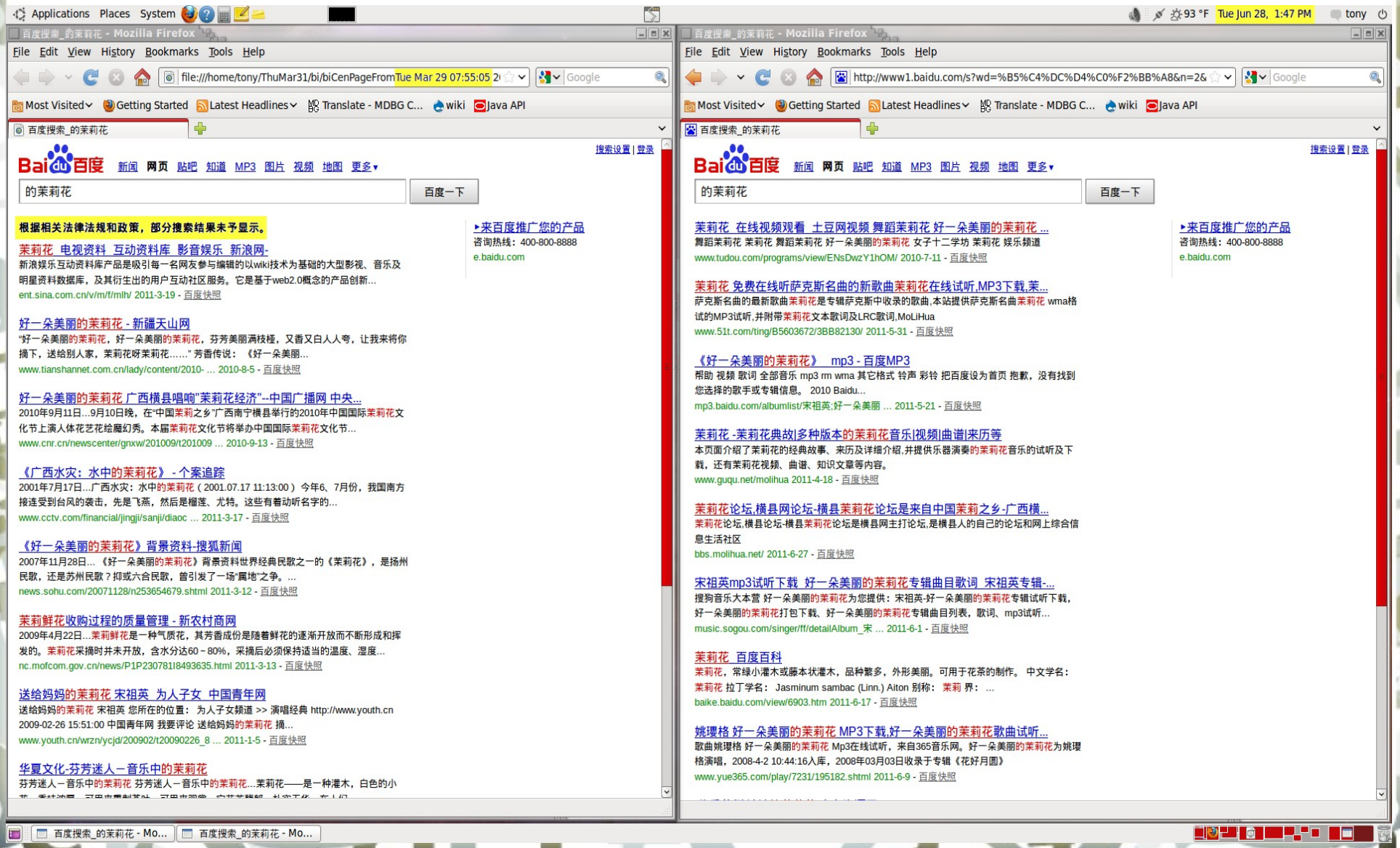
天涯博客 blog.tianya.cn

博客中国 每天五分钟 为思想加油 www.blogchina.com

twitter

facebook

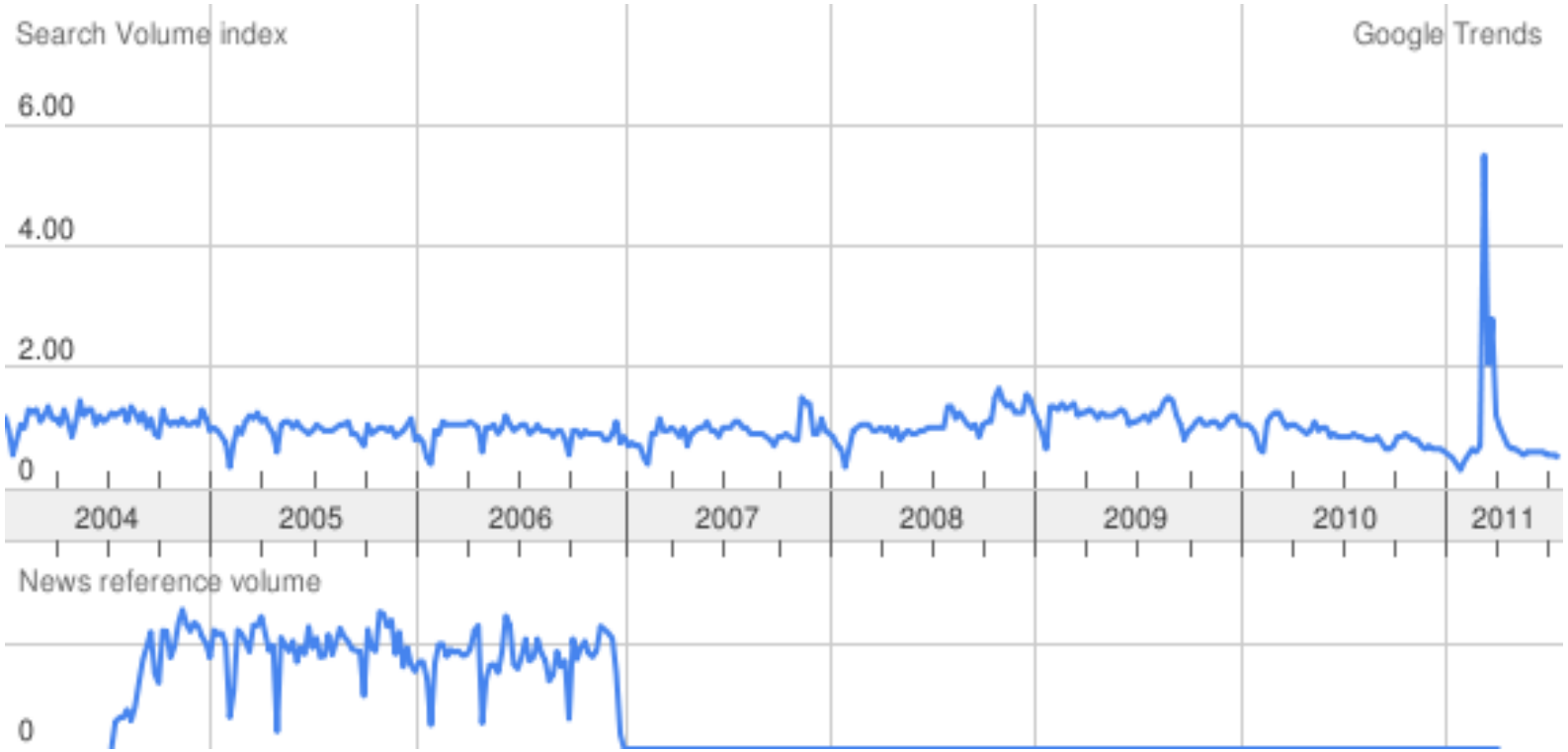
Censorship is dynamic



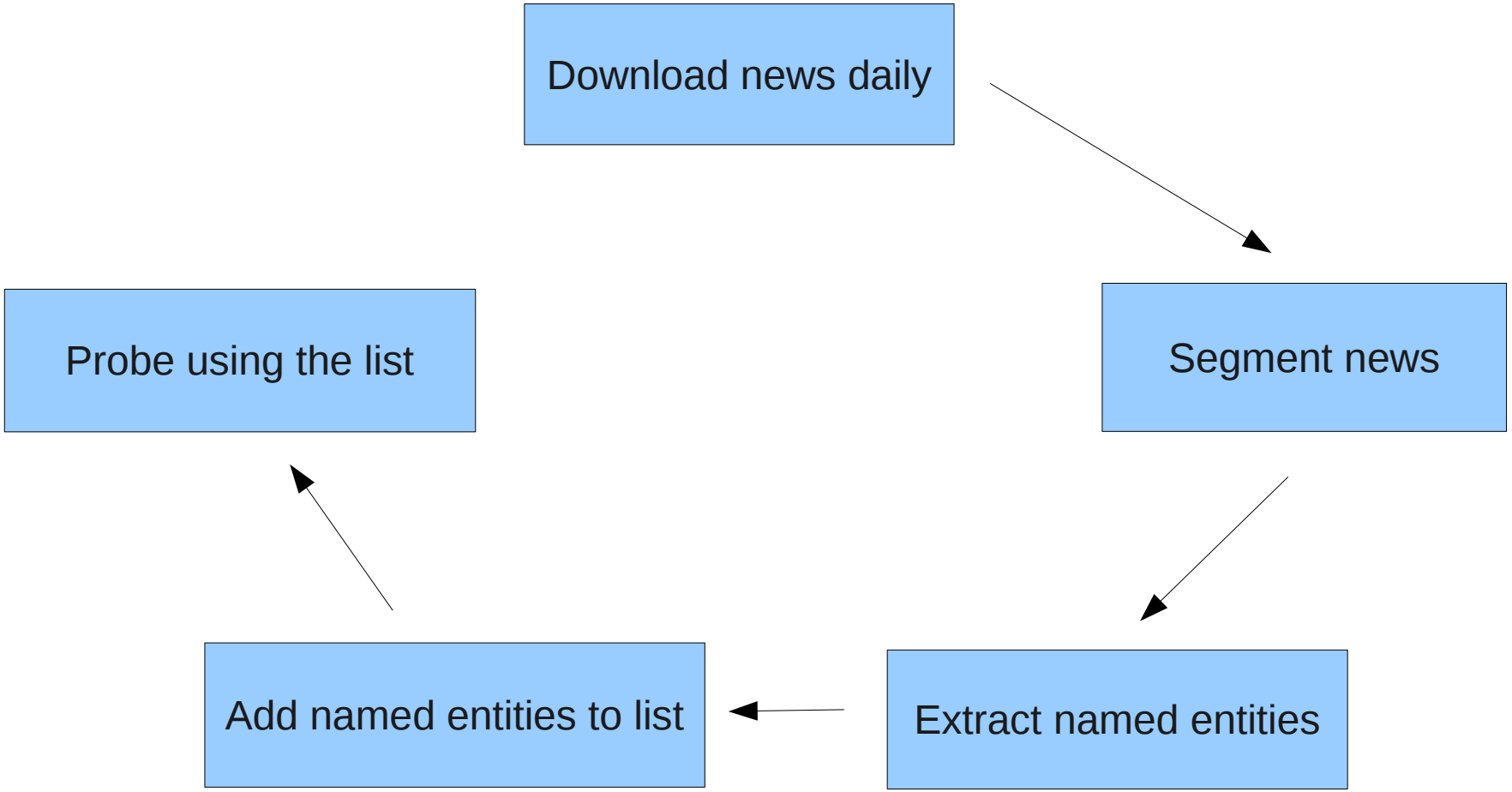
Tuesday, March 29th

Tuesday, June 28th

辐射


Search Volume index
Google Trends
6.00
4.00
2.00
0
2004
2005
2006
2007
2008
2009
2010
2011
News reference volume
0

High-level design

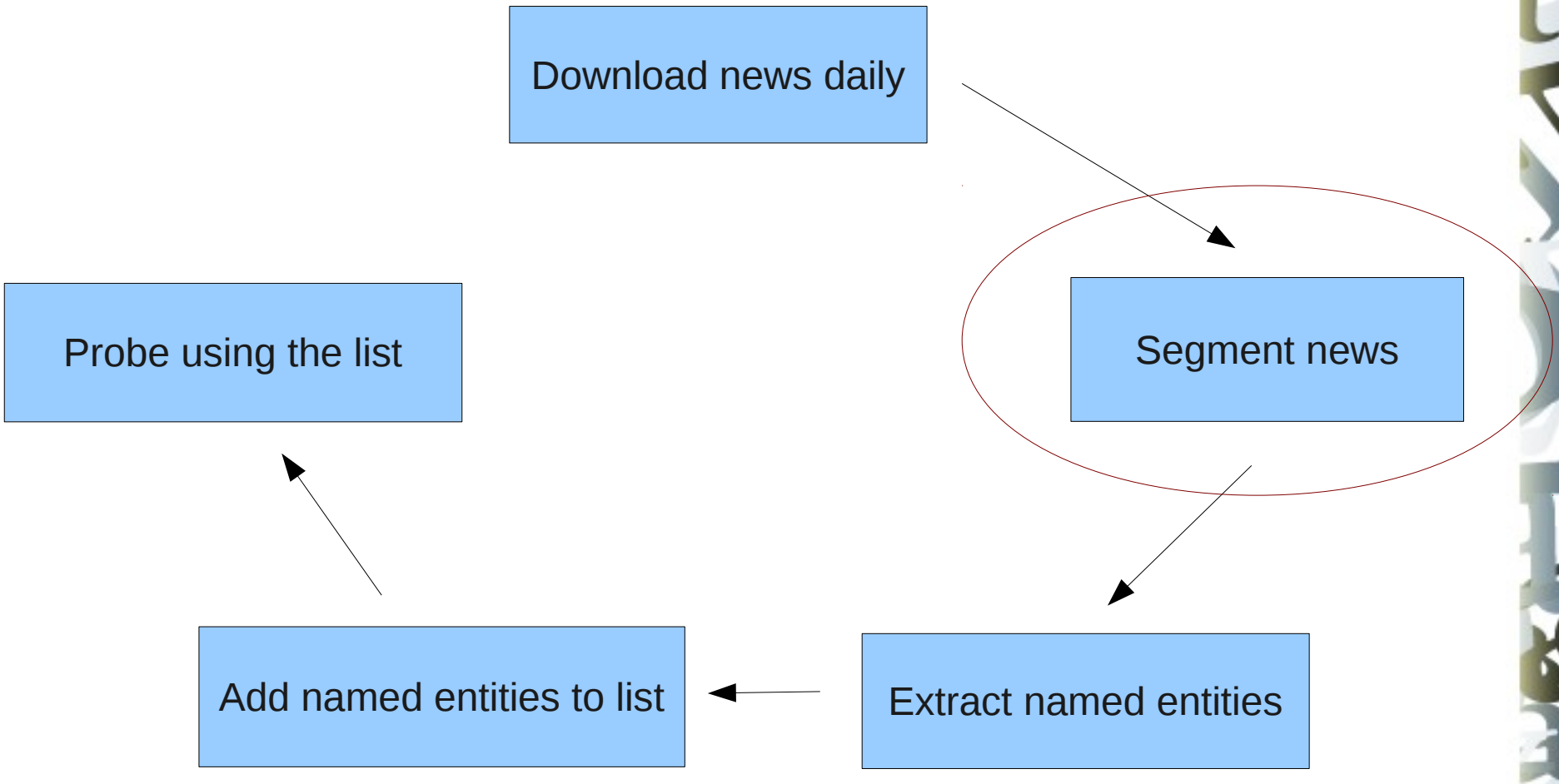


Sources

- http://news.wenxuecity.com/ ★
- http://www.epochtimes.com/ ★
- http://www.cnd.org/ ★
- http://www.popyard.org/
- http://www.rfa.org/ ★
- http://www.sina.com.cn/
- http://www.voanews.com/chinese/news ★

★ -Banned in China

High-level design



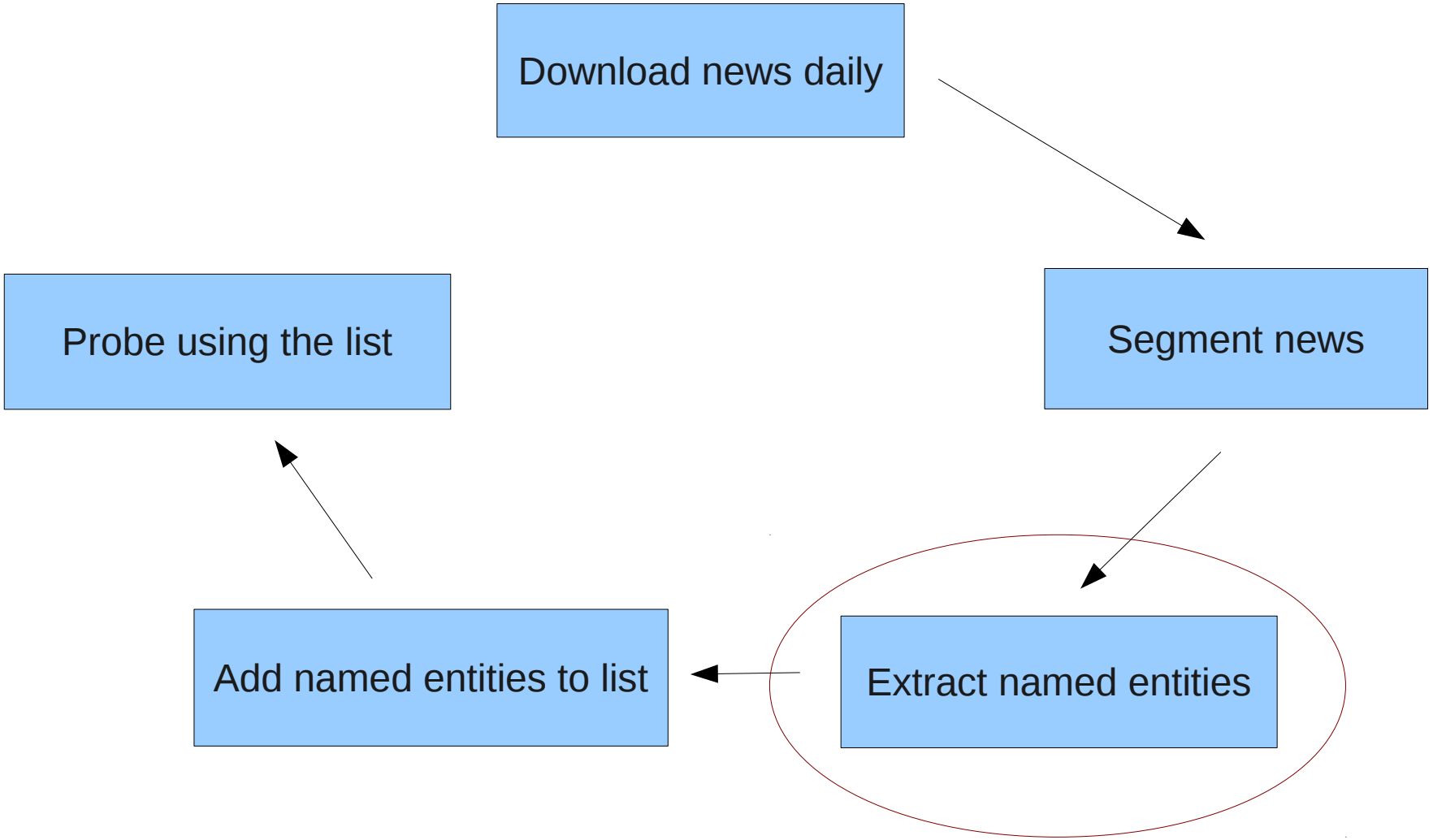
Segmentation

上周获取保候审的北京艺术家艾未未本周一被税务局追罚巨款后，他的妻子路青星期二对本台介绍了丈夫遭软禁及近况，并称已委托两位律师代理此案。浦志强律师告诉记者，已要求税务部门开听证会，解释处罚的理由和程序是否妥当，而现有材料与艾未未

上周 - 获取 - 保 - 候审 - 的 北京 - 艺术家 - 艾 - 未 - 未 本周 - 一 - 被 - 税务局 - 追 罚 - 巨款 后 ， - 他的 - 妻 子 - 路 - 青 - 星期二 - □本 - 台 介绍 - 了 - 丈夫 - 遭 - 软 禁 - 及 - 近况 - ， 并称 - 已 - 委托 - 两位 - 律师 - 代理 - 此 案 。 浦 - 志 - 強 - 律师 - 告 诉 - 记者 - ， 已 - 要求 - 税 务 - 部门 - 开 - 听证 - □ ,- 解释 - 处罚 - 的 - 理由 - 和 - 程序 - 是否 - 妥□ ,- 而 - 现 有 - 材料 - 与 - 艾 - 未 - 未 -

Segmenter written by Erik Peterson
erik AT mandarintools.com

High-level design



Named entity extraction

President Obama flew to New York last night to attend three back-to-back fundraisers. The first event was held at the home of former New Jersey governor and Goldman Sachs chair, Jon Corzine. The New York Times described the $35,800-a-plate event as Obama's first in a "kiss-and-make-up effort" with Wall Street donors. The Obama campaign is hoping to raise $1 billion for the 2012 election.

Place
Organization
Person

News headline from Democracynow.org

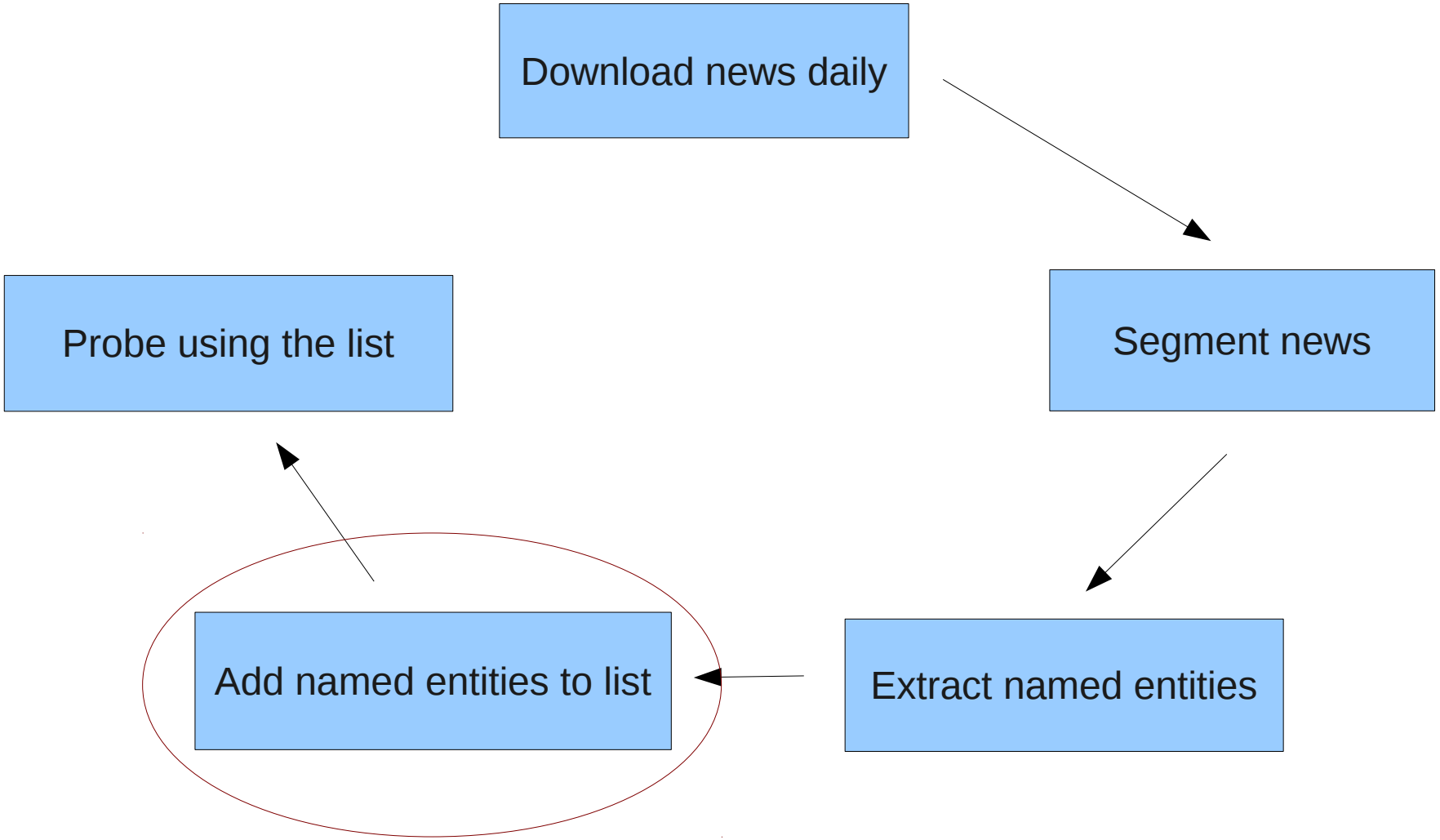
Named entity extraction

《中國時報》是台灣一□綜合性中文報紙 ,常被簡稱□《 中時》，由中國時報社編輯與發行
，與《聯合報》、《自由時報》、《蘋果日報》台灣版□列□台灣四大報。《中國時報》由
報人余紀忠創辦於 1950 年 ,現任社長□吳根成，總經理□蔡紹中，總編輯□王美玉。

Place
Organization
Person

The opennlp.maxent package was originally built by Jason Baldridge, Tom Morton, and Gann Bierner
http://maxent.sourceforge.net/about.html

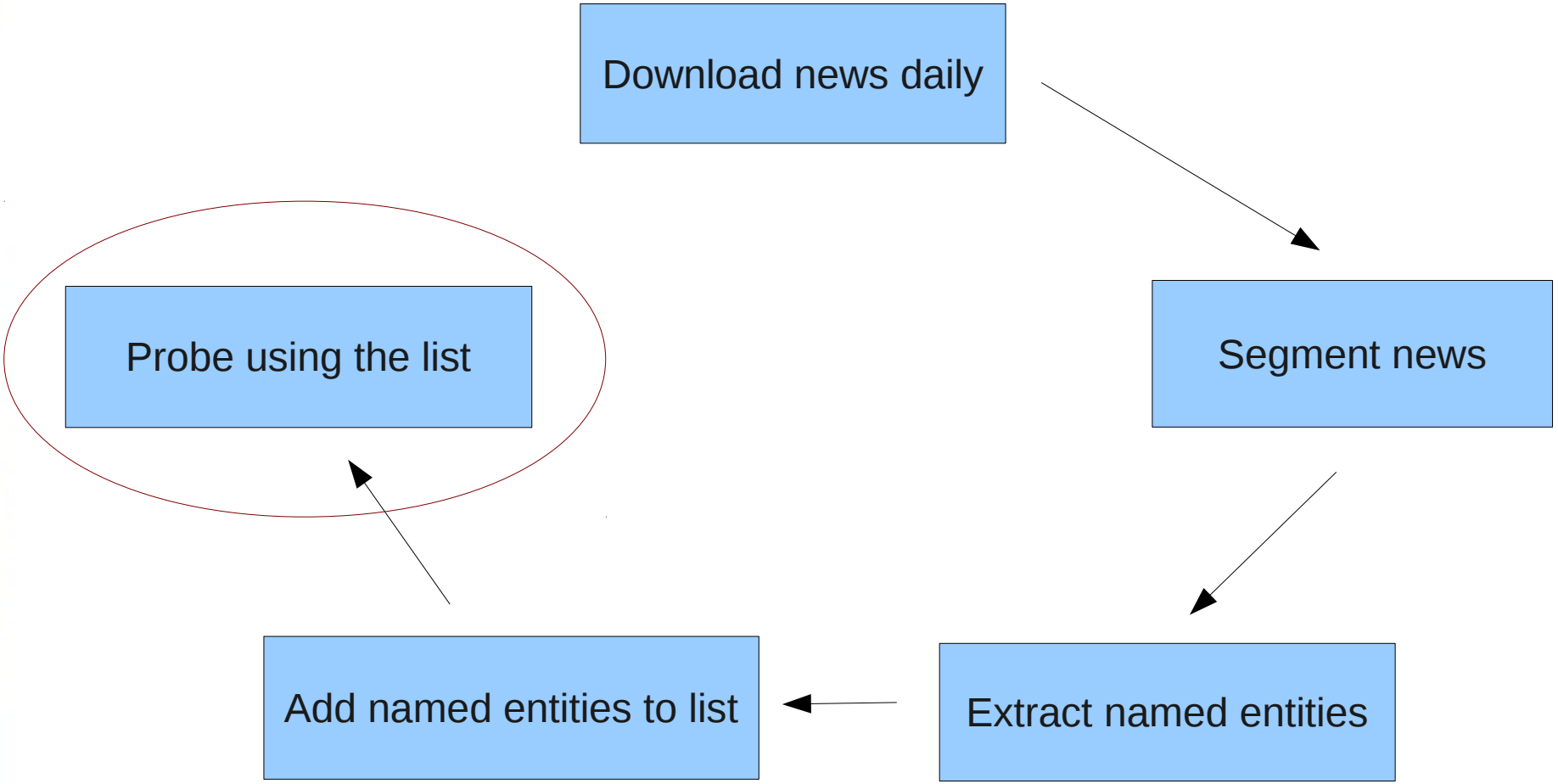
High-level design



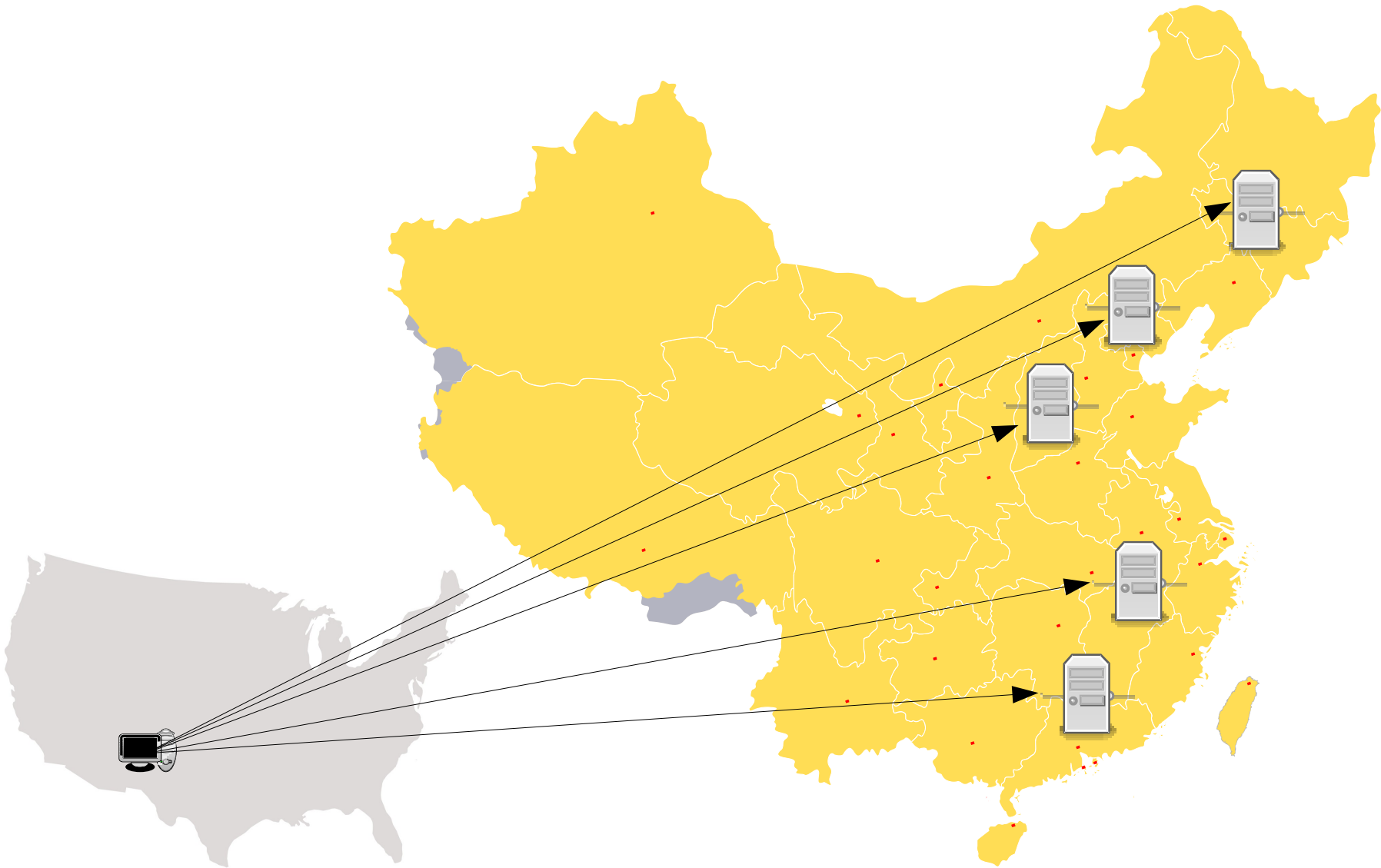
The list

- Priority 0-15
- 0 means blacklisted somewhere
- Words start out at priority 1
- After each probe move up
- After 15 probes, fall off the list
- Search a priority every 12 hours
- Only GET request censorship affects the list

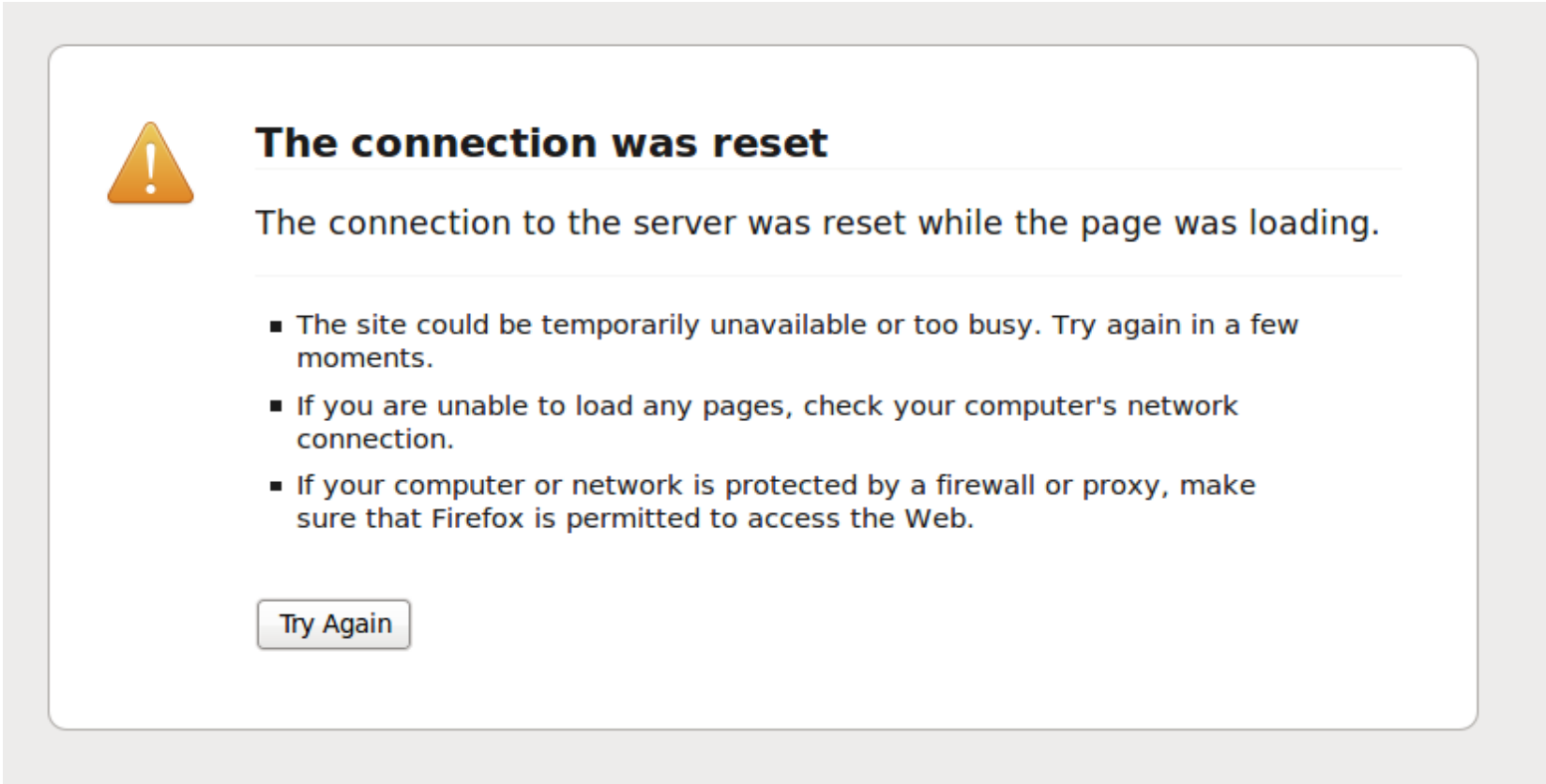
High-level design



1) GET request probing

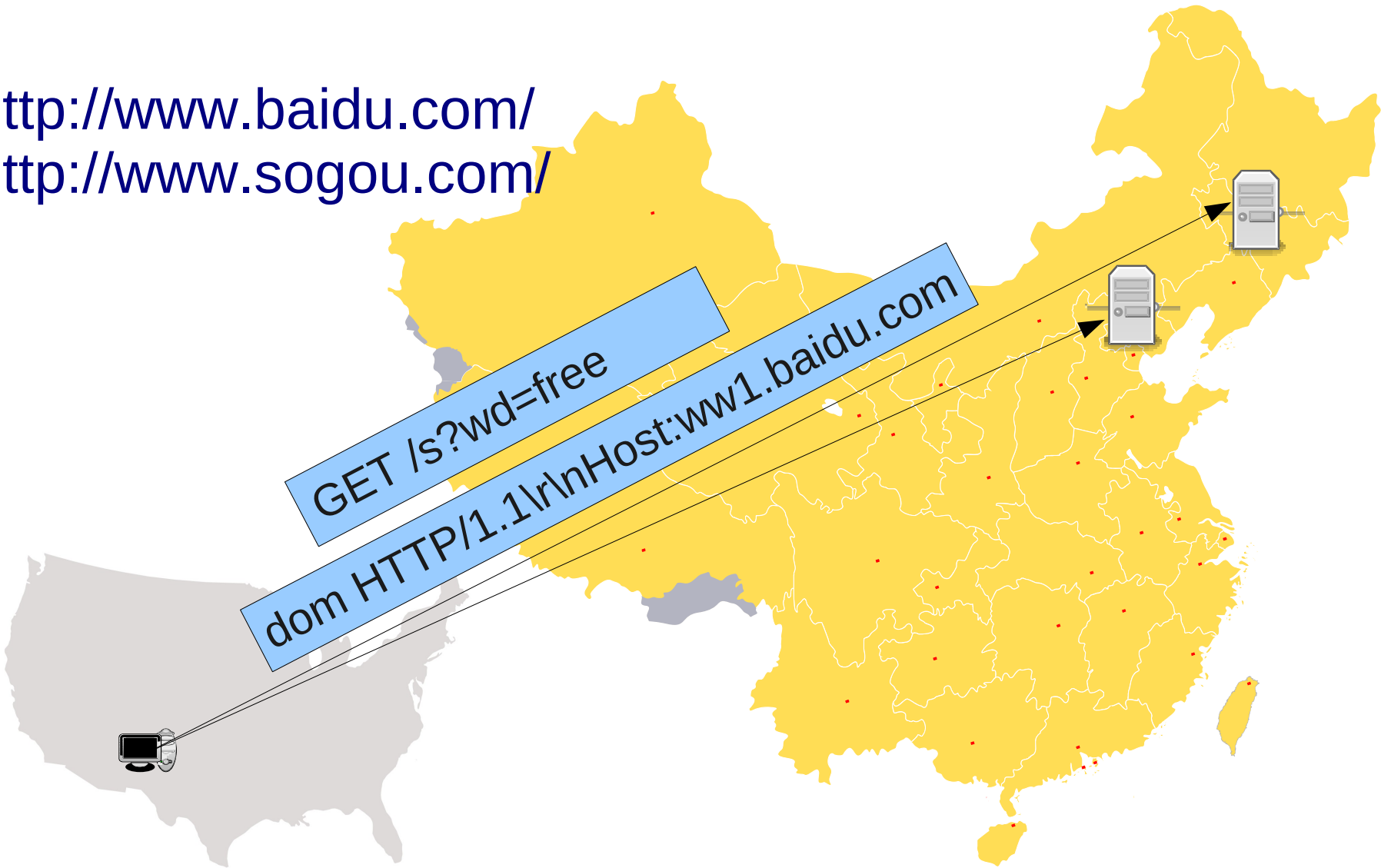


1) GET request probing



2) Search engine probing

http://www.baidu.com/
http://www.sogou.com/



2) Search engine probing

Implementation of NEE

伦敦
政治经济学
院

---

prev_1 prev_is_dict_term prev_Iverb prev_noun prev_adverb prev_verb 2 is_dict_term has_caps next_5 next_is_dict_term ?
prev_2 prev_is_dict_term prev_has_caps 5 is_dict_term next_1 next_is_dict_term next_noun ?
prev_5 prev_is_dict_term 1 is_dict_term noun next_2 next_is_dict_term next_verb ?

COMPLETE_PLACE NOT_PLACE NOT_PLACE	BEGINNING_NAME MIDDLE_NAME BEGINNING END_NAME	NOT_COMP BEGINNING_COMP END_COMP

What is maximum entropy modeling

- Constraints
- Find the maximal entropy model that meets the constraints

Why use maximum entropy?

- Can use diverse features
  - Flexibility of multiple languages
- Assume as little as possible about things you have not seen before
  - This sets it apart from maximum likelihood estimation

Topological sort

```
For context: prev_1 prev_has_punctuation 2 is_dict_term Tverb verb has_caps next_2 next_is_dict_term next_Iverb
NOT_NAME[1.0000]  BEGINNING_NAME[0.0000]  MIDDLE_NAME[0.0000]  END_NAME[0.0000]  COMPLETE_NAME[0.0000]

For context: prev_2 prev_is_dict_term prev_Tverb prev_verb prev_has_caps 2 is_dict_term Iverb next_7
NOT_NAME[0.9993]  BEGINNING_NAME[0.0006]  MIDDLE_NAME[0.0000]  END_NAME[0.0000]  COMPLETE_NAME[0.0000]

For context: prev_2 prev_is_dict_term prev_Iverb 7 next_1 next_has_punctuation
NOT_NAME[0.9993]  BEGINNING_NAME[0.0000]  MIDDLE_NAME[0.0000]  END_NAME[0.0000]  COMPLETE_NAME[0.0006]

For context: prev_7 1 has_punctuation next_1 next_is_dict_term
NOT_NAME[1.0000]  BEGINNING_NAME[0.0000]  MIDDLE_NAME[0.0000]  END_NAME[0.0000]  COMPLETE_NAME[0.0000]

For context: prev_1 prev_has_punctuation 1 is_dict_term next_2 next_is_dict_term next_noun
NOT_NAME[1.0000]  BEGINNING_NAME[0.0000]  MIDDLE_NAME[0.0000]  END_NAME[0.0000]  COMPLETE_NAME[0.0000]
```

Experimental methodology

Experimental methodology

Train

Test

Maximum Entropy results

Places

Specificity: 69.8%

Recall: 96.3%

Precision: 0.77%

Organizations

Specificity: 88.4%

Recall: 87.56%

Precision: 0.28%

People

Specificity: 83.4%

Recall: 89.63%

Precision: 0.42%

HTTP GET request blacklist additions

Text	Translation
dajiyuan	Pinyin for Epoch Times
罢课	Strike
freedom	Freedom
请愿书	Petition
华夏论坛	China Forum
学潮	Student protests

Search engine sensitive words

Text	Translation
茉莉花	Jasmine flower
诺贝尔	Nobel Prize
刘先生	Mr. Liu
挪威	Norway
七十七 ,77, 七七	77
王府井	Wangfujing

This is not a complete list

Conclusions

During the two months that we gathered data the GET request black list appeared to be static. Even the Jasmine Flower (茉莉花) keyword didn't appear as a blip on the GET request censorship.

Future Work

- Focus more on search engine and blogging censorship
- Improve the NEE
- Support other languages
- Add daily TOM-Skype list

Acknowledgments

- Anonymous FOCI reviewers
- Fletcher Hazlehurst, Veronika Strnadova, Leif Guillermo, Ronald Garduno, Terran Lane,Sergey Plis
- This material is based upon work supported by the National Science Foundation under Grant No.#0844880 and #1025447

Thank You

¿Questions?