

Automated Named Entity Extraction for Tracking Censorship of Current Events

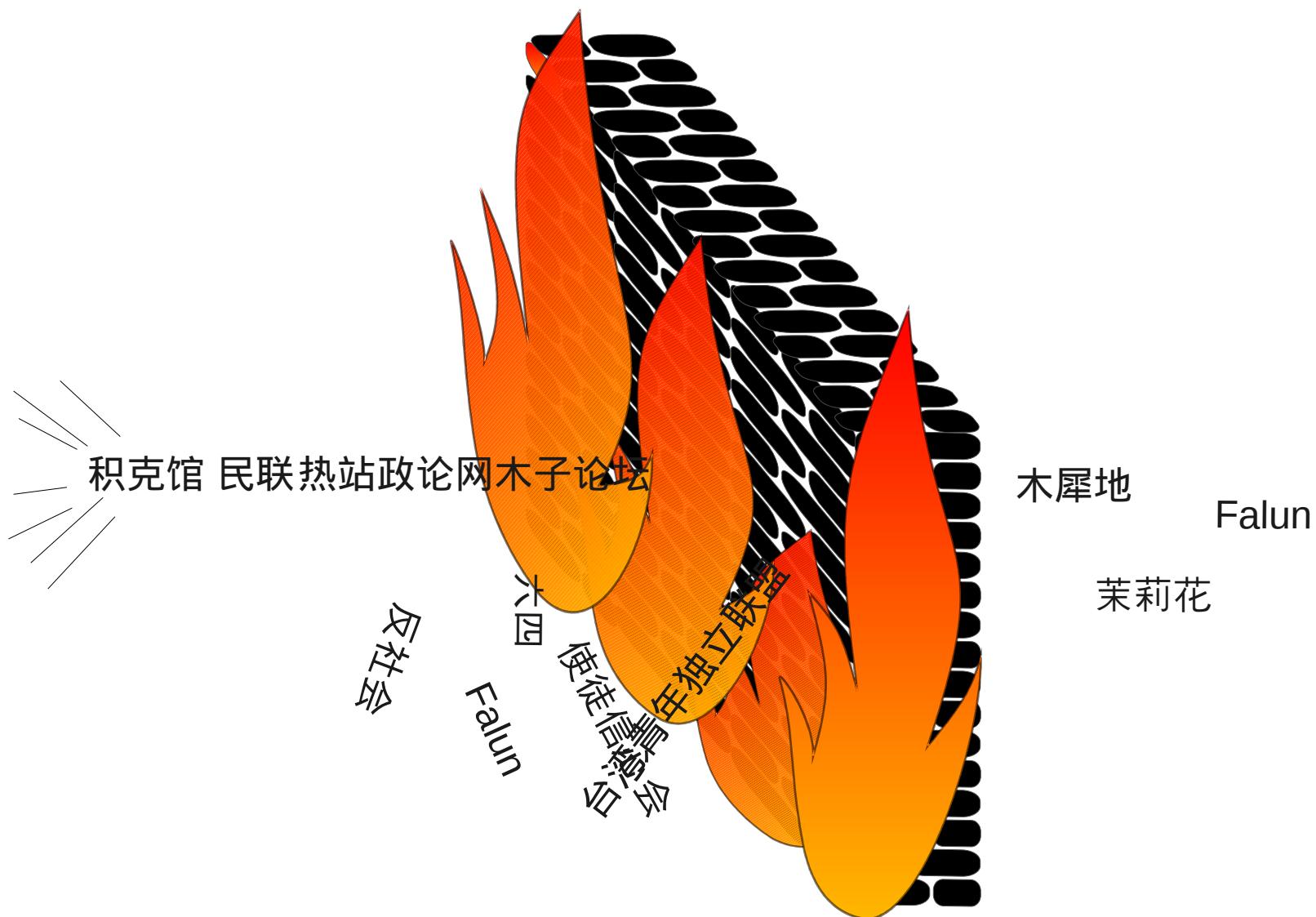
Antonio M. Espinoza & Jedidiah R. Crandall
Computer Science Department
University of New Mexico

{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



64memo.com

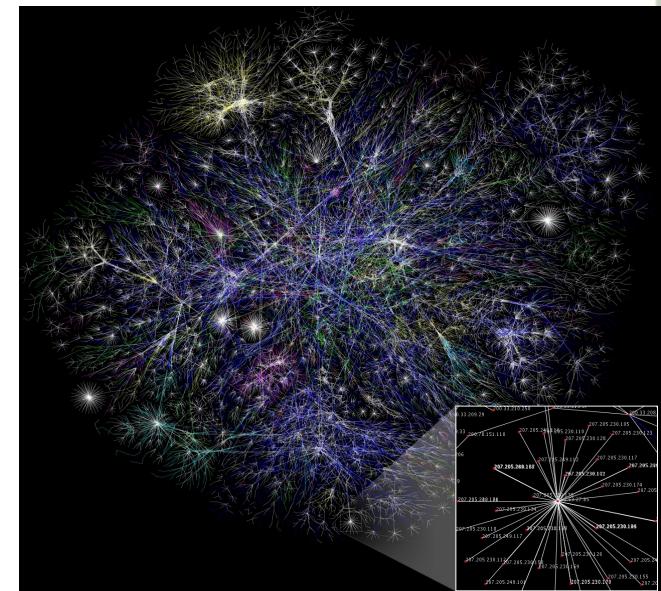
Stages

- 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



facebook

Censorship is dynamic

百度搜索_的茉莉花 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

file:///home/tony/ThuMar31/bi/biCenPageFromTue Mar 29 07:55:05 2011

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 音影娱乐 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/mfl/mflv/ 2011-3-19 - 百度快照

好一朵美丽的茉莉花 - 新疆天山网

“好一朵美丽的茉莉花，好一朵美丽的茉莉花，芬芳美丽满枝桠，又香又白人人夸，让我来将你摘下，送给别人家，茉莉花呀茉莉花……”芳香传说：《好一朵美丽...

www.tianshannet.com/lady/content/2010/ ... 2010-8-5 - 百度快照

好一朵美丽的茉莉花 广西横县唱响“茉莉花经济”--中国广播网 中央...

2010年9月11日...9月10日晚，在中国“茉莉之乡”广西南宁横县举行的2010年中国国际“茉莉花”文化节上演人体花艺绘廊幻秀，本届“茉莉花”文化节将举办中国国际“茉莉花”文化节...

www.cnr.cn/newscenter/gnxw/201009/t201009 ... 2010-9-13 - 百度快照

《广西水灾：水中的茉莉花》 - 一个案追踪

2001年7月17日...广西水灾：水中的茉莉花 (2001.07.17 11:13:00) 今年6、7月份，我国南方接连受到台风的袭击，先是飞燕，然后是榴莲、尤特。这些有着动听名字的...

www.cctv.com/financial/jingji/sanjidiao/ ... 2011-3-17 - 百度快照

《好一朵美丽的茉莉花》背景资料-搜狐新闻

2007年11月28日...《好一朵美丽的茉莉花》背景资料世界经典民歌之一的《茉莉花》，是扬州民歌，还是苏州民歌？抑或六合民歌，曾引发了一场“属地”之争。...

news.sohu.com/20071128/t25365467.shtml 2011-3-12 - 百度快照

茉莉鲜花收购过程的质量管理 - 新农村商网

2009年4月22日...茉莉鲜花是一种质花，其芳香成份是随着鲜花的逐渐开放而不断形成和挥发的。茉莉花采摘时并未开放，含水分达60~80%，采摘后必须保持适当的温度、湿度...

nc.mofcom.gov.cn/news/P1P230781I8493635.html 2011-3-13 - 百度快照

送给妈妈的茉莉花 宋祖英 为人文子 中国青年网

送给妈妈的茉莉花 宋祖英 您所在的位置：为人文子频道 >> 演唱经典 http://www.youth.cn

2009-02-26 15:51:00 中国青年网 我要评论 送给妈妈的茉莉花 摘...

www.youth.cn/wrzny/cjcd/200902/t20090226_8 ... 2011-1-5 - 百度快照

华夏文化-芬芳迷人—音乐中的茉莉花

芬芳迷人—音乐中的茉莉花...茉莉花——是一种灌木，白色的小花，花香浓郁，花期长，花语为“纯洁、忠贞、爱”。茉莉花茶是中国传统的名茶之一，具有提神醒脑、消暑解渴的功效。

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&

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

▶ 来百度推广您的产品
咨询热线：400-800-8888
e.baidu.com

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 女子十二学步 茉莉花 娱乐频道
www.tudou.com/programs/view/ENsDwZy1hOM/ 2010-7-11 - 百度快照

茉莉花 免费在线听萨克斯名曲的新歌曲 茉莉花在线试听,MP3下载,茉...
萨克斯名曲的最新歌曲 茉莉花 是专辑萨克斯中收录的歌曲,本站提供萨克斯名曲 茉莉花 wma格式的MP3试听,并附带 茉莉花 文本歌词及LRC歌词,MoliHua
www.51t.com/ting/B5603672/3BB82130/ 2011-5-31 - 百度快照

《好一朵美丽的茉莉花》_mp3 - 百度MP3
帮助 视频 歌词 全部音乐 mp3 rm 其它格式 铃声 彩铃 把百度设为首页 抱歉，没有找到您选择的歌手或专辑信息。2010 Baidu..
mp3.baidu.com/albumlist/宋祖英;好一朵美丽 ... 2011-5-21 - 百度快照

茉莉花 -茉莉花典故|多种版本的茉莉花音乐|视频|曲谱|来历等
本页面介绍了茉莉花的经典故事、来历及详细介绍，并提供乐器演奏的茉莉花音乐的试听及下载，还有茉莉花视频、曲谱、知识文章等内容。
www.guqu.net/molihua 2011-4-18 - 百度快照

茉莉花论坛,横县网论坛-横县茉莉花论坛是来自中国茉莉之乡-广西横...
茉莉花论坛,横县论坛-横县茉莉花论坛是横县网主打论坛,是横县人的自己的论坛和网上综合信息生活社区
bbs.molihua.net/ 2011-6-27 - 百度快照

宋祖英mp3试听下载 好一朵美丽的茉莉花专辑曲目歌词 宋祖英专辑...
搜狗音乐大营 好一朵美丽的茉莉花 为您提供：宋祖英-好一朵美丽的茉莉花 专辑试听下载，好一朵美丽的茉莉花 打包下载、好一朵美丽的茉莉花 专辑曲目列表、歌词、mp3试听...
music.sogou.com/singer/ff/detailAlbum_宋 ... 2011-6-1 - 百度快照

茉莉花 百度百科
茉莉花，常绿小灌木或藤本状灌木，品种繁多，外形美丽，可用于花茶的制作。 中文学名：
茉莉花 拉丁学名： Jasminum sambac (Linn.) Aiton 别称： 茉莉 界： ...
baike.baidu.com/view/6903.htm 2011-6-17 - 百度快照

姚璎格 好一朵美丽的茉莉花 MP3下载,好一朵美丽的茉莉花歌曲试听...
歌曲姚璎格 好一朵美丽的茉莉花 Mp3在线试听，来自365音乐网。好一朵美丽的茉莉花为姚璎格演唱，2008-4-2 10:44:16入库，2008年03月03日收录于专辑《花好月圆》
www.yue365.com/play/7231/195182.shtml 2011-6-9 - 百度快照

Tuesday, March 29th

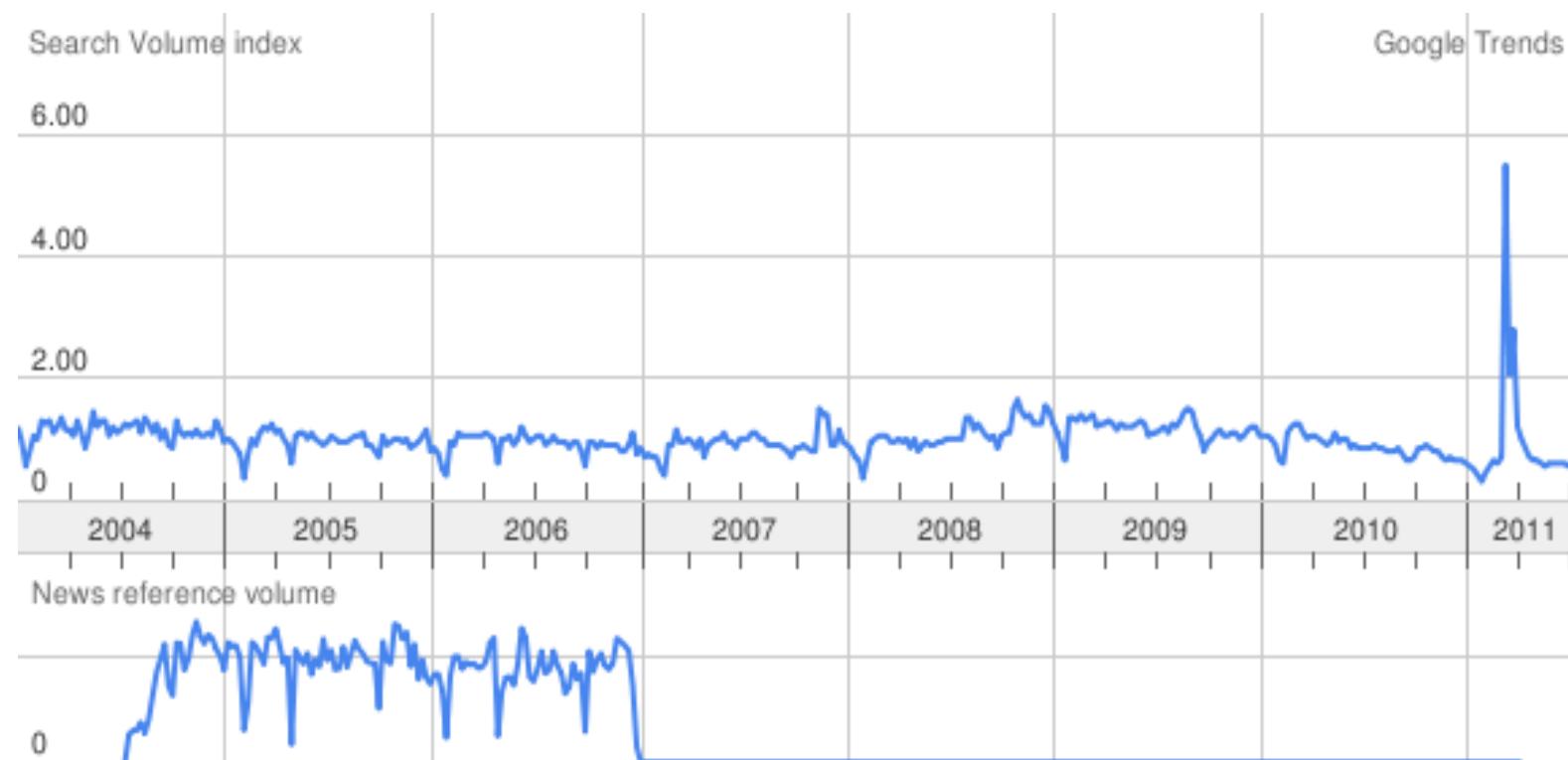
The screenshot shows a web browser window with the following details:

- Menu Bar:** File Edit View History Bookmarks Tools Help
- Address Bar:** file:///home/tony/ThuMar31/bi/biCenPageFromTue Mar 29 07:55:05 2011
- Toolbar:** Back, Forward, Stop, Home, Refresh
- Links:** Most Visited, Getting Started, Latest Headlines, Translate - MDBG C..., wiki, Java API
- Search Bar:** 百度搜索_的茉莉花
- Baidu Logo:** Baidu 百度
- Navigation Links:** 新闻, 网页, 贴吧, 知道, MP3, 图片, 视频, 地图, 更多
- Search Input:** 的茉莉花
- Search Button:** 百度一下
- Text Overlay:** 根据相关法律法规和政策，部分搜索结果未予显示。
- Text Overlay:** 茉莉花 电视资料 互动资料库 影音娱乐 新浪网-
- Text Overlay:** 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...
- Text Overlay:** ent.sina.com.cn/v/m/f/mlh/ 2011-3-19 - 百度快照
- Right Sidebar:** 来百度推广
咨询热线: 400-e.baidu.com

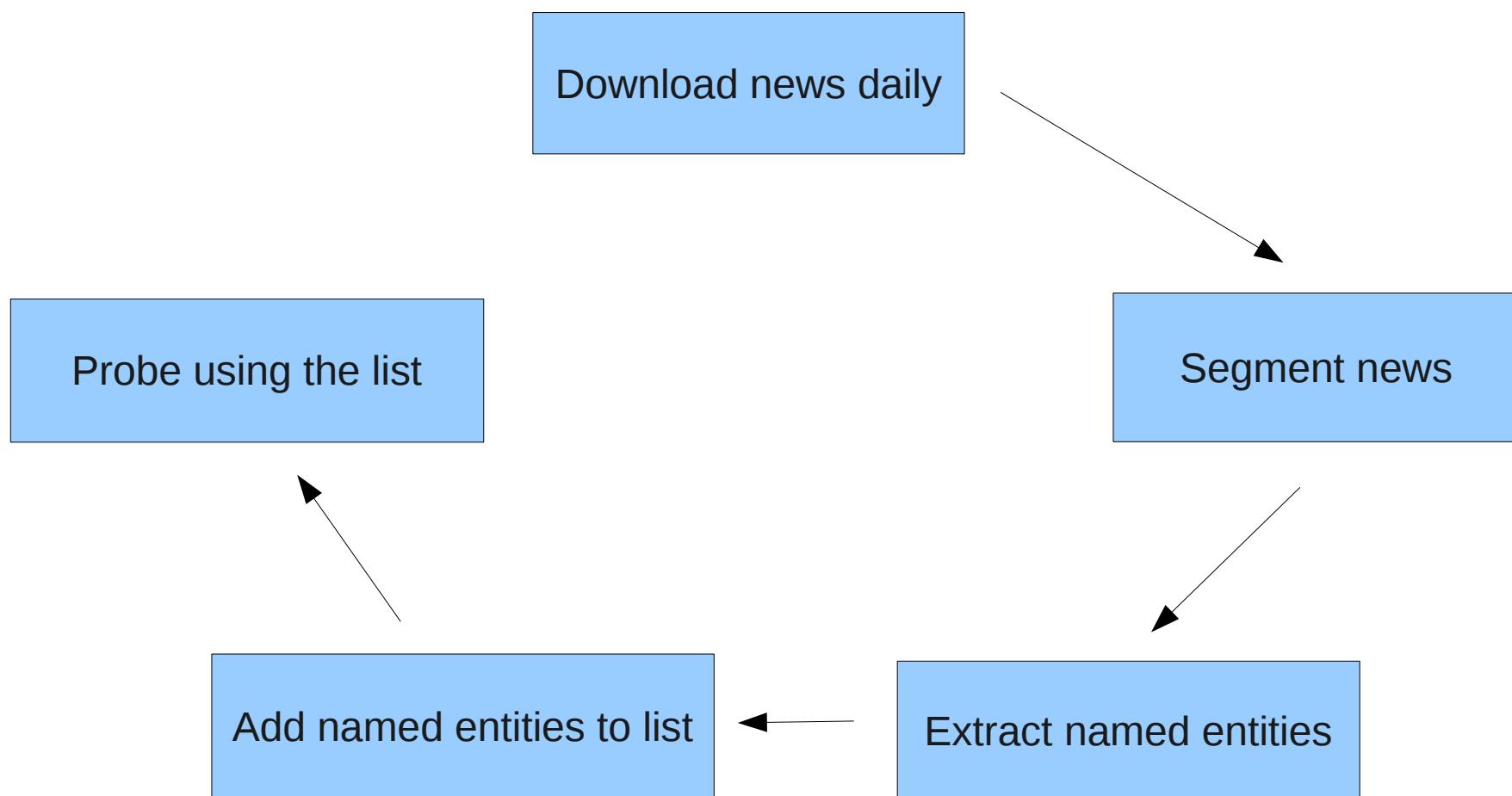
Tuesday, June 28th



辐射



High-level design



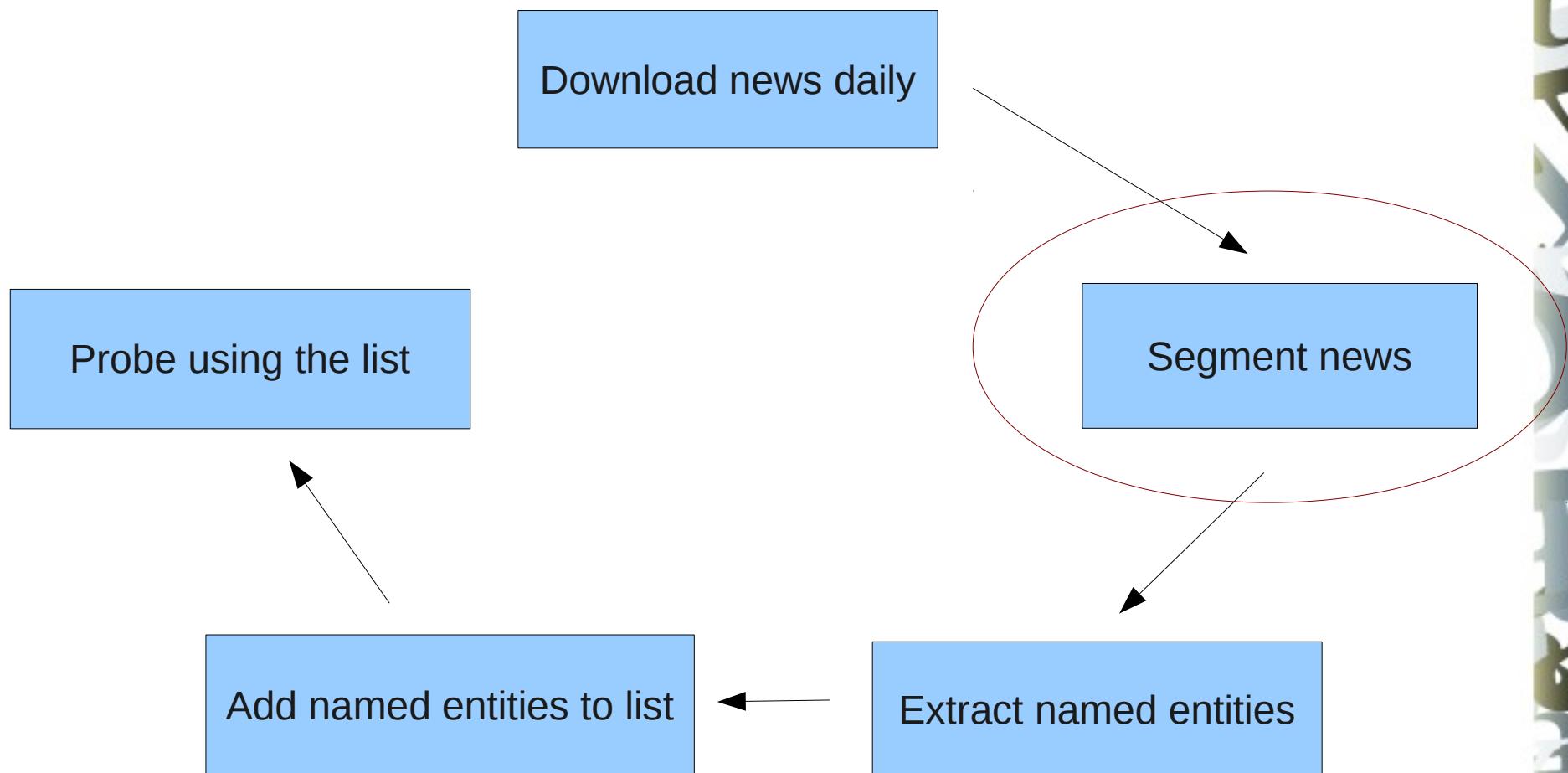
Sources

- <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

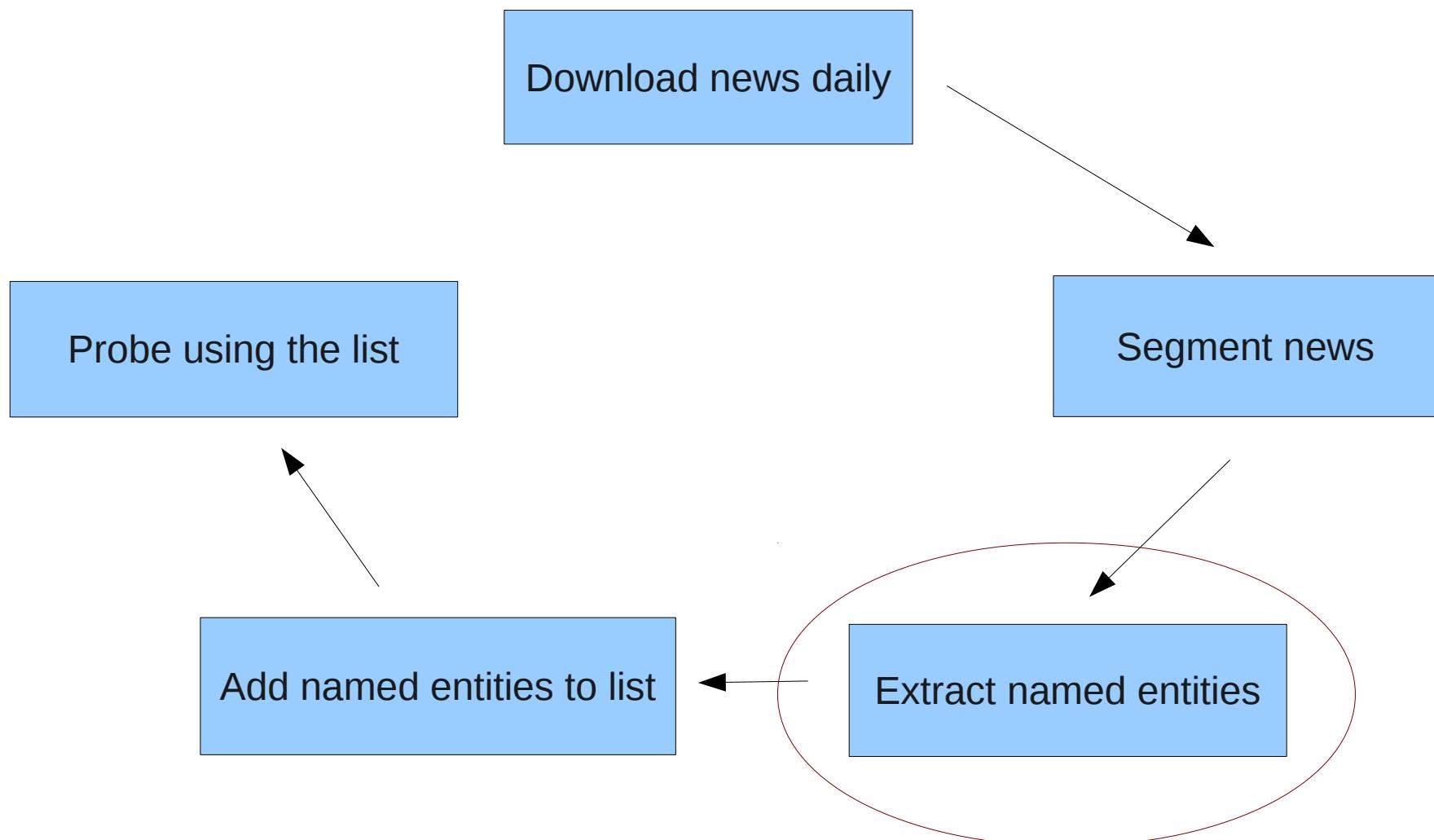


Segmentation

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 - 宜□ , - 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

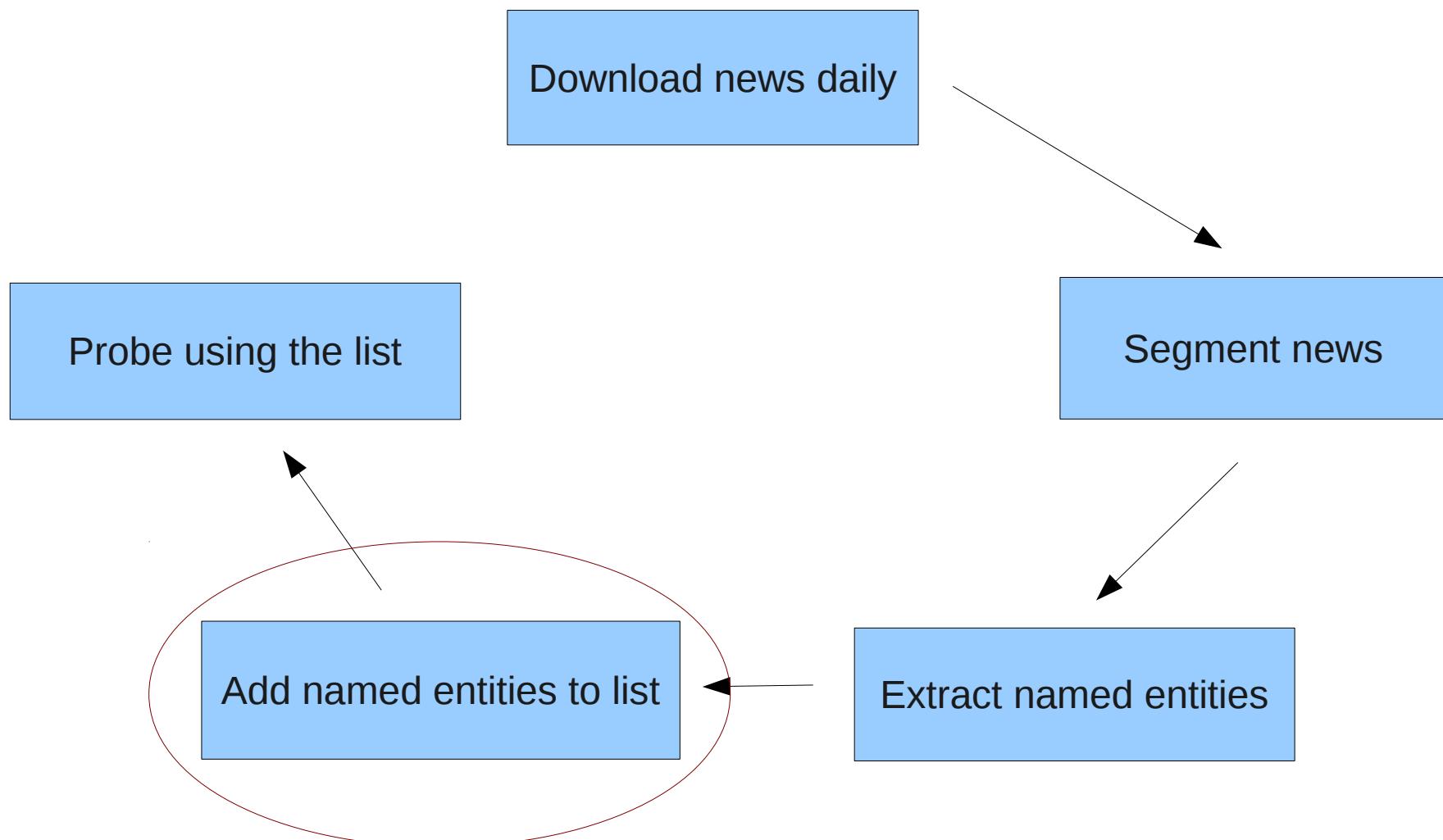
Named entity extraction

《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>

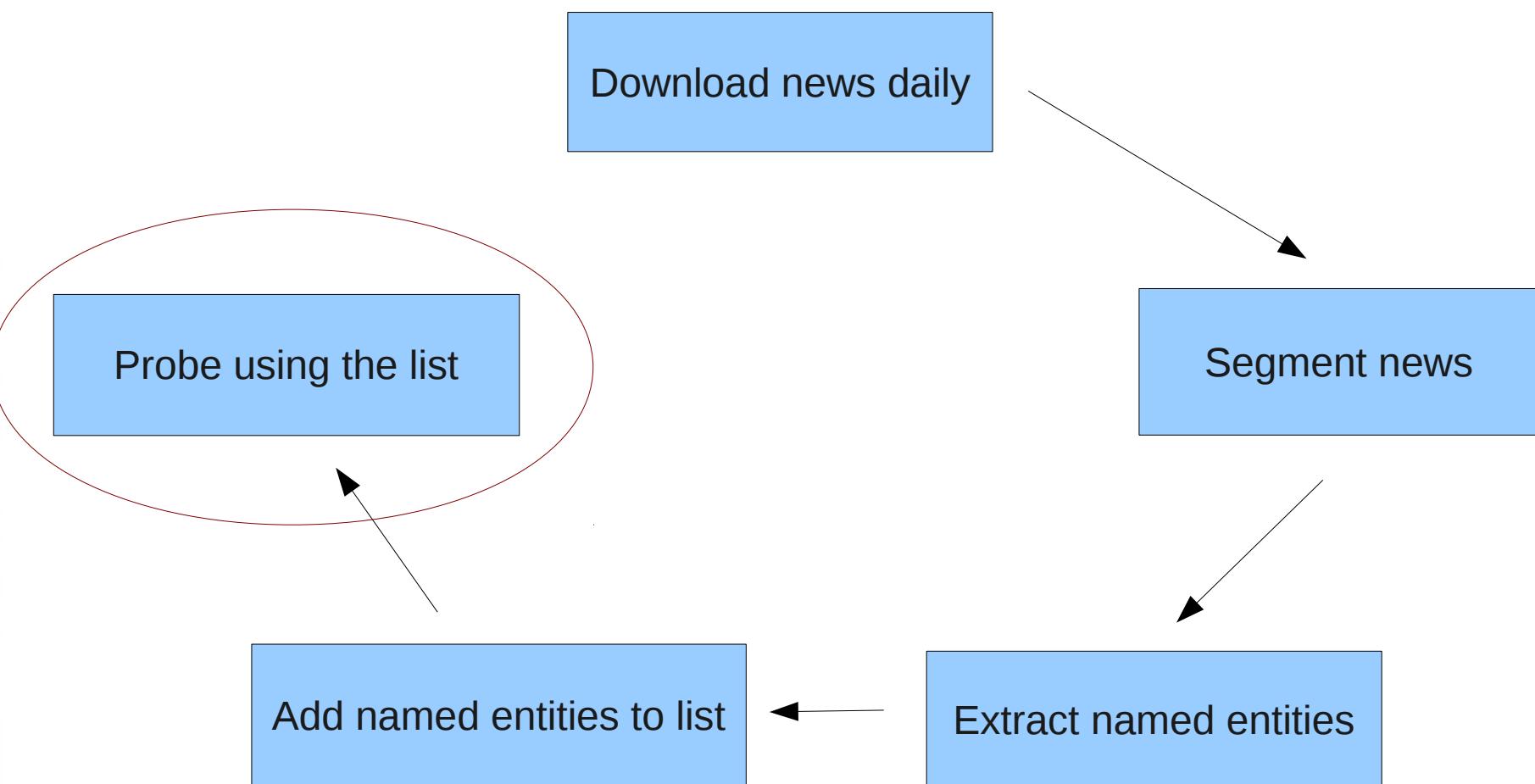
High-level design



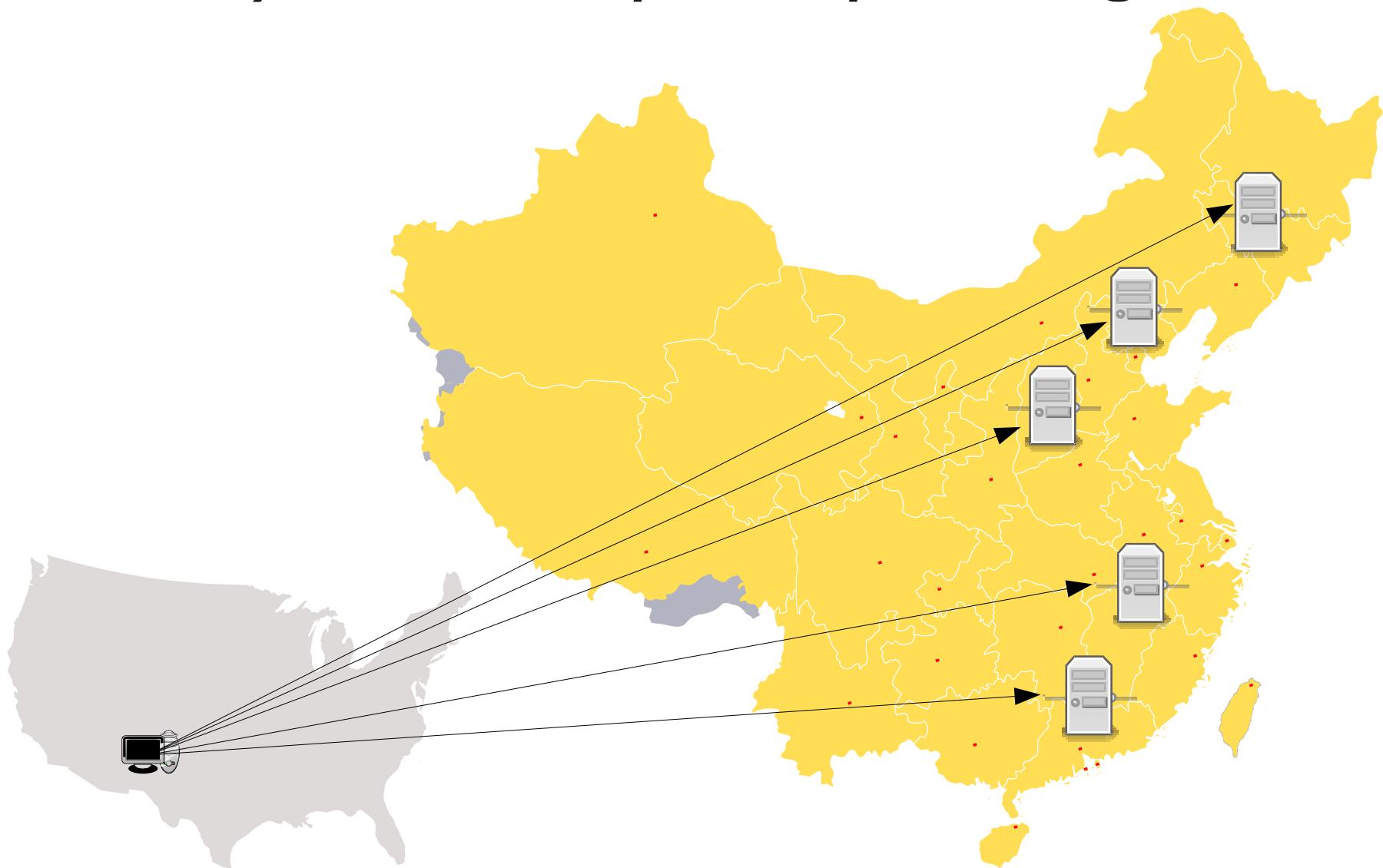
The list

- 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

 **The connection was reset**

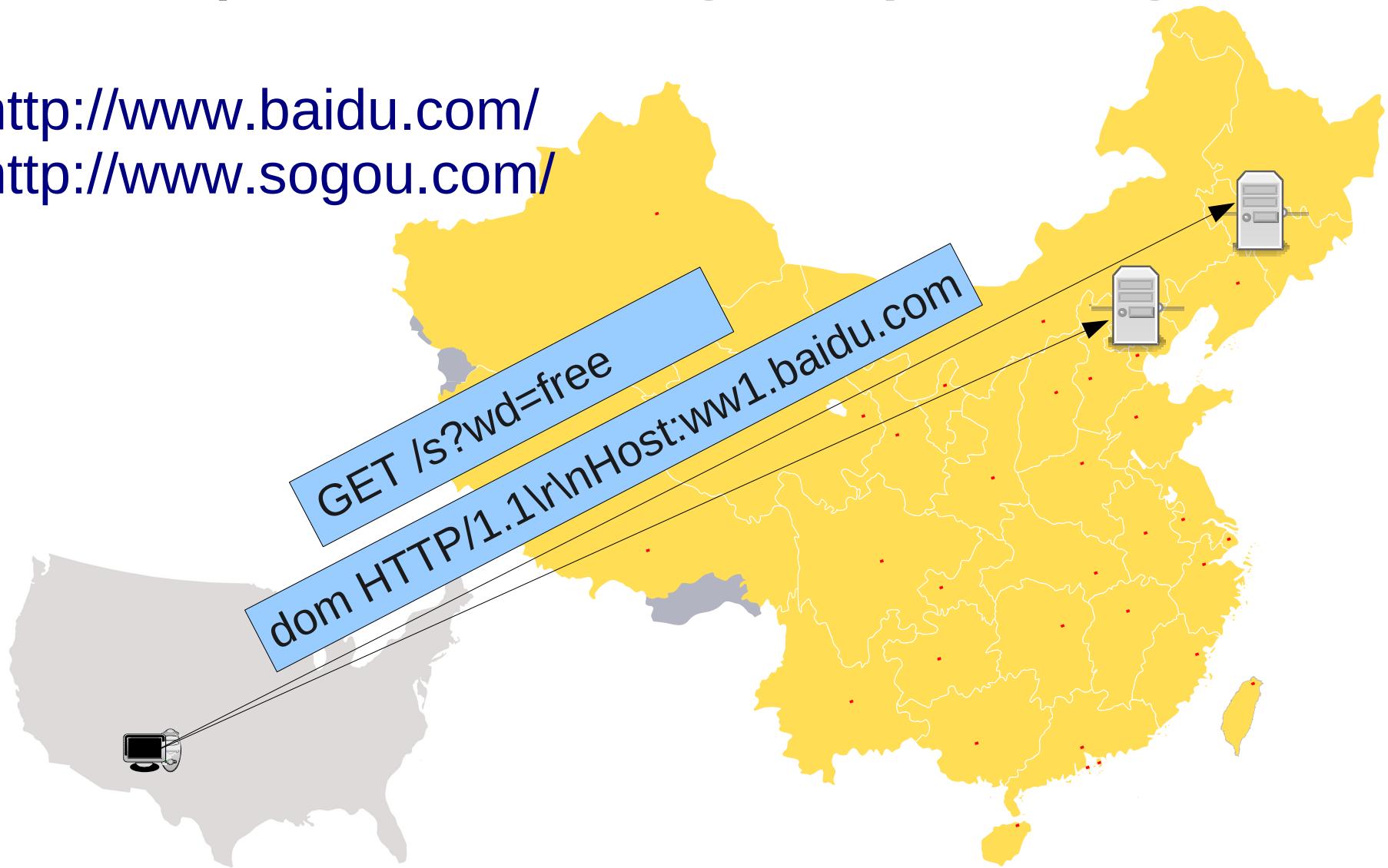
The connection to the server was reset while the page was loading.

- 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...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 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 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 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%

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
daijyuan	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?