

# Automated Named Entity Extraction for Tracking Censorship of Current Events

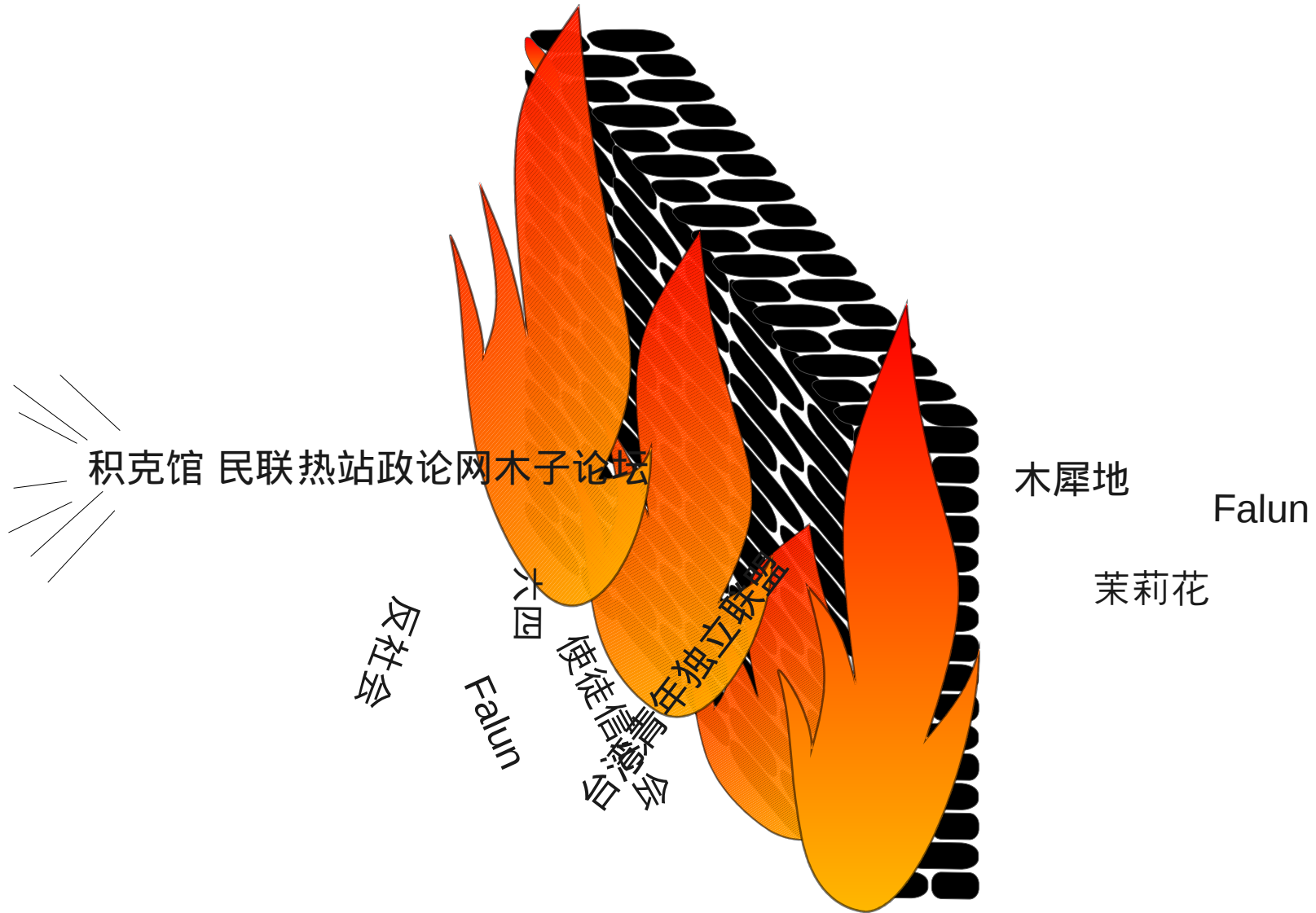
Antonio M. Espinoza & Jedidiah R. Crandall  
Computer Science Department  
University of New Mexico

{amajest,crandall}@cs.unm.edu

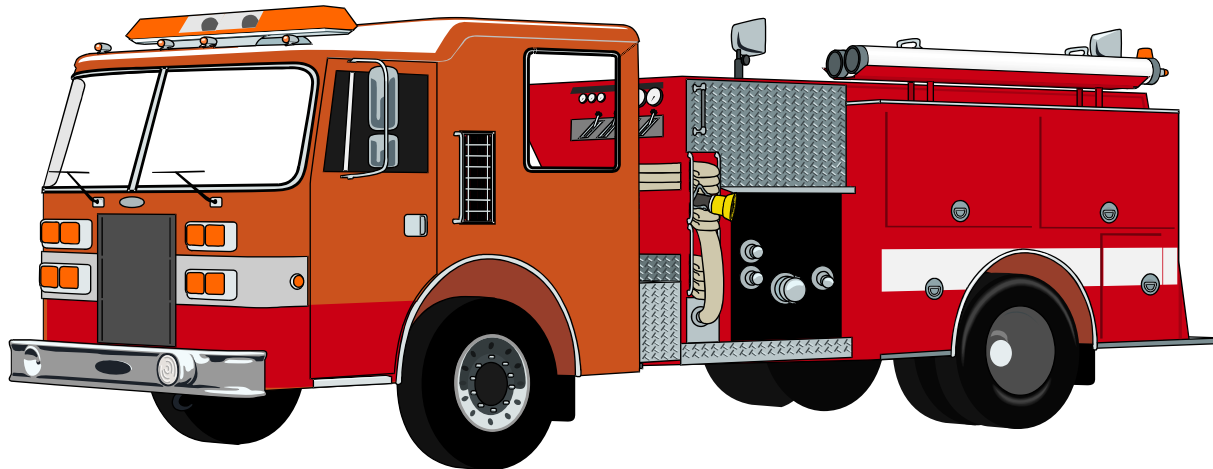


THE UNIVERSITY *of*  
NEW MEXICO

# 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 22<sup>nd</sup>, 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

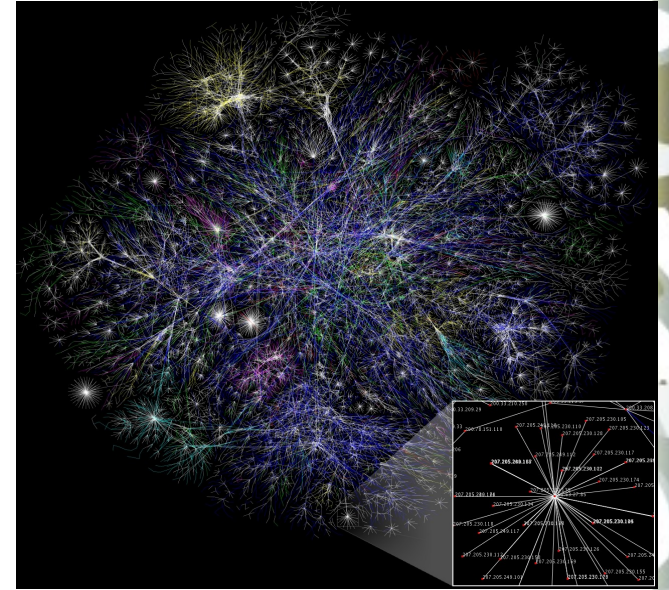
- Plan protest



- Establish demands

- Advertisement

- Action



**myspace**

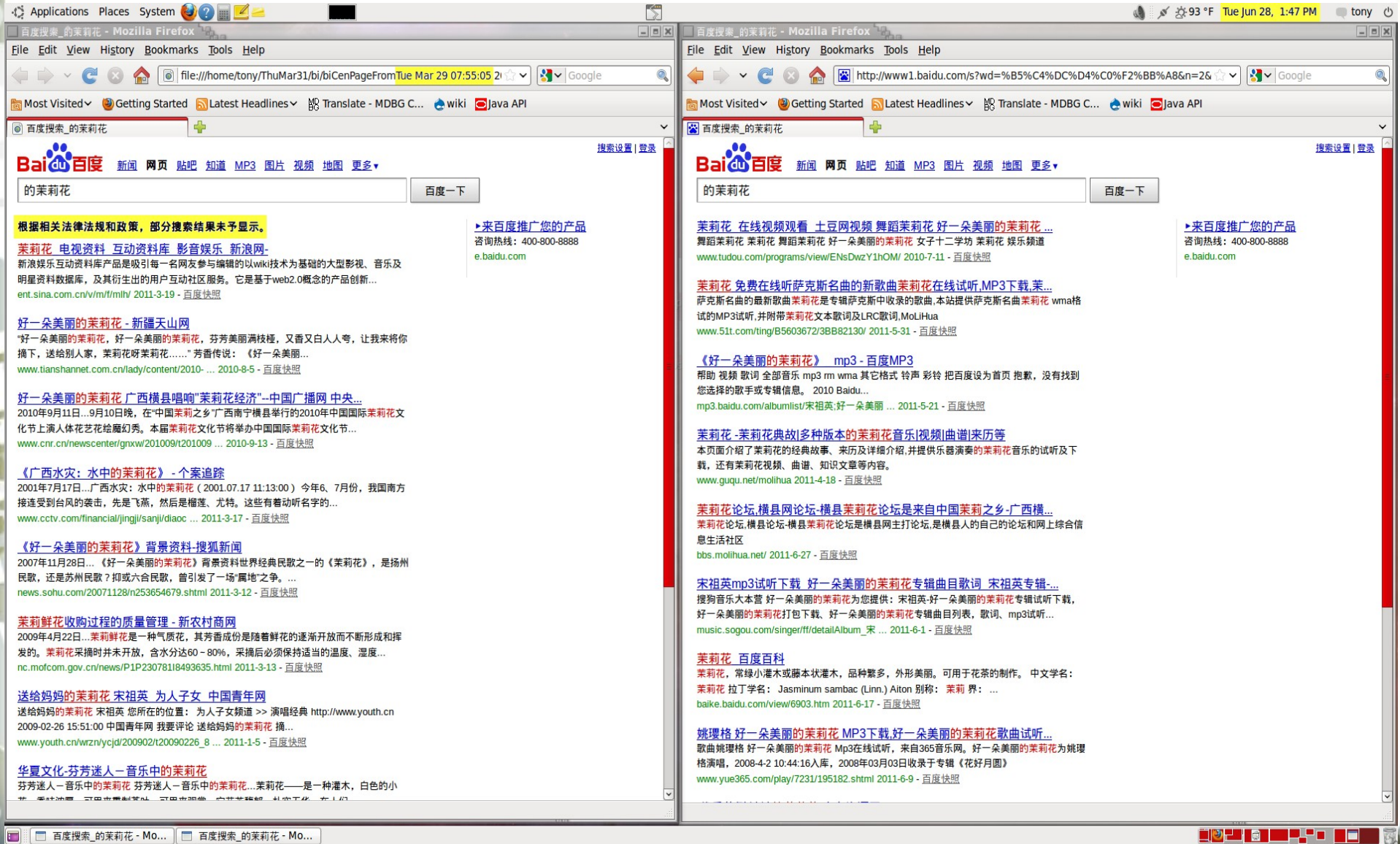


**twitter** 

**facebook**



# Censorship is dynamic



# Tuesday, March 29<sup>th</sup>

File Edit View History Bookmarks Tools Help

file:///home/tony/ThuMar31/bi/biCenPageFromTue Mar 29 07:55:05 21

Most Visited Getting Started Latest Headlines Translate - MDBG C... wiki Java API

百度搜索\_的茉莉花

**Baidu 百度** 新闻 网页 贴吧 知道 MP3 图片 视频 地图 更多

的茉莉花 百度一下

根据相关法律法规和政策，部分搜索结果未予显示。

[茉莉花 电视资料 互动资料库 影音娱乐 新浪网-](#)

新浪娱乐互动资料库产品是吸引每一名网友参与编辑的以wiki技术为基础的大型影视、音乐及明星资料数据库，及其衍生出的用户互动社区服务。它是基于web2.0概念的产品创新...

[ent.sina.com.cn/v/m/f/mlhv](http://ent.sina.com.cn/v/m/f/mlhv) 2011-3-19 - [百度快照](#)

来百度推广  
咨询热线：400-  
[e.baidu.com](http://e.baidu.com)

# Tuesday, June 28<sup>th</sup>

93 °F Tue Jun 28, 1:47 PM

百度搜索\_的茉莉花 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www1.baidu.com/s?wd=%B5%C4%DC%D4%C0%F2%BB%A8&n=2& Google

Most Visited Getting Started Latest Headlines Translate - MDBG C... wiki Java API

百度搜索\_的茉莉花

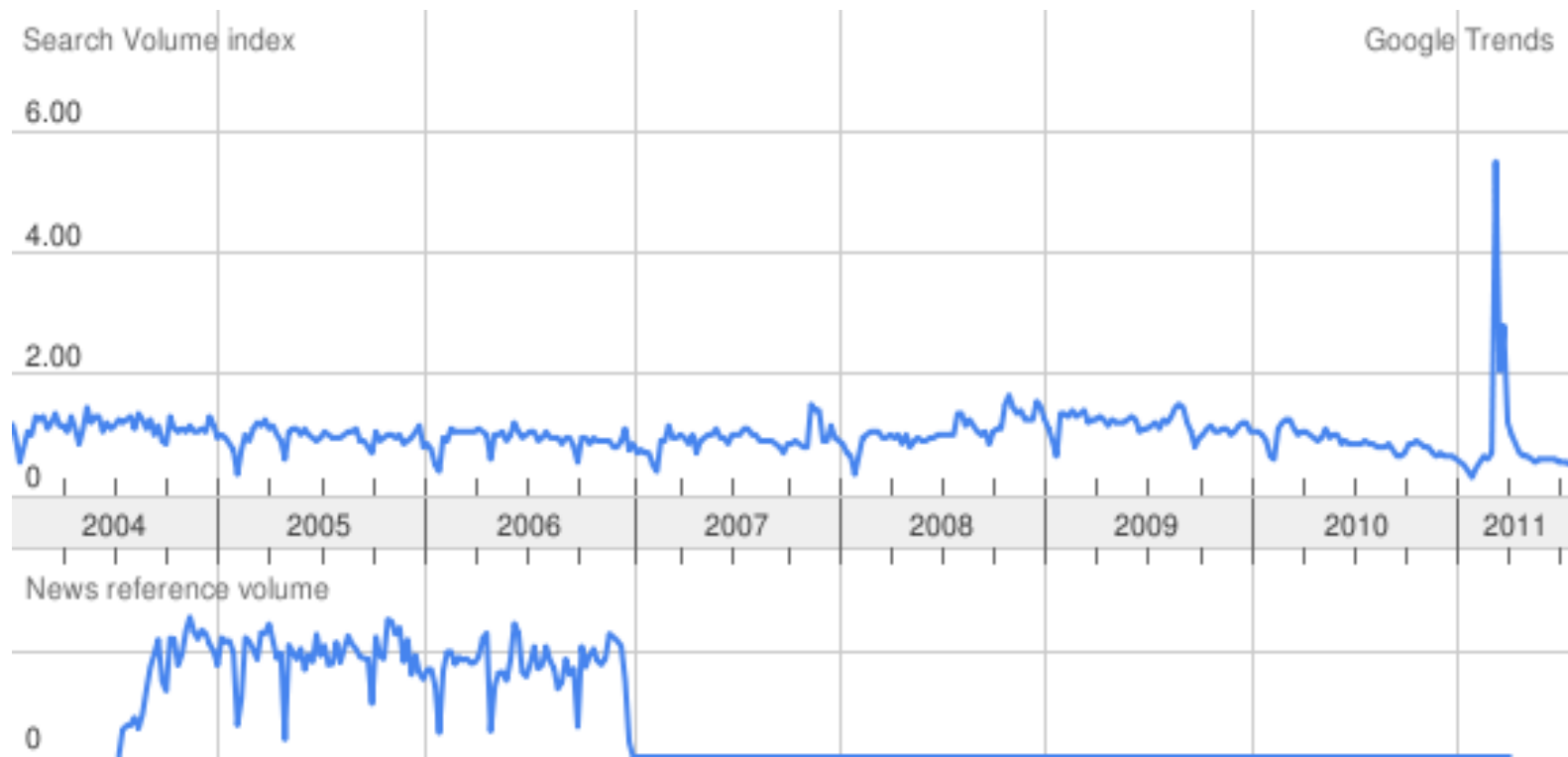
**Baidu 百度** 新闻 网页 贴吧 知道 MP3 图片 视频 地图 更多

的茉莉花 百度一下

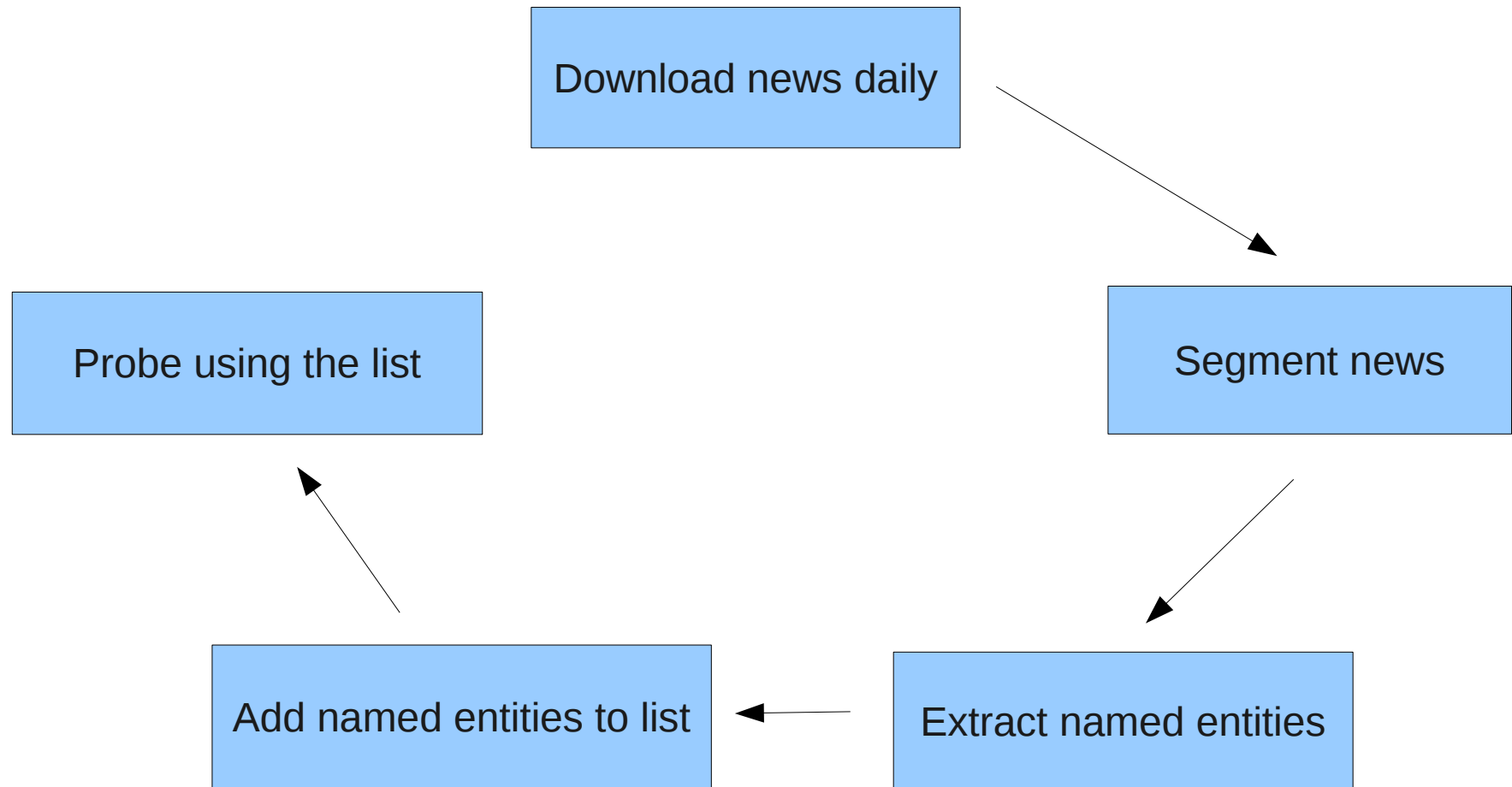
[茉莉花 在线视频观看 土豆网视频 舞蹈茉莉花 好一朵美丽的茉莉花 ...](#)  
舞蹈茉莉花 茉莉花 舞蹈茉莉花 好一朵美丽的茉莉花 女子十二学坊 茉莉花 娱乐频道  
[www.tudou.com/programs/view/ENsDwzY1hOM/](http://www.tudou.com/programs/view/ENsDwzY1hOM/) 2010-7-11 - 百度快照

来百度推广您的产品  
咨询热线: 400-800-8888  
[e.baidu.com](http://e.baidu.com)

# 辐射



# High-level design



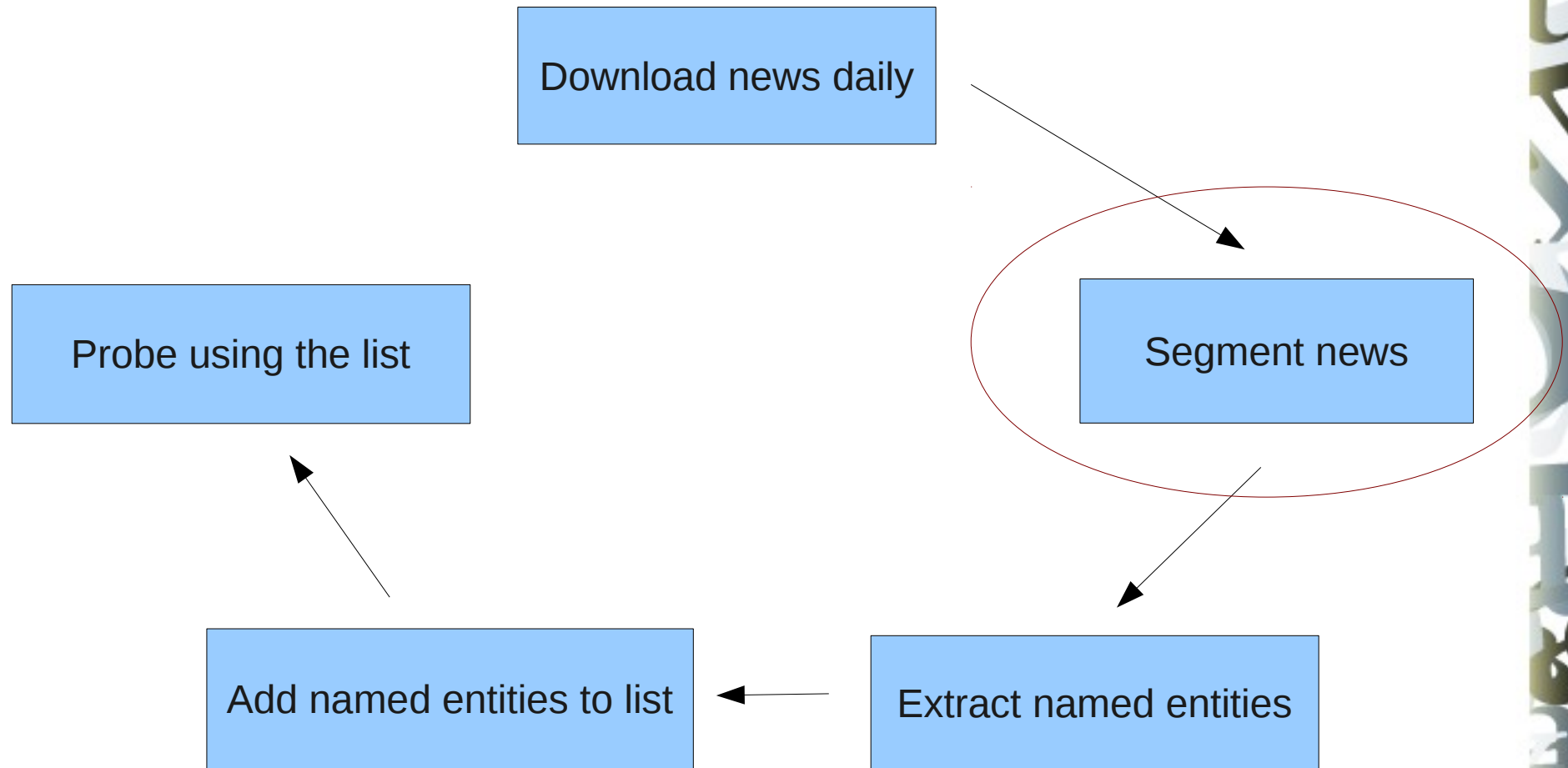
# Sources

- <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



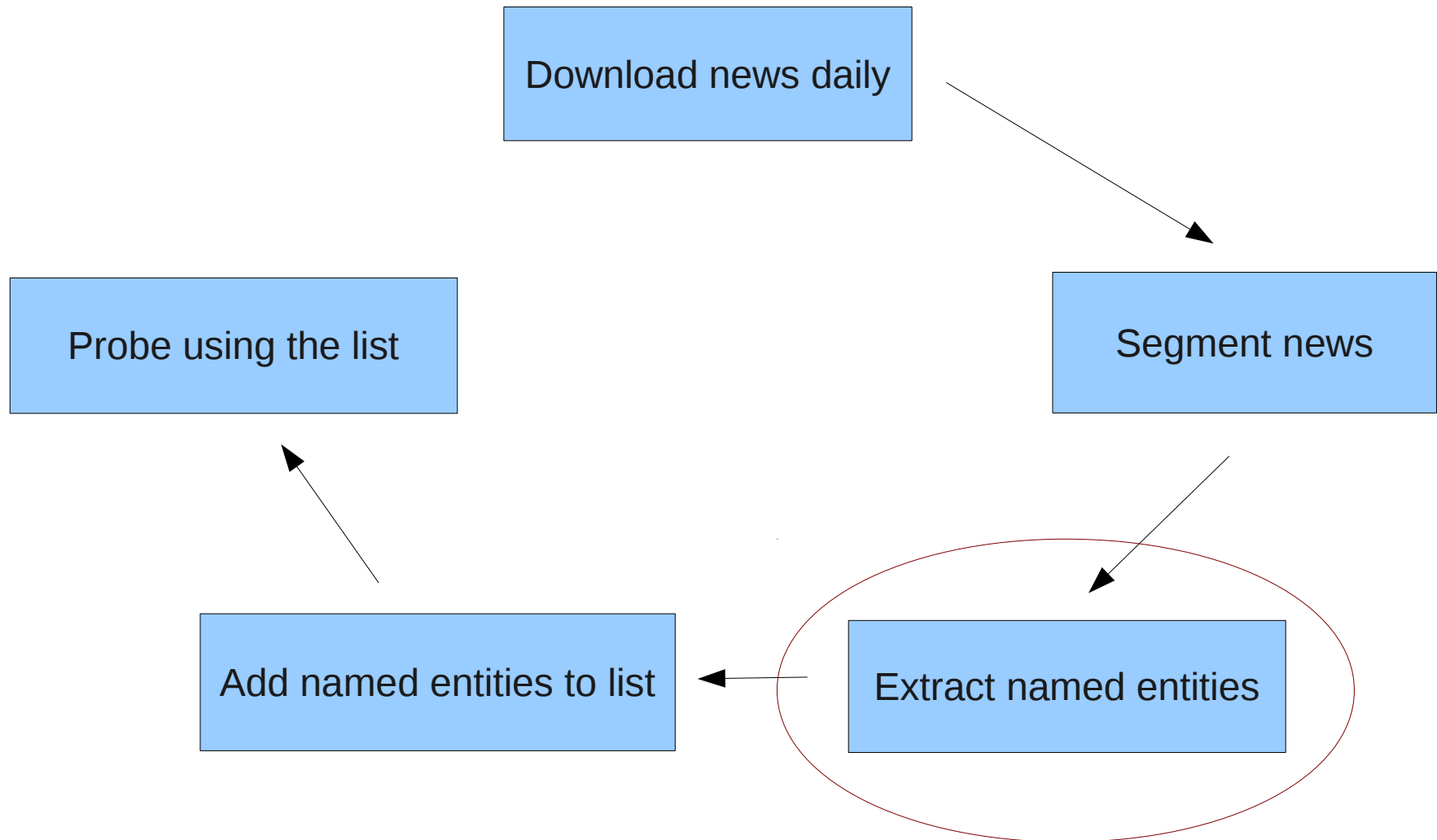
# Segmentation

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 - 强 - 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 - 当 □ ， - 而 - 现  
有 - 材 - 料 - 与 - 艾 - 未 - 未 -



# 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

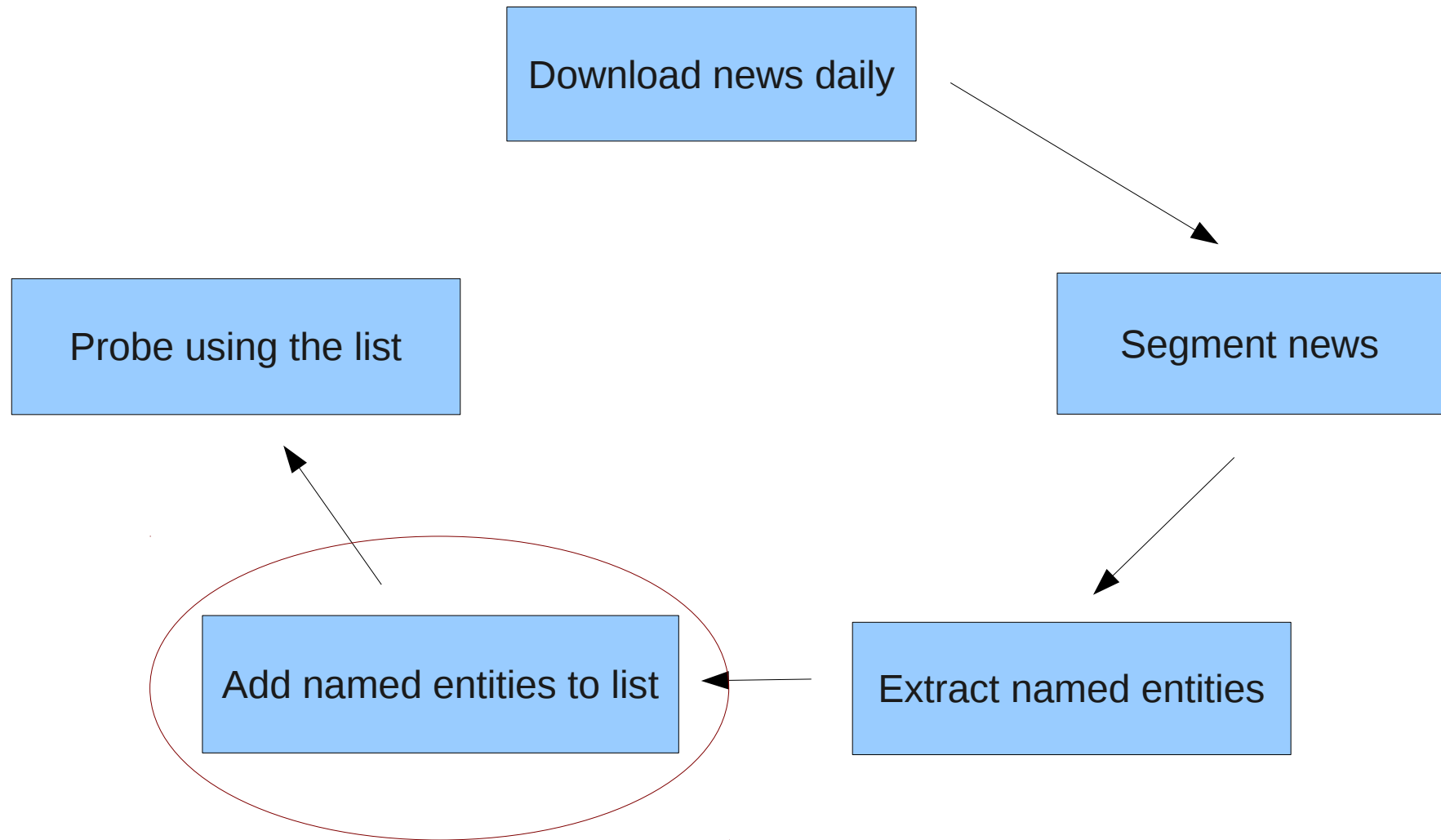
# Named entity extraction

《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。

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

Place  
Organization  
Person

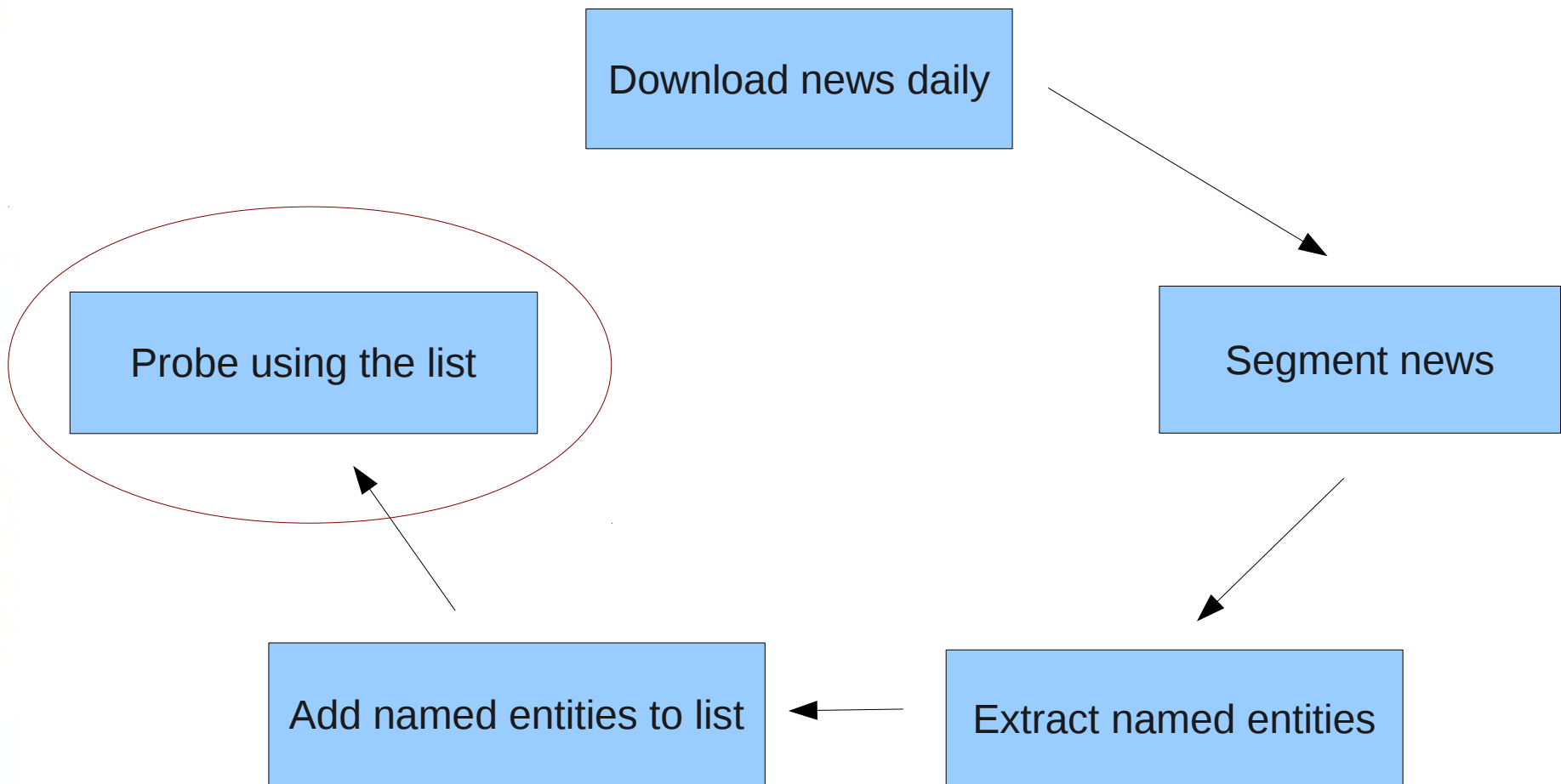
# High-level design



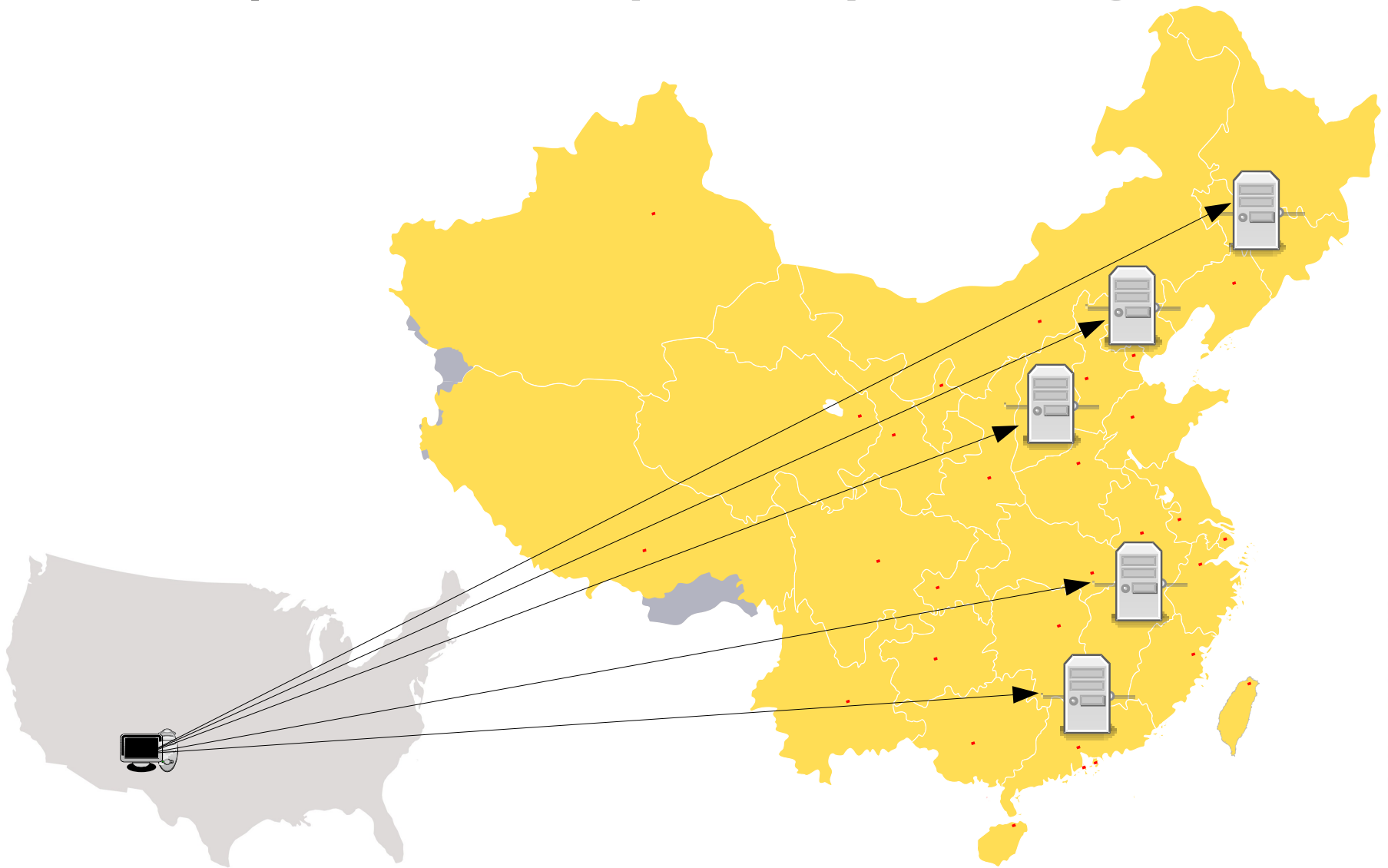
# The list

- 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



## **The connection was reset**

The connection to the server was reset while the page was loading.

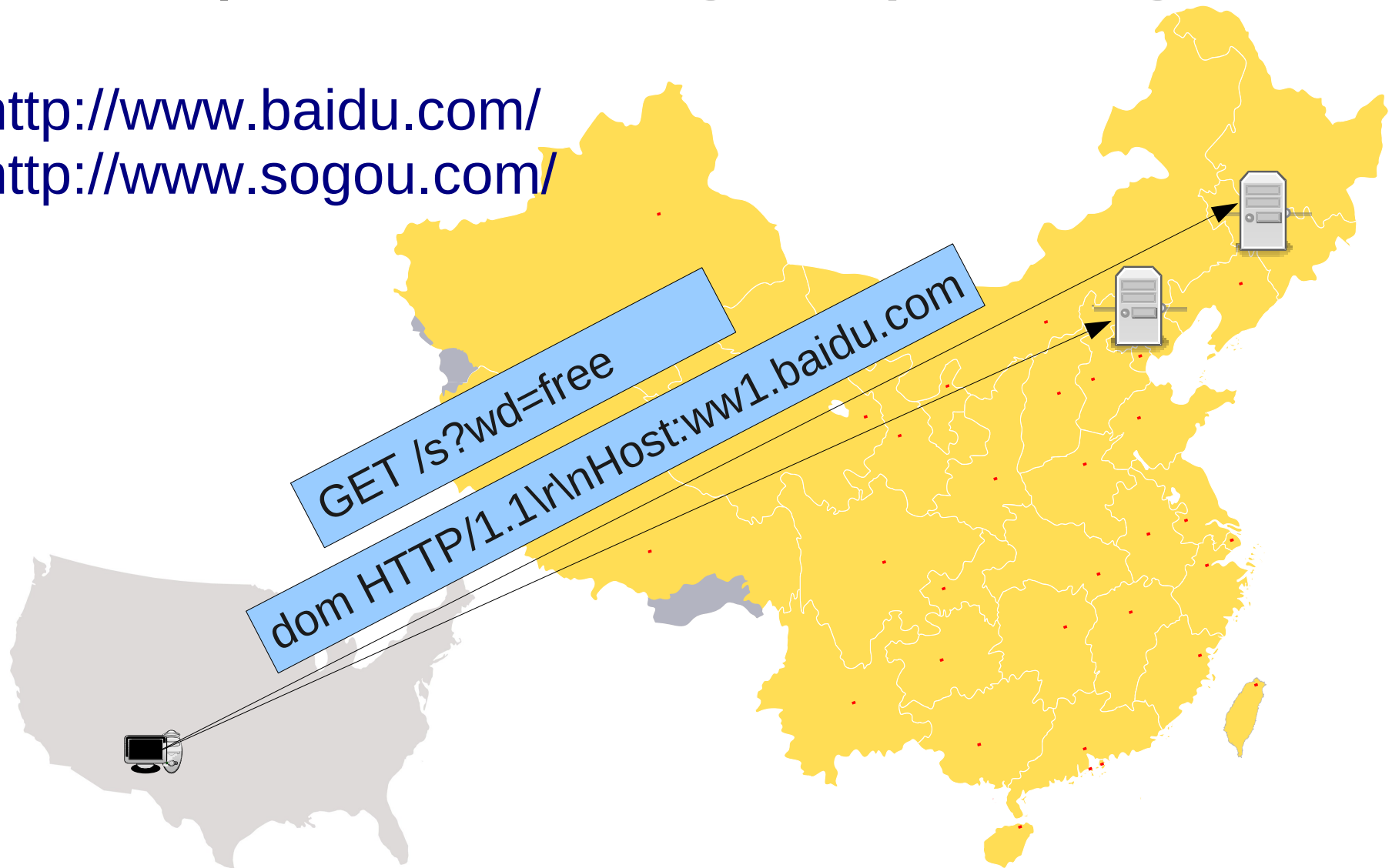
- The site could be temporarily unavailable or too busy. Try again in a few moments.
- If you are unable to load any pages, check your computer's network connection.
- If your computer or network is protected by a firewall or proxy, make sure that Firefox is permitted to access the Web.

Try Again



## 2) Search engine probing

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



## 2) Search engine probing

Baidu 百度

新闻 网页 贴吧 知道 MP3 图片 视频 地图 更多 ▾

茉莉花革命

百度一下

根据相关法律法规和政策，部分搜索结果未予显示。

[2011年一月互联网十大热词出炉 茉莉花革命上榜 社会频道 新华网](#)

2011年2月9日...2011年一月互联网十大热词出炉 茉莉花革命上榜 2011年02月09日 08:44:33 来

源：人民网-人民日报海外版 【字号 大小】 【留言】 【打印】 【关闭】 ...

[news.xinhuanet.com/society/2011-02/09/c\\_1 ...](http://news.xinhuanet.com/society/2011-02/09/c_1...) 2011-2-9 - 百度快照

[阿拉伯世界忧虑茉莉花革命蔓延](#)

2011年1月19日...阿拉伯世界忧虑茉莉花革命蔓延,这接连而来的自焚事件被外界认为是受到了突尼斯“茉莉花革命”的鼓励。埃及作为阿拉伯世界大国，突尼斯前总统本·阿里逃亡沙特...

[jingji.cntv.cn/20110119/100279.shtml](http://jingji.cntv.cn/20110119/100279.shtml) 2011-6-17 - 百度快照

[2011年一月互联网十大热词出炉 茉莉花革命上榜 --地方--人民网](#)

2011年2月9日...2011年一月互联网十大热词出炉 茉莉花革命上榜 2011年02月09日15:00 ... 茉

莉花革命：突尼斯某街头小贩遭到执法人员的粗暴对待随即自焚抗议。该事件...

[zb.people.com.cn/GB/119638/13878488.html](http://zb.people.com.cn/GB/119638/13878488.html) 2011-2-9 - 百度快照

[稳定，是中国人最想要的--观点--人民网](#)

2011年3月10日...“中国的茉莉花革命”。他们利用自己的传播优势，...就会知道他们到底是想要革命还是想要稳定。那些...

[opinion.people.com.cn/GB/14111310.html](http://opinion.people.com.cn/GB/14111310.html) 2011-3-10 - 百度快照

# Implementation of NEE

伦敦  
政治经济学  
院

---

prev\_1 prev\_is\_dict\_term prev\_lverb 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





# Maximum Entropy results

## Places

Specificity: 69.8%

Recall: 96.3%

Precision: 0.77%

## People

Specificity: 83.4%

Recall: 89.63%

Precision: 0.42%

## Organizations

Specificity: 88.4%

Recall: 87.56%

Precision: 0.28%

# 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?