



**BASIS**  
TECHNOLOGY

**GOVERNMENT**  
**USERS CONFERENCE**

JUNE 8-9  
2010  
CHANTILLY, VA

# Chinese Text Analysis

Joe Ho 何嘉栩

Principal Software Engineer  
Basis Technology Corp.



Human Language Technology From Arabia to Afghanistan

# Overview

- Introduction to Chinese language
- Processing Chinese text
- Software Solution for Chinese
  - Language Identifier
  - Chinese Script Converter
  - Chinese Language Analyzer
  - Rosette Entity Extractor
  - Chinese Name Matching

# Chinese vs. English - Script

- English: Latin-based, alphabets  
A .. Z
- Chinese: ideographic characters  
一 .. 龔

# Chinese vs. English – Character Set

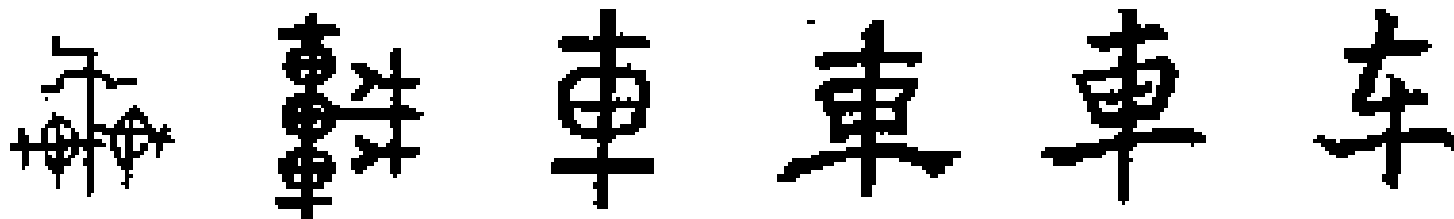
- English: ASCII, 26 letters
- Chinese: GB18030 in People's Republic of China (PRC), Big5 in Taiwan
- Simplified and Traditional Characters
- Chinese has about 49K characters, 4k is usually enough for common use

# Chinese vs. English – Word

- English:  
Words are separated by spaces
- Chinese:  
Words are **not** separated by spaces

# Introduction to Chinese Language

- Long history, first developed 4500 years ago
- Chinese character is ideographic - expresses meaning



# Introduction to Chinese Language

- Chinese character dictionary – lookup by pinyin, radical and stroke count
  - Radical: 心 Pinyin: xīn
  - Word: 志 Pinyin: zhì
- Writing – Each character is generally written left to right, top to bottom.

# Chinese Homonyms

- Reading – Many words have the same pronunciation. This is a problem when transcribing from sound to text, especially names. For example: Lín Fēng

林锋

林风

林峰

琳枫

琳丰

# Chinese Homonyms

- They all sound the same!
- Problem with large population (China!) e.g. school, military, bank account and city registry.

# Chinese Language Scripts

- Script: Simplified (简体) and Traditional (繁體)
  - Traditional: 國
  - Simplified: 国
- Simplified Chinese (SC) used in PRC and Singapore
- Traditional Chinese (TC) used in Taiwan and Hong Kong

# Chinese Language Scripts

- About 6764 characters are different between SC & TC
- The rest of the characters are the same
- Script conversion is required for information exchange
- Basis Technology RLP Chinese Script Converter handles script conversion

# Chinese Character Sets - PRC

- Chinese character set standard in PRC
  - GB2312 – Established in 1980
  - GBK – Microsoft Windows 95 and up
  - GB18030 – Established in 2000

# Chinese Character Sets - Taiwan

- Chinese character set standard in Taiwan
  - Big5

# Chinese Character Sets – Hong Kong

- Character set standard in Hong Kong SAR
  - HKSCS-2004

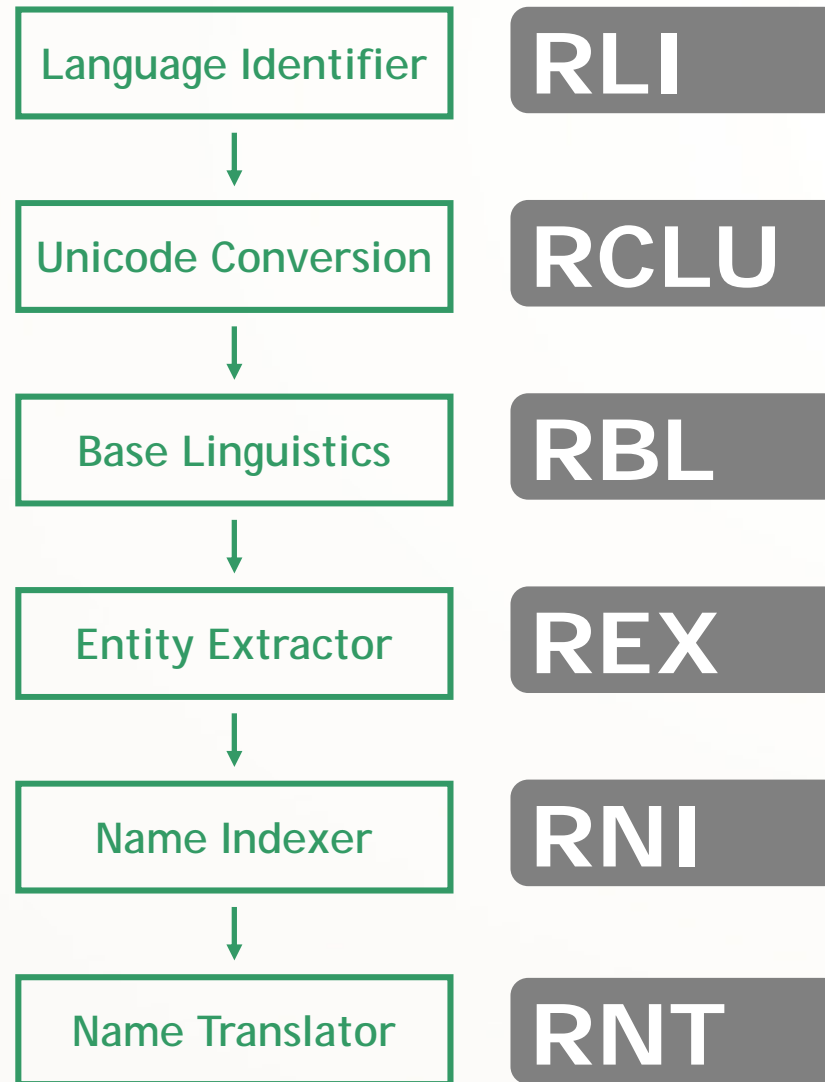
# Processing Chinese on the Internet

- Segmentation of Chinese text without spaces
- Simplified (SC) vs Traditional (TC)
- Handle different Chinese character sets in China, Taiwan and Hong Kong

# Processing Chinese on the Internet

- Gradual mixing of SC and TC text in Hong Kong and Guangdong province
- New/trendy words in SMS, blogs and e-mail. Mixing letters and numbers
- Chinese SMS: TMD, 7456, 今天GG、MM都上哪了, 一个也没来, 我只好也886。

# Rosette<sup>®</sup> Linguistics Platform (RLP)



# RLI identifies different encodings: GB2312

## Language

### Detected Language

Chinese, Simplified

## Rosette Language Identifier

Language	Chinese, Simplified
Script	Han (Simplified variant)
MIME Type	text/plain
Encoding	GB2312
Length	167

新华网日内瓦3月13日电（记者杨益）中国代表团副团长沈永祥在日内瓦举行的联合国人权理事会第七次会议上说，缅甸问题的根本解决主要依靠缅甸政府和人民的自身努力。联合国人权理事会第七次会议当天审议缅甸状况特别报告。中国代表团副团长说，政治稳定、经济发展、社会和谐、民主进步的缅甸不仅符合东南亚所有国家利益，也符合整个国际社会的利益。

# Language Boundary Locator

The New York Philharmonic Orchestra will make a historic trip to North Korea on February, it has announced. Dominique de Villepin a été nommé Premier ministre ce mardi en fin de matinée par Jacques Chirac. The orchestra's president and executive director, Zarin Mehta said it would play in the capital Pyongyang on February 26. In August, the reclusive communist country's Ministry of Culture sent an invitation to the orchestra at Lincoln Center in Manhattan.

朝鲜外务省发言人11月1日在平壤宣布，朝鲜将重返六方会谈，但前提条件是朝鲜与美国在六方会谈框架内讨论解除美国对朝鲜金融制裁问题。针对朝鲜方面的动向，各方均表示欢迎。美联社11月1日报道说：“长期以来一直拒绝与平壤进行直接对话的美国总统布什认为，各方达成一致、同意恢复六方会谈应归功于中国的斡旋。

L'ancien ministre de l'Intérieur, qui n'a jamais participé à une élection, a déjeuné avec les députés UMP et UDF à l'invitation du président de l'Assemblée nationale, Jean-Louis Debré.

In October, Mehta spent six days in North Korea exploring venues and other arrangements for a concert in Pyongyang.

# Language Boundary Locator

The New York Philharmonic Orchestra will make a historic trip to North Korea on February, it has announced. Dominique de Villepin a été nommé Premier ministre ce mardi en fin de matinée par Jacques Chirac. The orchestra's president and executive director, Zarin Mehta said it would play in the capital Pyongyang on February 26. In August, the reclusive communist country's Ministry of Culture sent an invitation to the orchestra at Lincoln Center in Manhattan.

朝鲜外务省发言人11月1日在平壤宣布，朝鲜将重返六方会谈，但前提条件是朝鲜与美国在六方会谈框架内讨论解除美国对朝鲜金融制裁问题。针对朝鲜方面的动向，各方均表示欢迎。美联社11月1日报道说：“长期以来一直拒绝与平壤进行直接对话的美国总统布什认为，各方达成一致、同意恢复六方会谈应归功于中国的斡旋。

L'ancien ministre de l'Intérieur, qui n'a jamais participé à une élection, a déjeuné avec les députés UMP et U du président de l'Assemblée nationale, Jean-Louis Borloo. In October, Mehta spent six days in North Korea and other arrangements for a concert in Pyongyang.

## Rosette Language Boundary Locator

Detected Languages (# characters)

Chinese, Simplified	150
English	483
French	285
<b>Total</b>	<b>918</b>

# Language Identifier detects Chinese scripts

國際在線消息（記者王鑫）：羅馬尼亞衛生部26日發表公報稱，罕見低溫天氣在過去兩天又導致該國8人死亡，羅全國凍死人數已上升至31人。羅經濟部25日下午宣布自即日起全國能源供應進入緊急狀態，部分耗能企業將暫時關閉，以確保居民用氣、採暖和熱水的正常供應。

新华网联合国 1 月 2 2 日电（记者 白洁 王湘江）第 6 4 届联合国大会 2 2 日一致通过决议，呼吁 1 9 2 个成员国尽快响应联合国发起的海地救援紧急募捐呼吁，强调各国应对联合国主导的救灾工作予以支持。

联大当天在纽约联合国总部就海地地震举行全体会议。第 6 4 届联大代理主席、哈萨克斯坦常驻联合国代表艾季莫娃在致辞中说，海地灾后的长期重建和发展工作需要国际社会在未来几个月甚至几年内长期关注。

# Language Identifier detects Chinese scripts

國際在線消息（記者王鑫）：羅馬尼亞衛生部26日發表公報稱，罕見低溫天氣在過去兩天又導致該國8人死亡，羅全國凍死人數已上升至31人。羅經濟部25日下午宣布自即日起全國能源供應進入緊急狀態，部分耗能企業將暫時關閉，以確保居民用氣、採暖和熱水的正常供應。

新华网联合国 1 月 2 2 日电（记者 白洁 王湘江）第 6 4 届联合国大会 2 2 日一致通过决议，呼吁 1 9 2 个成员国尽快响应联合国发起的海地救援紧急募捐呼吁，强调各国应对联合国主导的救灾工作予以支持。

联大当天在纽约联合国总部就海地地震举行全体会议。第 6 4 届联大代理主席、哈萨克斯坦常驻联合国代表艾季莫娃在致辞中说，海地灾后的长期重建和发展工作需要国际社会在未来几个月甚至几年内长期关注。

## Rosette Language Boundary Locator

Detected Languages (# characters)

Chinese, Simplified	193
Chinese, Traditional	128
<b>Total</b>	<b>321</b>

# Chinese Script Converter

《星島日報》報道，中信泰富集團主席榮智健動向成疑，公司更傳出會急售資產。不過，據接近榮智健的消息指出，榮智健對於今次公司犯下嚴重錯誤連累小股東深感歉意，但他無意在這個時候辭去主席一職，亦無意在市道低迷下賤賣資產。

《香港商報》報道，中國人民銀行行長周小川昨天指出，全球經濟明顯減速會影響到中國經濟發展，而國內經濟局勢也存在一些突出矛盾和問題。他說，央行將繼續調控利率，保持市場有足夠資金流動。周小川的講話暗示，中國貨幣政策在由「從緊」轉向「靈活審慎」之後，還有鬆動的餘地。

# Chinese Script Converter

《星島日報》報道，中信泰富集團主席榮智健動向成疑，公司更傳出會急售資產。不過，據接近榮智健的消息指出，榮智健對於今次公司犯下嚴重錯誤連累小股東深感歉意，但他無意在這個時候辭去主席一職，亦無意在市道低迷下賤賣資產。

《香港商報》報道，中國人民銀行行長速會影響到中國經濟發展，而國內經濟他說，央行將繼續調控利率，保持市場示，中國貨幣政策在由「從緊」轉向地。

#	Original Text	Simplified
1	《	《
2	星島	星島
3	日報	日报
4	》	》
5	報道	报道
6	,	,
7	中信泰富	中信泰富
8	集團	集团
9	主席	主席
10	榮	荣
11	智健	智健

# Rosette Base Linguistics Chinese

- Segments Chinese text into words
- Chinese text does not contain spaces
- Words are harder to identify, index and search

# Rosette Base Linguistics Chinese

- Segmentation is the first step for Text Processing, Machine Translation, Text-to-Speech and Linguistic Analysis.
- Segmentation Standard GB13715  
中国国家标准GB13715 “信息处理用现代汉语分词规范”

# Segments Chinese text into words

- Input text

新华社北京5月30日电 国家主席胡锦涛30日上午在人民大会堂接受了上海合作组织成员国哈萨克斯坦、吉尔吉斯斯坦、俄罗斯、塔吉克斯坦、乌兹别克斯坦和中国记者的联合采访。

- Output text

新华社/北京/5月/30日/电/国家/主席/胡  
/锦涛/30日/上午/在/人民/大会堂/接受/  
了/上海/合作/组织/成员国/哈萨克斯坦//  
吉尔吉斯斯坦//俄罗斯//塔吉克斯坦//  
/乌兹别克斯坦/和/中国/记者/的/联合/采访/  
/。

# Word Breaking for Better Search

- Searching for 学生 (Student)
- Result: 北京大学生物系 (Peking University Department of Biology)
- Why are we getting this result?
- Bigram tokenization:
  - 北京/京大/大学/学生/生物/物系
  - *False Positive*
- RBL-Chinese Tokenization:
  - 北京大学/生物系
- RBL-Chinese Decomponding:
  - 北京大学 => 北京/大学  
(Beijing/University)

# Rosette Base Linguistics Chinese

中新网保定3月15日报道(记者张永利)

回族人信奉伊斯兰教。伊斯兰教在东南亚等地区形成过程中曾起过重要作用。他们多数从事农业,有的兼营牧业、手工业。伊斯兰教教徒生活在全国各地。回族人爱整洁、爱鲜花,这种性格在大多数人家的院落中就能体现出来。各地回族人都有喜欢饮茶习惯,其他民族兄弟到家里作客,会深深感到他们好客大方的性格。

在马来西亚和文莱,作为国教的伊斯兰教已经深深地。马来西亚的回教党已经成为国内最大的反对党,国家伊斯兰教选民和教界是任何政党竞选时必须争取的对

## Rosette Base Linguistics

### Part of Speech

A	adjective
D	adverb
EOS	sentence final punctuation
J	conjunction
NC	common noun
NP	proper noun
NR	pronoun
NT	temporal noun
PL	particle
PR	preposition
PUNCT	non-sentence-final punctuation
V	verb
W	derivational suffix
WL	direction word

# Rosette Entity Extractor (REX)

- What is an entity? Person, Organization, Location, Date, Time, Religion and Title, etc.
- Name, e.g. 胡锦涛, 奥巴马
- Location, e.g. 北京, 上海, 美国, 华盛顿

# Rosette Entity Extractor (REX)

- Organization: 微软, 联合国
- Temporal – Date: 2010年6月8日
- REX uses statistical modeling to learn patterns from large Chinese corpora
- Language model is built into REX, will recognize new entities in text

# RLP Entity Extraction

中新网保定3月15日报道(记者张永利)

回族人信奉伊斯兰教。伊斯兰教在东南亚等地区形成过程中曾起过重要作用。他们多数从事农业,有的兼营牧业、手工业。伊斯兰教教徒生活在全国各地。回族人爱整洁、爱鲜花,这种性格在大多数人家的院落中就能体现出来。各地回族人都有喜欢饮茶习惯,其他民族兄弟到家里作客,会深深感到他们好客大方的性格。

在马来西亚和文莱,作为国教的伊斯兰教已经深深地打上了政治的烙印。马来西亚的回教党已经成为国内最大的反对党,国家总统由回族人担任,伊斯兰教选民和教界是任何政党竞选时必须争取的对象。

## Rosette Entity Extractor

Named Entity (# instances)

LOCATION	5
ORGANIZATION	1
PERSON	1
RELIGION	5
TEMPORAL:DATE	1
TITLE	2

# Chinese Names

- What is a Chinese Name
  - Surname, then given name
  - 黄满池 is Huang Man Chi (黄 surname, 满池 given name)
  - In PRC standard: Huang Manchi (note given name is one Latin word)
  - In English given name, surname order: Manchi Huang

# Challenges in Chinese Name Matching

- Name may have different Latin transliteration depends on locale.
  - 黃 – Huang (PRC, Taiwan), Wong (Hong Kong)
  - 陳 – Chen (PRC, Taiwan), Chan (Hong Kong), Tan (Singapore)
  - 周 – Zhou, Chow, Chau, Chou
  - 張 – Zhang, Chang, Cheung, Teo, Teoh

# Chinese Name – Variations

- Chinese can have western names, for example, the first lady of Taiwan: 周美青 Christine Chow Mei-ching.
- Traditionally, Chinese women retain their maiden name as surname after marriage
- However, some add the husband's surname to their maiden name: Christine Chow Ma, whose husband is President Ma Ying-jeou of Taiwan.
- In Hong Kong, e.g. 范徐丽泰, Fan Hsu Lai-tai, President of the Legislative Council of Hong Kong, 1998-2008. Her maiden name is 徐丽泰

# Multilingual Chinese Name Matching

- Possible matches of 黄满池 in Latin script
  - Huang Man Chi
  - Wong Mun Chi
  - Wang Man Chi
  - Wong Moon Chi
  - Wong Mun Chee

# Multilingual Chinese Name Matching

- Rosette Name Translator (RNT) and Name Indexer (RNI) allow cross-script name translation and matching.

# Rosette Product in Action



Basic Search

Rosette Linguistics Platform

Rosette Name Indexer

Basic tokenizers (stemming)

Search



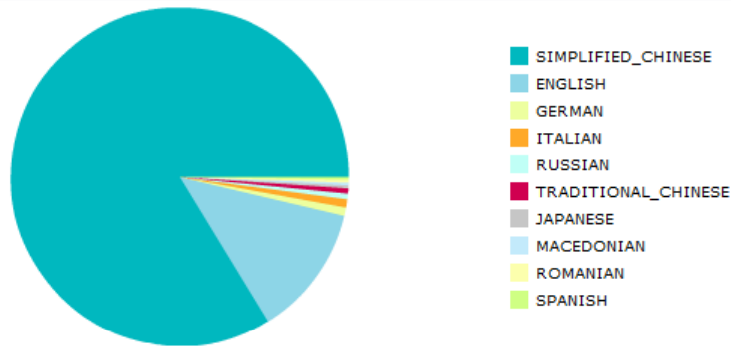
[Home](#)

[Relationship Map](#)

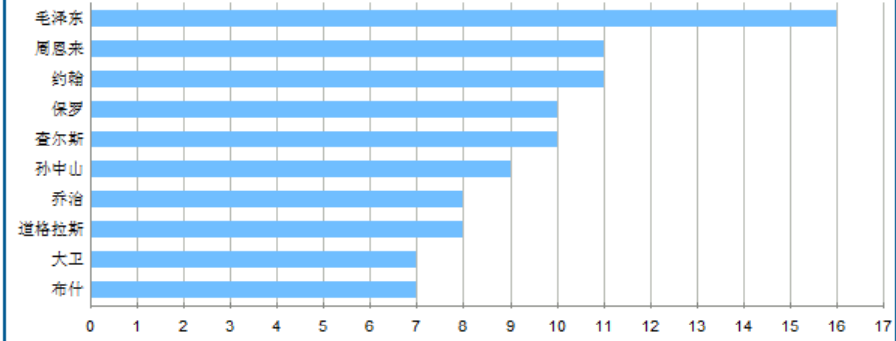
[Administration](#)

[File Viewer](#)

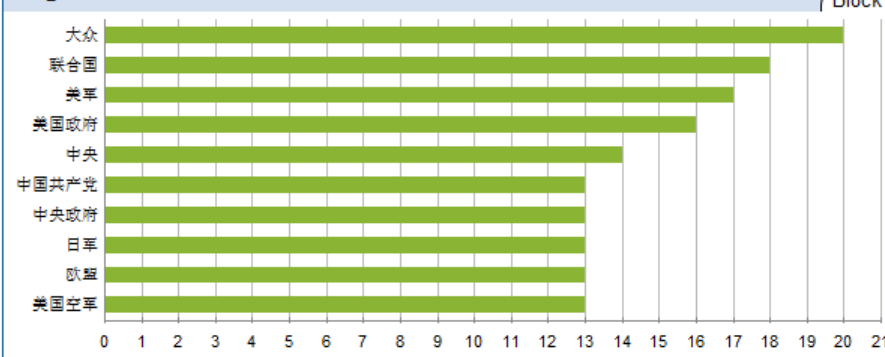
### Language Statistics



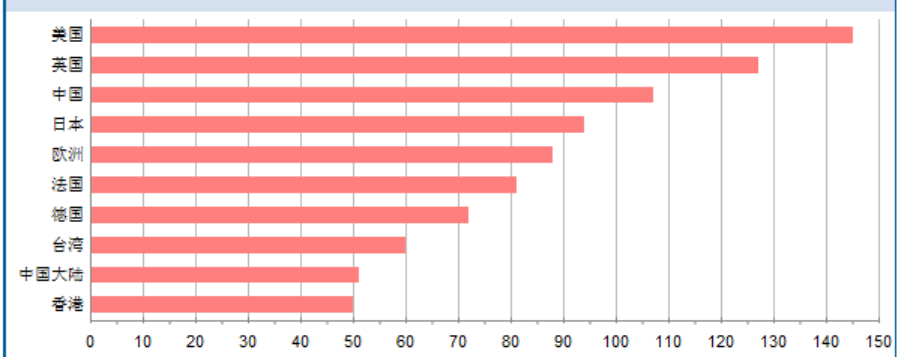
### Person Statistics



### Organization Statistics



### Location Statistics



# Chinese Script Converter for cross-



Basic Search

Rosette Linguistics Platform

Rosette Name Indexer

Rosette Linguistics Platform ( language identification, stemming and  
lemmatization )

胡锦涛

Search



Home > Search for [ 胡锦涛 ]

Filter Results By	
<b>Language</b>	
SIMPLIFIED_CHINESE	87
ENGLISH	1
GERMAN	1
<b>Person</b>	
胡锦涛	87
江泽民	23
温家宝	22
毛泽东	20
邓小平	17
<b>Location</b>	
中国	65
中华人民共和国	48
北京	37
中国大陆	34
美国	33
<b>Organization</b>	
中国共产党	33
中共	18
中共中央	17
中共中央政治局	16
中国政府	16

Results 1 - 10 of 87

Page: 1 of 9

## [刘永清.](#)

刘永清. 刘永清, 中国主席胡锦涛的夫人, 中国重庆人。曾担任北京市城乡规划委员会副主任。毕业于重庆巴蜀中学, 后考入清华大学水利系1959年级学习, 是胡锦涛大学期间同班同学。刘永清当年

2010-01-29T23:53:47.383Z

## [两岸一中.](#)

两岸一中. 两岸一中, 出现于2005年, 台湾的亲国民党主席宋楚瑜访问中国大陆, 进行搭桥之旅时。宋楚瑜和胡锦涛会面之后, 双方提出「两岸一中」的主张, 就是两岸同属一个中国之意。事实上宋楚瑜

2010-01-29T23:03:29.122Z

## [1989年拉萨事件.](#)

人民解放军部队以武力镇压的事件。这次事件中死伤人数不明。据林和立在《胡锦涛时代的中国政治: 新领袖, 新挑战》一书中描写, 尽管时任西藏自治区党委书记的胡锦涛向中共中央申请了戒严, 但是之后几天他并没有

2010-01-29T22:55:00.501Z

## [2006中国意大利年.](#)

2006中国意大利年. 中意文化年, 2006年是中意文化年。2004年12月, 意大利总统钱皮在访华期间与胡锦涛主席共同宣布举办“2006中国意大利年”活动。

2010-01-29T23:35:14.577Z

## [国际奥林匹克委员会第120次全会.](#)

国际奥林匹克委员会第120次全会. 国际奥林匹克委员会第120次全会于2008年8月5日至8月7日在北京举行。开幕式4日晚在中国国家大剧院举行。胡锦涛出席开幕式并发表致辞。国际奥林匹克

2010-01-30T05:24:03.13Z

# Rosette Name Indexer search: Hu



Basic Search

Rosette Linguistics Platform

Rosette Name Indexer

Rosette Name Indexer ( cross-language fuzzy name search )

Hu Jintao

Search



[Home](#) > Search for [ Hu Jintao ]

Filter Results By	
<b>Language</b>	
SIMPLIFIED_CHINESE	109
ENGLISH	7
TRADITIONAL_CHINESE	2
GERMAN	1
<b>Person</b>	
胡锦涛	87
江泽民	23
温家宝	22
毛泽东	20
邓小平	18
<b>Location</b>	
中国	80
中华人民共和国	52
北京	46
美国	43
中国大陆	39
<b>Organization</b>	
中国共产党	33
中共	22
中国政府	20
中共中央	18
中共中央政治局	17

Results 1 - 10 of 114

Page: 1 of 12

## [Chinese President inspects reconstruction in quake-hit Shaanxi Chinese President Hu J...](#)

Chinese President inspects reconstruction in quake-hit Shaanxi Chinese President **Hu Jintao**

2010-01-29T22:58:54.648Z

## [Chinese president wraps up Central Asia trip Chinese President Hu Jintao wrapped up h...](#)

Chinese president wraps up Central Asia trip Chinese President **Hu Jintao** wrapped up his tour of two

2010-01-29T22:58:54.101Z

## [Hu attends inauguration of China-Central Asia gas pipeline Chinese President Hu Jinta...](#)

**Hu** attends inauguration of China-Central Asia gas pipeline Chinese President **Hu Jintao** and his

2010-01-29T22:58:54.336Z

## [刘永清.](#)

刘永清. 刘永清, 中国主席胡锦涛的夫人, 中国重庆人。曾担任北京市城乡规划委员会副主任。毕业于重庆巴蜀中学, 后考入清华大学水利系1959年级学习, 是胡锦涛大学期间同班同学。刘永清当年

2010-01-29T23:53:47.383Z

## [两岸一中.](#)

两岸一中. 两岸一中, 出现于2005年, 台湾的亲国民党主席宋楚瑜访问中国大陆, 进行搭桥之旅时。宋楚瑜和胡锦涛会面之后, 双方提出「两岸一中」的主张, 就是两岸同属一个中国之意。事实上宋楚瑜

2010-01-29T23:03:29.122Z

## [1989年拉萨事件.](#)

人民解放军部队以武力镇压的事件。这次事件中死伤人数不明。据林和立在《胡锦涛时代的中国政治: 新领袖, 新挑战》一书中描写。尽管时任西藏自治区党委书记的胡锦涛向中共中央申请了戒严, 但是之后几天他并没有

# Questions and Answers

Thank you for coming to learn about  
our technology and products.

Questions?

More information:

<http://www.basistech.com>  
[productinquiries@basistech.com](mailto:productinquiries@basistech.com)



**BASIS**  
TECHNOLOGY

**GOVERNMENT**  
**USERS CONFERENCE**

JUNE 8-9  
2010  
CHANTILLY, VA

Thank You!

谢谢



Human Language Technology From Arabia to Afghanistan