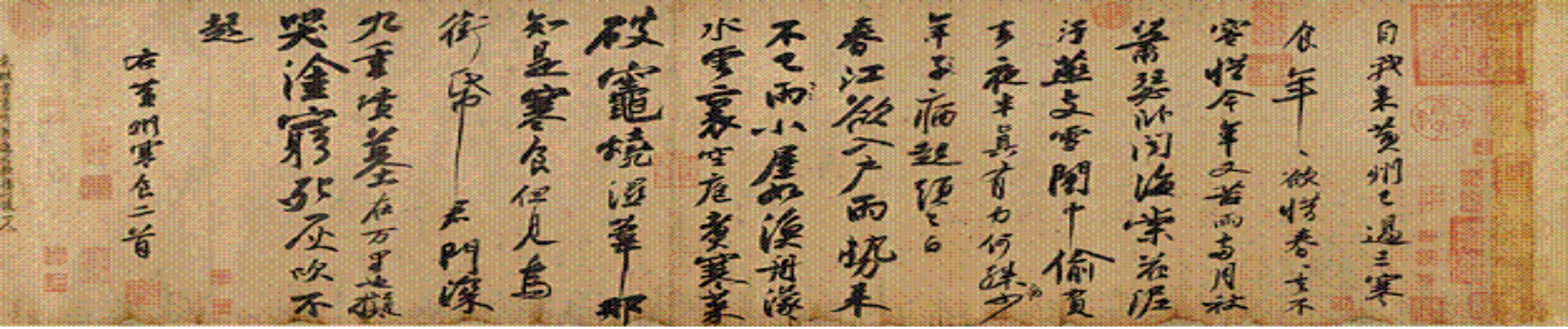


计算语言学
Computational Linguistics

教师：孙茂松

Tel:62781286

Email:sms@tsinghua.edu.cn

TA：**林衍凯**

Email:linyankai423@qq.com

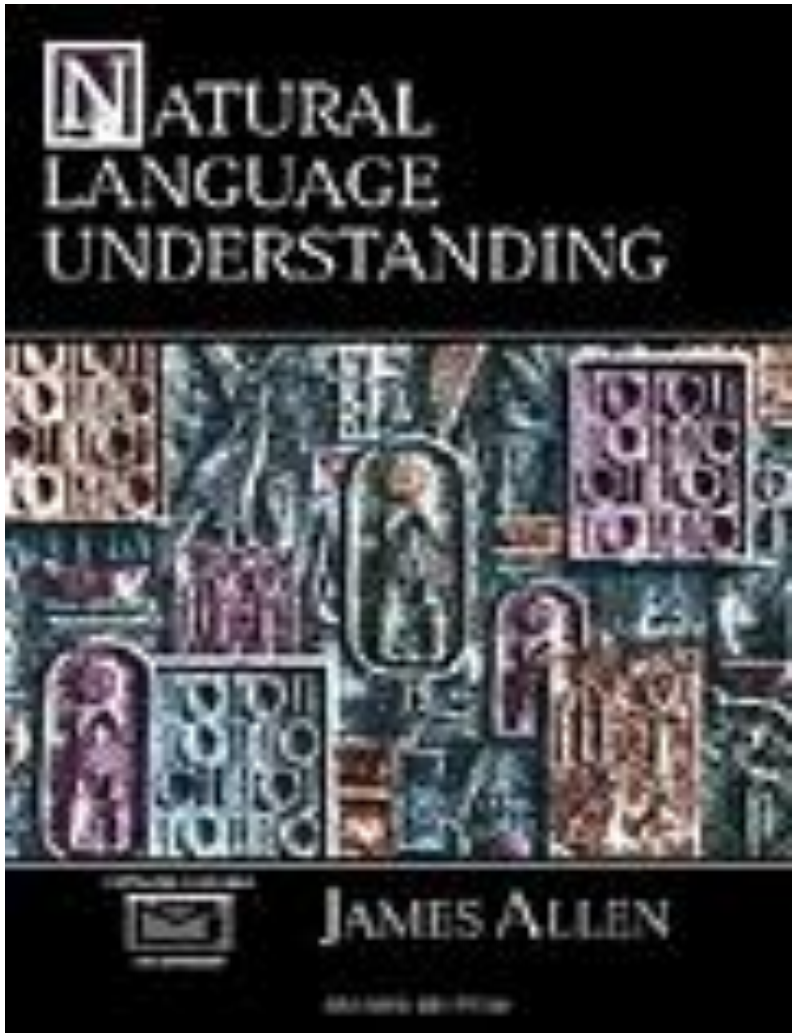
郑重声明

- 此课件仅供选修清华大学计算机系研究生课《计算语言学》(70240052)的学生个人学习使用，所以只允许学生将其下载、存贮在自己的电脑中。未经孙茂松本人同意，任何人不得以任何方式扩散之（包括不得放到任何服务器上）。否则，由此可能引起的一切涉及知识产权的法律责任，概由该人负责。
- 此课件仅限孙茂松本人讲课使用。除孙茂松本人外，凡授课过程中，PTT文件显示此《郑重声明》之情形，即为侵权使用。

主要参考书

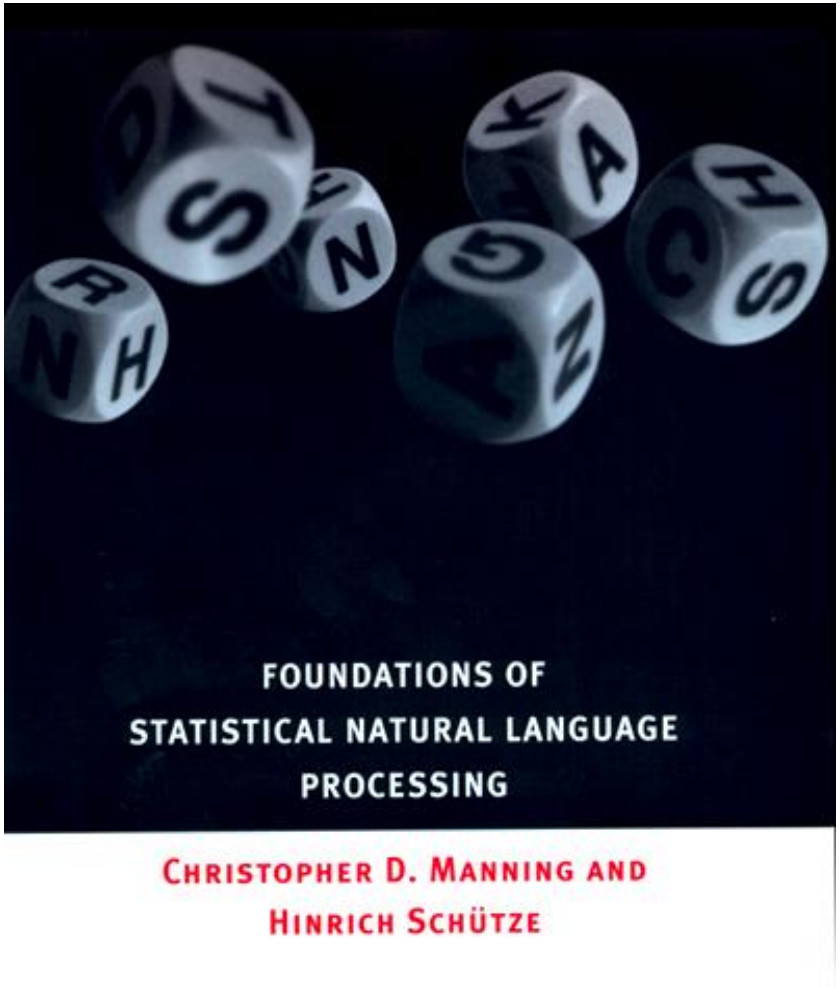
http://www.uni-giessen.de/~g91062/Seminare/gk-cl/Allen95/al1995co.htm

James Allen, Natural Language Understanding (2nd edition), *The Benjamin/Cummings Publishing Company, Inc.* 1994



主要参考书

http://nlp.stanford.edu/fsnlp/
Chris Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, MA: May 1999.



要求

- 成绩：编程作业(+DEMO) + seminar + 笔头作业 + 不定期的课堂小测验

  对CL不感兴趣的同学，请不要选。

  作业未按时交，该次成绩以零分计（一般电子版交）
- 不分组：(一人一组)

纪律：

- 鼓励讨论，严禁抄袭\拷贝：

  第一次发现，成绩记0分，并警告；

  第二次发现，整个课程不通过。

  （以上均无论抄与被抄者）。
- 不得迟到，早退，不得吃东西。关手机。

第一章 引言
(Part 1)

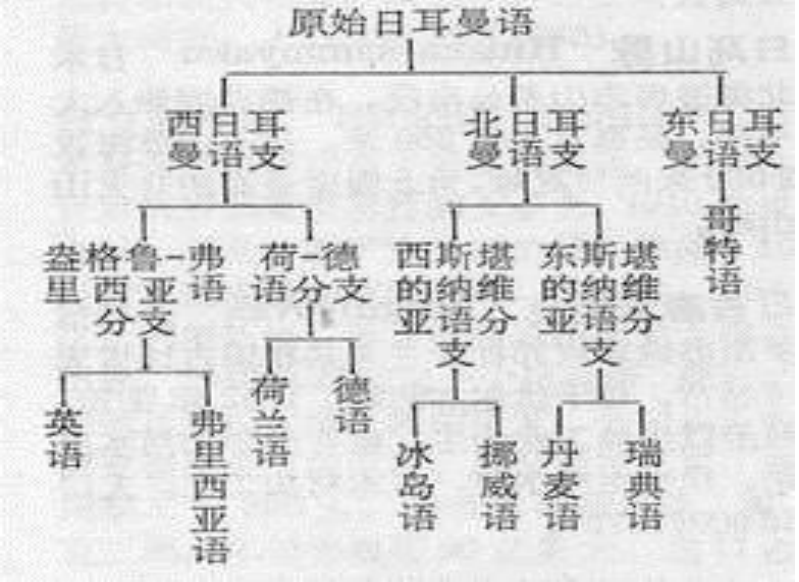
1.1. What is Computational Linguistics

- Object: natural language 6000种（3000-10000种不等）
- The role of language:

  * the most natural means of communication between humans

The tower of Babel是《圣经》之《旧约 创世纪》第11章（1-9）中一座通天塔，由挪亚（Noa）的后代在示拿(Shinnar)所建

ATLANTIC
OCEAN
germanic
baltic
slavic
celtic
romance
armenian
iranian
BURUSHASKI
indic
BASQUE
albanian
greek
NAHALI
INDIAN
OCEAN


原始日耳曼语
西日耳曼语支
北日耳曼语支
东日耳曼语支
哥特语
盎格鲁-弗里西亚支
荷-德分支
西斯堪维纳亚的语支
东斯堪维纳亚的语支
英语
弗里西亚语
荷兰语
德语
冰岛语
挪威语
丹麦语
瑞典语

阿尔泰山脉
天山山脉
大兴安岭
小兴安岭
长白山脉
阴山山脉
祁连山脉
昆仑山脉
喜马拉雅山脉
横断山脉
秦岭
大巴山脉
南岭
地图教学网
www.ditu.cn

1.1. What is Computational Linguistics

- The role of language:

  * record of knowledge

    “搜索”产生的权力 http://www.sina.com.cn 2006年02月17日

  Google发现宝马德国网站通过“技术作弊”人为地提高它在其搜索引擎中的排名，涉嫌欺骗用户，便在引擎中删除宝马汽车的德国网站。而宝马则认为Google没有事先通知就采取公开批评和封杀的行为有违常规，并同时与Google进行联系，希望能尽快解决。

  ，“由于Google制定了信息社会在未来一个阶段的游戏规则，所有人在信息时代寻找一席之地的时候，都必须遵守Google的游戏规则”。 人们向Google索求信息，却不觉间投票给了它很大的支配性权力。

  * 语言与思维

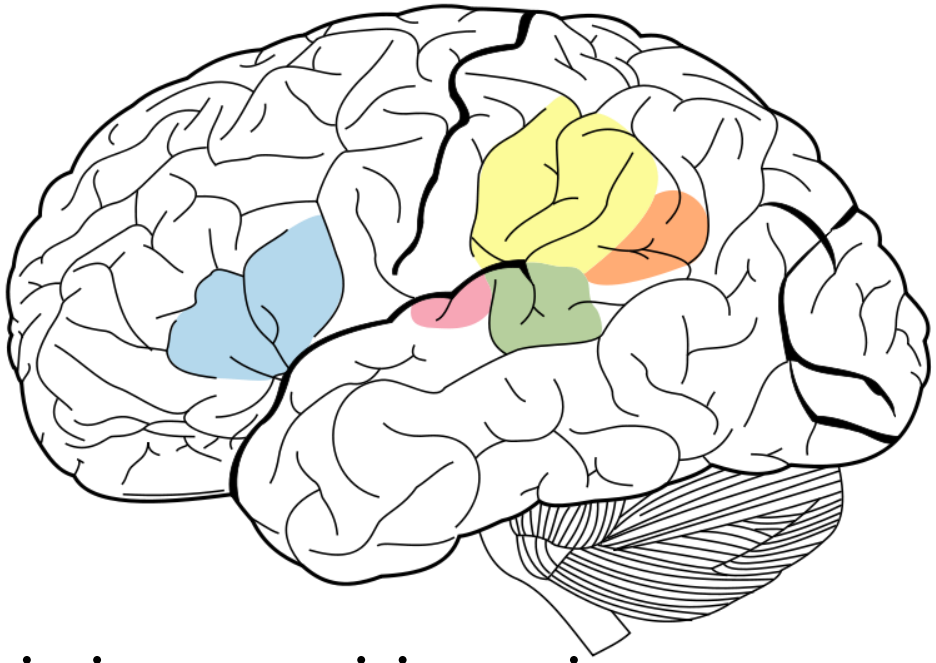
1.1. What is Computational Linguistics

- Linguistics

  Psycholinguistics

  Philosophy

  Computational Linguistics



- Academic disciplines:

  linguistics, psycholinguistics, cognitive science, computer science, …

  Linguistics + Computation (computer)

1.1. What is Computational Linguistics

- The task: to create computational models of language
  * scientific motivation
    the nature of linguistic communication
    validation (testing) of linguistic models
  * practical motivation
    Most of human knowledge is recoded in linguistic form,…: capability of intelligence for machines
    NLP capabilities would revolutionize the way computers are used
    Effective human-machine communication (too complex to everyone now!）

1.2. Applications of CL

- information retrieval, extraction, filtering, classification and summarization, Search engine, digital library, e-commerce etc.

query: keywords

* English: *bank*
* Chinese word segmentation:

和服 | 务 | 于三日后裁制完毕。
(kimono must)
这个酒店的设施 | 和 | 服务 | 是一流的。
(and service)

* unknown word processing:

高海燕 (storm petrel)

Google搜索："和服" - ★欢迎访问■娱乐网★ ★http://flash1968.51.net★

Google

高级搜索 使用偏好 语言工具 搜索建议

"和服" Google搜索

◉ 搜索所有网站 ○ 搜索所有中文网页 ○ 搜索中文(简体)网页

所有网站 图像 新闻群组 网页目录

已向英特网搜索"和服". 共约有690,000项查询结果，这是第21-30项 。 搜索用时0.25秒。

和服見聞録
□□□□□□□、和服□関□□話題□□□□□広□□□□□
□□□掲載□□□□□□思□□□。 □□□□□□□面白□□
便□□掲載□□□□思□□□□□、□□□□□□□。 /B ...
www.kimono-taizen.com/watch.htm - 11k - 网页快照 - 类似网页
[www.kimono-taizen.com站内的其它相关信息]

Yahoo! JAPAN - 芸術□人文 > □□□□□□□ > ...
□□□ > 芸術□人文 > □□□□□□□ > □□□□□□、服飾 > 和服、□□
□, □□□□推薦□依頼. 全検索 □□□□□□以下□□検索, Yahoo!□□□□
□. ...
www.yahoo.co.jp/Arts/Design_Arts/Fashion/Kimono/ - 13k - 网页快照 - 类似网页

代表团手册 - 为代表团提供的设施和服务
www.un.org/chinese/ga/55/facilit.htm - 21k - 网页快照 - 类似网页

代表团手册 - 为代表团提供的设施和服务
... 五. 为代表团提供的设施和服务. ...

文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹 媒体
地址(D) http://www.google.com/search?hl=zh-CN&newwindow=1&q=%E5%92%8C%E6%9C%8D&btnG=%E6%90%9C%E7%B4%A2&lr=
转到 链接

登录

Google

网页 图片 资讯 论坛 更多 »

和服 搜索 高级搜索 使用偏好

搜索所有网页 搜索所有中文网页 搜索简体中文网页

网页

约有**2,000,000**项符合**和服**的查询结果，以下是第**1-10**项。（搜索用时 **0.03** 秒）

建议： 可直接输入拼音， Google会自动提示最符合的中文关键词 [细节]

和服的穿法

江户时代以前，和服的“浴”字是指“热水”，那时候人们把和服称作“汤帷子（Yukatabira）”。 帷子指单衣，即用于单穿的衣物。 ... 汤帷子被称为“和服”始于江户时代。后来，人们逐渐 改变了入浴的形式，洗浴时不再穿衣，而是浸泡在热水中。于是，和服便成为洗浴 ...

www.tourism.metro.tokyo.jp/chinese/yukata/index.html - 4k - 网页快照 - 类似网页

千龙网--娱乐--日本女星和服秀

推荐 我要说说. 日本女星和服秀. http://www.qianlong.com/ 2004-06-16 08:58:31. 和服是日本民族的传统服装，女式和服色彩缤纷艳丽美不胜收，穿和服体现温婉亲民的一面 也是日本女星的宣传法宝。看看这些女星的和服，你喜欢谁的打扮呢？ ...

ent.qianlong.com/4543/2004/06/15/999@2108733.htm - 38k - 网页快照 - 类似网页

锻灿清供应信息、锻灿清供应商，尽在阿里巴巴网上批发市场

阿里巴巴锻灿清市场是全球最大的锻灿清网上贸易市场，这里有我们为您精选的锻灿清供应信息、 公司黄页、行业资讯、价格行情、展会信息等，阿里巴巴将尽我们所能满足你所有的采购需求！

search.china.alibaba.com/search/offer_search.htm?keywords=**和服**&tracelog=pd_info_promo - 83k - 网页快照 - 类似网页

和服众多的和服商品、和服行情，尽在阿里巴巴市场

阿里巴巴和服市场是全球最大的和服采购中心，这里汇集了大量的和服求购信息、采购公司，快捷的行业资讯、价格行情、展会信息等，阿里巴巴将尽我们所能满足你所有的销售需求！

search.china.alibaba.com/offer/**和服**.html - 215k - 补充材料 - 网页快照 - 类似网页

赞助商链接

流行新款和服
女装/童装/内衣等数千万件在线商品
eBay易趣，全球领先网上交易社区!
www.eBay.com.cn

500元/年起,在此展示广告
在此展示您的广告，500元/年起
先看效果,后付款.咨询:021-50399807
www.5ien.cn

Internet

文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹 媒体
地址(D) http://www.google.com/search?hl=zh-CN&newwindow=1&q=%E5%BC%A0%E6%83%A0&btnG=%E6%90%9C%E7%B4%A2&lr= 转到 链接

登录

网页 图片 资讯 论坛 更多 »

Google 张惠 搜索 高级搜索 使用偏好

搜索所有网页 搜索所有中文网页 搜索简体中文网页

网页

约有**2,370,000**项符合**张惠**的查询结果，以下是第**1-10**项。（搜索用时 **0.27** 秒）

您是不是要找： 张辉 张慧

张惠妹 影音娱乐 新浪网

张惠妹. ... 最新专辑：《勇敢》 主要作品：《发烧》 《给我感觉》《姊妹》 《牵 手》《Bad Boy》 《站在高岗上》《听海》 《一想到你呀》《骄傲》 张惠妹歌曲下载 ▪ 星光无限--Fans地带. 昵称：阿妹，A-mei 小名：卡兹(Katsu) 英文名：Sherry 生日：8月9 ...

ent.sina.com.cn/s/h/f/zhm1.html - 59k - 网页快照 - 类似网页

::张惠妹内地妹迷第一站::[妹力中国]::太阳的部落-音乐无国界，爱妹到永远

www.ameicn.com/ - 2k - 网页快照 - 类似网页

繁體: 張惠妹 - 國際歌友會官方网站即將來臨! 很快 ...

English:. Zhang Hui Mei - International Fan Club Official Website is launching soon! In no time, A*Mei's latest News, In-house pictures and etc will be up in iamei.com to supporters and lovers of A*Mei. ...

www.iamei.com/ - 9k - 网页快照 - 类似网页

张惠妹：魅力永远四射 TOM娱乐

张惠妹：魅力永远四射. ... 包小柏包小松公开批评张惠妹嘲笑为“亏本货” (2005-2-23) >> 张惠妹从乐坛天后到被公开劝退人气暴跌三迷团 (2005-2-23) >> 包小柏炮轰阿妹该休息 孙燕姿越洋声援张惠妹 (2005-2-20) >> 咸鱼与“鲜”鱼两重从包氏兄弟炮轰阿妹说开去 ...

ent.tom.com/Archive/1129/1573/2003/7/29-30377.html - 37k - 网页快照 - 类似网页

高手网flash音乐动画网:张惠妹flashMTV专辑

小游戏:在线小游戏,迷你小游戏,flash小游戏,搞笑flash,flash音乐,网站内容健康丰富,休闲娱乐为主.

完毕 Internet

(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)

后退 搜索 收藏夹 媒体

(D) http://image.baidu.com/i?tn=baiduimage&ct=201326592&cl=2&lm=-1&pv=&word=%C0%EE%D7%D3&z=0 转到 链接

aidu 图片

新闻 网页 贴吧 知道 MP3 图片

李子 百度搜索 帮助 | 高级搜索

新闻图片 全部图片 大图 中图 小图 壁纸

度设为首页

百度一下，找到相关图片约3,630张，用时0.002秒

法欢乐谷一魔幻灵异惊艳

心灵匹配成就幸福婚姻

06超女彩铃官方下载

表达情意的最好方法是？

392×509 124k gif

李子

260×195 14k jpg

北京胜荣科技发展有限公司 - 李子 ...

100×100 8k jpg

标签/好吃|Yupoo.com|全球中文领域...

800×853 152k jpg

李子侯图片列表-油画艺术-大千世界...

800×826 118k jpg

李子侯图片列表-油画艺术-大千世界...

800×853 122k jpg

李子侯图片列表-油画艺术-大千世界...

800×1022 120k jpg

李子侯图片列表-油画艺术-大千世界...

105×150 10k jpg

华夏旅游网-锦绣中华-沈阳

Internet

(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)

后退 搜索 收藏夹 媒体

(D) http://image.baidu.com/i?tn=baiduimage&ct=201326592&cl=2&lm=-1&pv=&word=%D3%F1%C3%D7&z=0 转到 链接

新闻 网页 贴吧 知道 MP3 图片

玉米 百度搜索 帮助 | 高级搜索

新闻图片 全部图片 大图 中图 小图 壁纸

度设为首页

百度一下，找到相关图片约41,300张，用时0.004秒

法欢乐谷一魔幻灵异惊艳 心灵匹配成就幸福婚姻 海通生物饲料玉米蛋白粉.. 益尔鲜甜玉米罐头 06超女彩铃官方下载

米脱粒机--淄博兴民机.. 玉米脱粒机--大兴机械厂 玉米膨化机--北京洋工机械 玉米胚芽膨化--北京中农.. 表达情意的最好方法是?

1024x768 206k jpg

奔向NB的生活_2006 July

1134x854 176k jpg

黑龙江省建三江农垦茂源农产品有限...

947x710 106k jpg

玉米

390x295 30k jpg

玉米－食品图库_食品图片－食品伙伴网

400x266 30k jpg

图文：玉米由坐变站_大挥玉米棒支...

301x282 22k jpg

奉贤区教育资源库>>>图片素材>>>植物

338x450 28k jpg

李宇春_139|要你好玩!

500x375 28k jpg

李宇春_139|要你好玩!

Internet

1.2. Applications of CL

- machine translation

ogle 翻译 - Windows Internet Explorer
http://translate.google.cn/translate?js=y&prev=_t&hl=zh-CN&ie=UTF-8&u=htt
Live Search
Google 翻译
页面(P)
ogle 翻译 谷歌
http://ajnad.aljazeera.net/banners/default.aspx?BannerT
新功能！ 用 Google 工具栏即时翻译
阿拉伯文 > 英语 翻译
工具提示：

Retaken Harf Sufian

Yemeni army evacuates Malahit "tactical"

موبايل
mobile

متعلقات
- Yemeni army regain Harf Sufian
- Houthi rebellion .. Interventions that fuel conflict
- Warnings of the internationalization of the events of Yemen
- Renewed clashes in Saada, Yemen
- Fierce fighting in Saada and the army confirms offer

أهم أخبار الصفحة الرئيسية

完毕，但网页上有错误。
Internet | 保护模式: 禁用
100%

oogle 翻译 - Windows Internet Explorer
http://translate.google.cn/translate?js=y&prev=_t&hl=zh-CN&ie=UTF-8&u=htt
Live Search
Google 翻译
页面(P)
oogle 翻译
http://ajnad.aljazeera.net/banners/default.aspx?BannerT
新功能！ 用 Google 工具栏
工具提
阿拉伯文 > 英语 翻译

Houthis said they were forced to evacuate the Yemeni army Almlahiz (Reuters)

Military sources said that the Yemeni government forces withdrew from the area of tactical Malahit Saada, while the Ministry of Defense confirmed that its forces took control of most of the strongholds of Huthi who's President Ali Abdullah Saleh to overwhelm them as soon as possible.

The Yemeni military sources denied he talked about on the Houthis forced to retreat from the area Malahit. She re-based on the outskirts of the region to curb the entry and exit Huthis so as not to become its easy targets.

In an interview with Al Jazeera, Mohammed Abdel-Salam, the spokesman Houthi Yemeni forces were forced to retreat from the Malahit and other areas, adding that government forces "did not make one iota" in the face of forces.

For its part, military sources said the site "September net" of the Ministry of Defense said that the Yemeni military and security has tightened its

- U.S. plays down hopes the Tripartite Summit
- Zelaya calls for reconciliation in Honduras
- Climate Summit in New York today
- Ahmadinejad: No one will dare to attack us

完毕，但网页上有错误。
Internet | 保护模式: 禁用
100%

ogle 翻译 - Windows Internet Explorer
http://translate.google.cn/translate?hl=zh-CN&sl=ar&tl=zh-CN&u=http%3A%2F
Live Search
Google 翻译
页面(P)
ogle 翻译 谷歌
http://ajnad.aljazeera.net/banners/default.aspx?BannerT
新功能！ 用 Google 工具栏即时翻
阿拉伯文 > 中文(简体) 翻译
工具提示：

More than 3,700,000 visitors per calendar month
visit Al Jazeera's Youtube channels

星期四，29/8/1430啊- 20/8/2009，相应的（最近更新）在17时46分（麦加），14时46分（格林尼治）

موبايل
mobile

家：阿拉伯

重新夺回哈弗苏菲安

也门军队撤走Malahit“战术”

متعلقات
- 也门军队重新哈弗苏菲安
- Houthi叛乱.. 助长冲突的干预措施
- 警告对也门的活动国际化
- 在萨阿达，也门再度发生冲突
- 激烈的战斗中萨阿达和军方证实提供

أهم أخبار الصفحة الرئيسية
- 美国次向下希望三方首脑会

Internet | 保护模式: 禁用
100%

oogle 翻译 - Windows Internet Explorer
http://translate.google.cn/translate?hl=zh-CN&sl=ar&tl=zh-CN&u=http%3A%2F
Live Search
Google 翻译
页面(P)
oogle 翻译
http://ajnad.aljazeera.net/banners/default.aspx?BannerT
新功能！ 用 Google 工具栏即时
阿拉伯文 > 中文(简体) 翻译
工具提示：

Houthis说，他们被迫撤离也门军队Almlahiz（路透社）

军方人士说，也门政府军撤出的战术Malahit萨阿达地区，而国防部证实，它的部队注意到了胡迪的据点，大多数控制谁总统阿里阿卜杜拉萨利赫压倒他们尽快。

也门军方消息来源否认他约在被迫退出该地区Malahit的Houthis交谈。 她重新基于对本区域的郊区，制止出入境Huthis，以免成为其容易攻击的目标。

在接受阿拉伯半岛电视台，穆罕默德阿卜杜勒采访时，萨拉姆，发言人Houthi也门军队被迫撤退从Malahit和其他领域，他又说，政府部队“并没有在部队面临的一个丝毫”。

就其部分，军方人士称，这一网站“9月净国防部”说，也门军队和安全，加紧了对Houthis在萨达和奥姆兰省的大部分据点的控制权，并造成巨大损失，现在是过滤若干地区的叛乱分子所描述的口袋，加强它们。

当地人士说，萨阿达附近地区的城市，已经出现零星的战斗和炮击目标网站Houthi。 她指出，

议
- 塞拉亚在洪都拉斯呼吁和解
- 气候峰会今天在纽约
- 内贾德：没有人敢攻击我们

完毕，但网页上有错误。
Internet | 保护模式: 禁用
100%

1.2. Applications of CL

- cross-lingual information retrieval
- question-answering systems: query a database
- tutoring system
- language model for speech recognition (The listening machine)

  Gong1Shi4 (Pinyin):

  公式(formula)　　工事(work)

  攻势(offensive)　　公事(public affair)

  数学公式,　　摧毁了敌方工事,

  猛烈的攻势,　　例行公事

1.2. Applications of CL

- spoken language control of a machine: mobile phone
- language model for OCR

input: 我们要振奋精神

top one given by pure OCR: 我优要扳奋精种

output (recovered by the language model):

我们要振奋精神

top ten candidates: similarity measures given by OCR:

我	钱	线	载	哦	栽	哉	裁	劣	绥	397	682	700	722	774	781	787	815	838	851
优	仍	**们**	仿	伦	奶	砧	犯	扔	妨	868	878	929	947	953	964	979	984	1009	1010
要	耍	密	穷	安	壁	驻	努	窑	垂	627	650	730	747	749	802	808	818	836	838
扳	报	叔	嵌	奴	**振**	技	寂	叙	蔽	663	709	743	746	755	772	799	815	822	824
奋	夯	杏	蚕	香	脊	秀	吞	吝	番	192	381	393	436	438	471	507	534	543	544
精	猜	指	洁	括	治	捐	活	冶	桔	756	787	791	799	824	826	836	875	885	886
种	**神**	衬	祥	科	钟	拌	样	拎	补	463	548	555	575	636	663	671	681	689	694

1.2. Applications of CL

- language model for TTS (the talking machine)

  一行白鹭上青天 *(a line of)*

  单信一行于昨日抵达香港 *(delegation)*

  你 | 要 | 先 | 搞懂 | 文章 | 大意 (n) *(main points)*

  你 | 千万 | 别 | 大意 (adj) *(careless)*

- proof-reading

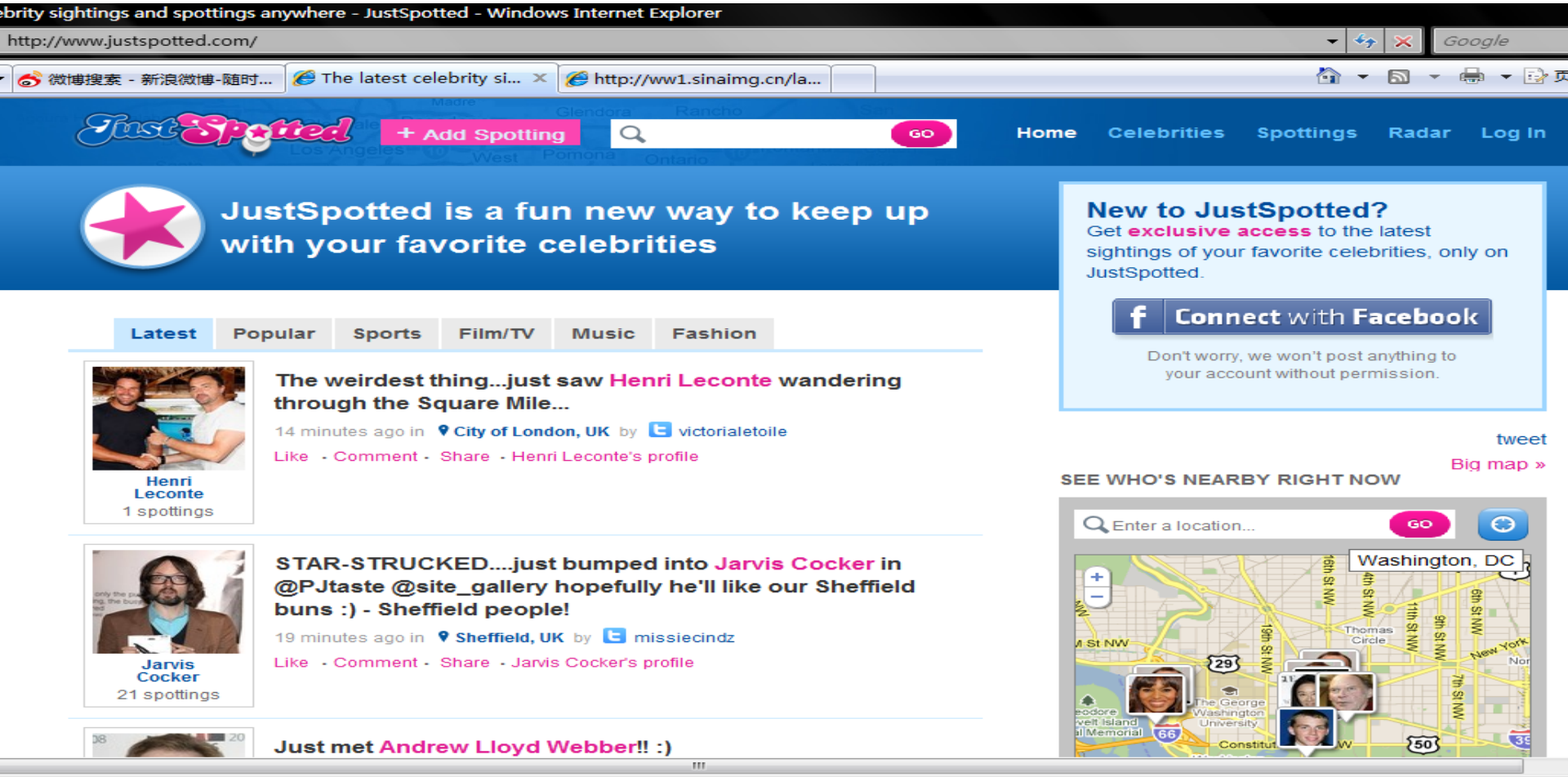
  你己(已)不小了。

- traditional and simplified Chinese character conversion

  发展~头发　后来~皇后

新的应用示例：Twitter用于确定明星位置

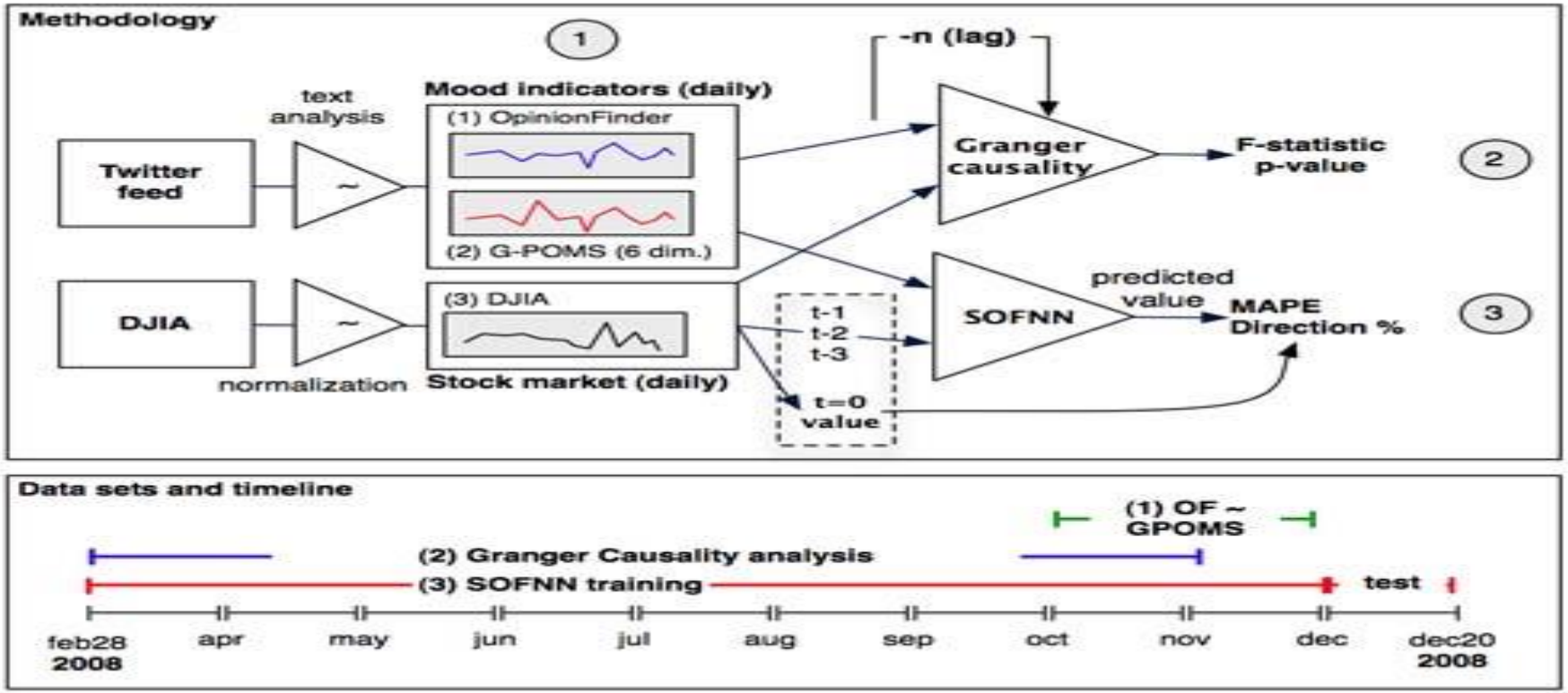
JustSpotted.com与Twitter合作，获取每天超过5000万条公开tweet信息，使用自然语言技术筛选包含明星动向的tweet信息，确定明星目前所在位置。将覆盖7000位明星。



新的应用示例： Twitter 用于预测股市行情

根据行为经济学的原理，情绪对个人的行为和决策有深远影响。公众情绪有可能影响响相关的经济指标。

印第安纳大学研究结果表明，根据 twitter 上收集到的公众情绪状态，道琼斯指数预测的准确率大大提高，预测每天收市时的升降准确率达到87.6%。



新的应用示例： Twitter 用于评估幸福感

东北大学、哈佛大学:

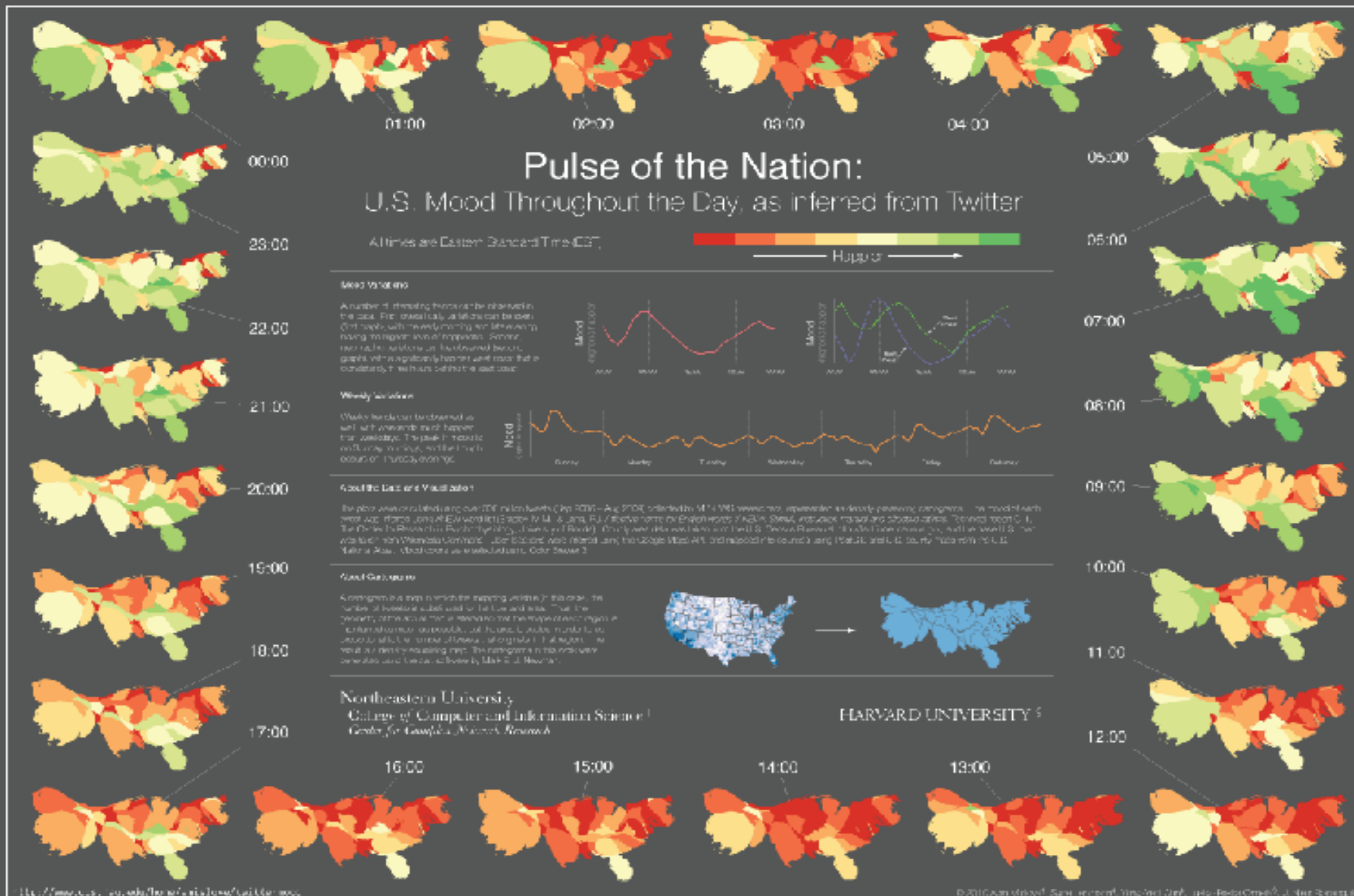
Twitter Mood Map(Pulse of the Nation)

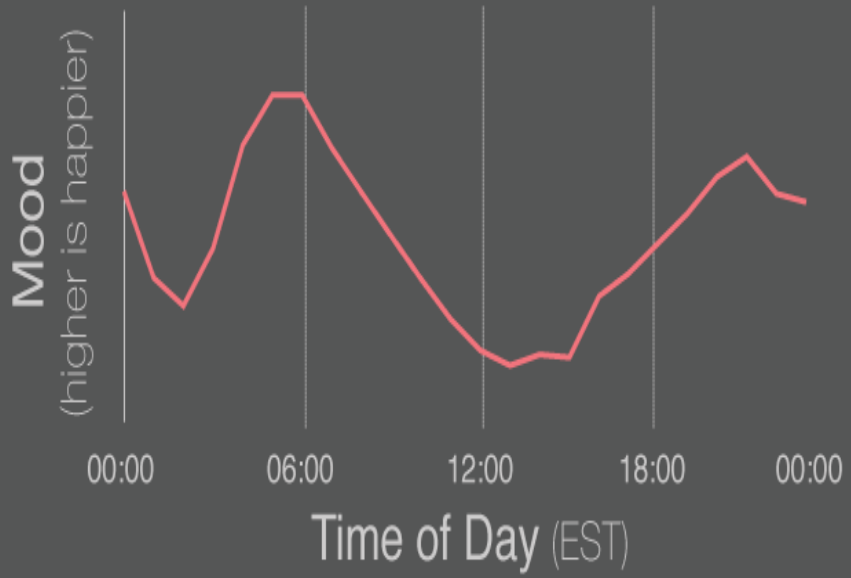
早上和傍晚最快乐，下午最郁闷；

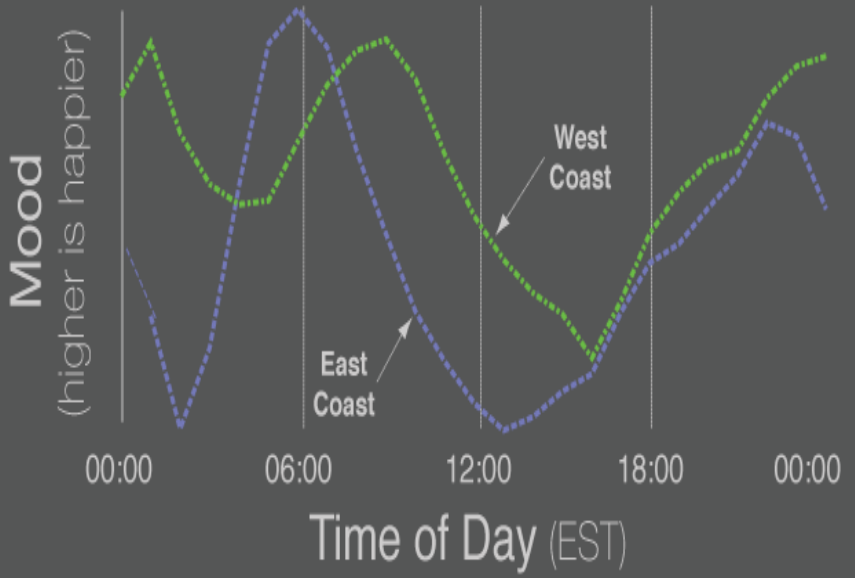
周六最快乐，好心情在周日早上达到巅峰，周四晚上跌至谷底

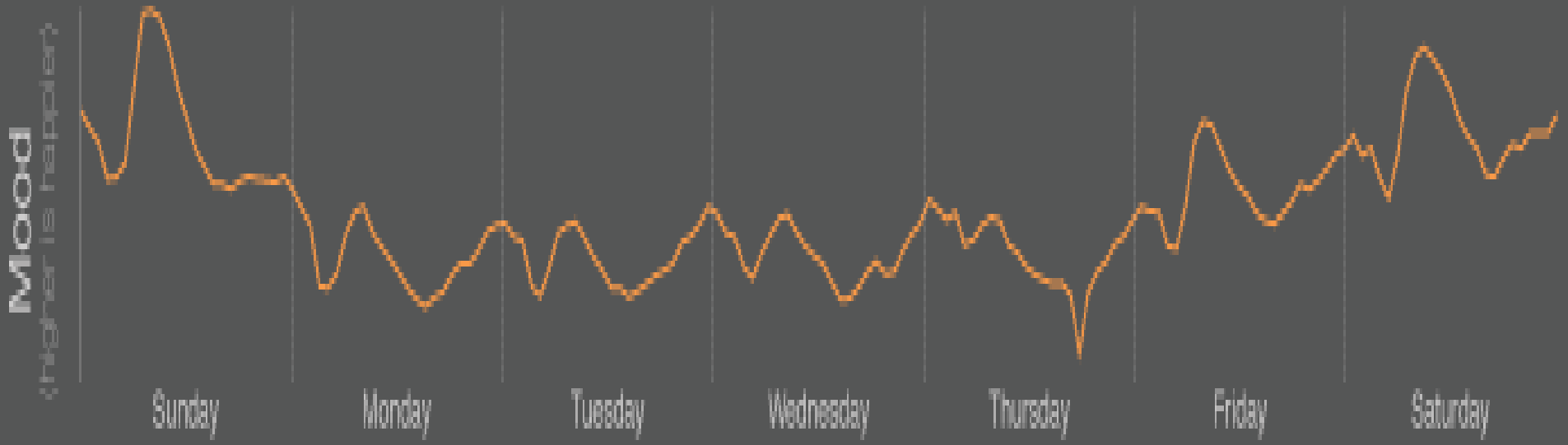
U.S. Mood Throughout the Day inferred from Twitter

Click for high-resolution PDF version (11MB)




Mood
(higher is happier)
00:00
06:00
12:00
18:00
00:00
Time of Day (EST)

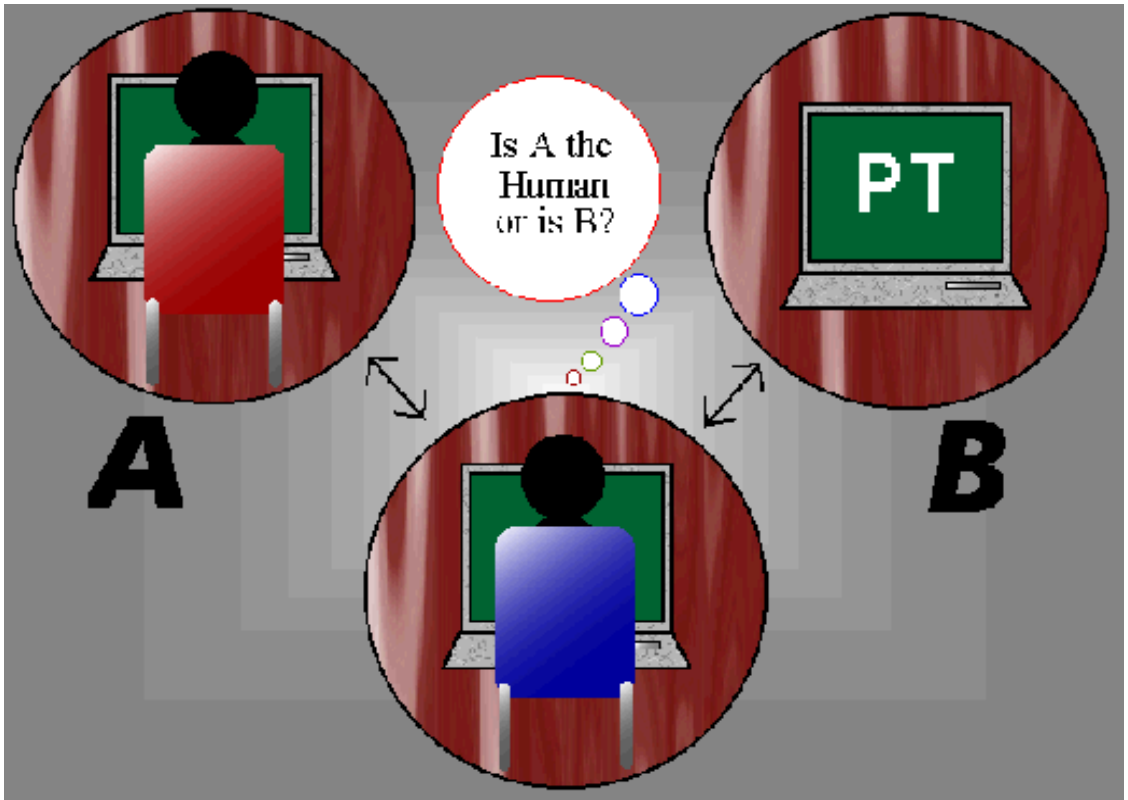

Mood
(higher is happier)
West Coast
East Coast
00:00
06:00
12:00
18:00
00:00
Time of Day (EST)


Mood
(higher is happier)
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday

1.3. Evaluating Language Understanding Systems

Black box evaluation: Turing test “Can machines think?”

The Imitation Game



1.3. Evaluating Language Understanding Systems

Alan Turing: Founder of computer science, mathematician, philosopher

“计算机之父”、“人工智能之父”、“破译之父”

http://www.turing.org.uk/turing/

1.3. Evaluating Language Understanding Systems

1912 (23 June): Birth, Paddington, London
1926-31: Sherborne School
1931-34: Undergraduate at King's College, Cambridge University
1932-35: Quantum mechanics, probability, logic
1935: Elected fellow of King's College, Cambridge
1936: The Turing machine, computability, universal machine
1936-38: Princeton University. Ph.D. Logic, algebra, number theory
1938-39: Return to Cambridge. Introduced to German Enigma cipher machine
1939-40: The Bombe, machine for Enigma decryption
1939-42: Breaking of U-boat Enigma, saving battle of the Atlantic
使用了分析英格玛机的数学结构的方法，也使用了德文交流中的一些规律性，例如德国人会严格地在每天早晨的天气预报发布中用到“WETTER”这个词（即德语中的“天气”）
1943-45: Chief Anglo-American crypto consultant. Electronic work.
1945: National Physical Laboratory, London
1946: Computer and software design leading the world.
1947-48: Programming, neural nets, and artificial intelligence
1948: Manchester University
1949: First serious mathematical use of a computer
1950: The Turing Test for machine intelligence
1951: Elected FRS. Non-linear theory of biological growth
1952: Arrested as a homosexual, loss of security clearance
1953-54: Unfinished work in biology and physics
1954 (7 June): Death (suicide) by cyanide poisoning, Wilmslow, Cheshire.

1.3. Evaluating Language Understanding Systems

German military Enigma machine

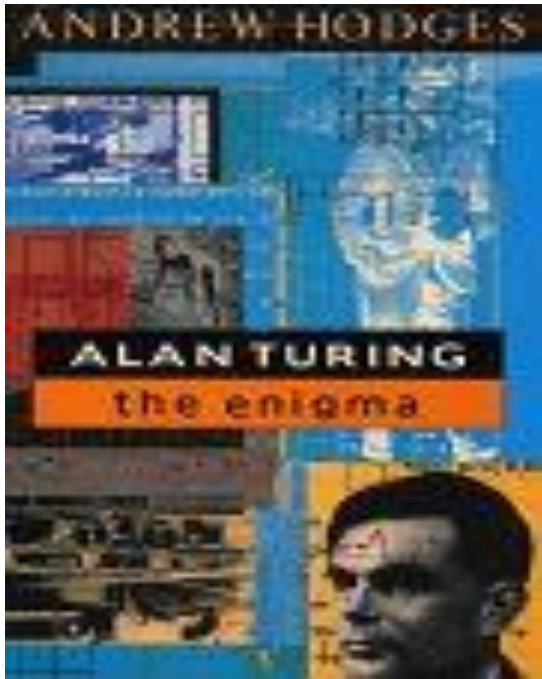
26个字母能替代8万亿个谜文字母。如果改动接线，变化会超过2千5百万亿。

图灵的“图灵炸弹”解码机主要由继电器构成，由光电阅读器直接读入密码。平均每天破译敌方三千条密码。有些重要信息直接送达英国首相丘吉尔。

2009年9月，布朗首相代表英国政府向艾伦·图灵作出了明确的道歉。“如果没有他的卓越贡献，二战的历史也许会被重写”

1.3. Evaluating Language Understanding Systems

牛津大学著名数学家安德鲁·哈吉斯在为图灵写的一部脍炙人口的传记《谜一样的图灵》(Alan Turing: The Enigma)中这样描述到：“图灵似乎是上天派来的一个使者，匆匆而来，匆匆而去，为人间留下了智慧，留下了深邃的思想，后人必须为之思索几十年、上百年甚至永远。”



1.3. Evaluating Language Understanding Systems

ELIZA: in the mid-1960s at MIT: patient and therapist

User:	**ELIZA:**
Man are all alike.	In what way?
They're always bugging us about something or other.	Can you think of a specific example?
Well, my boyfriend made me come here.	Your boyfriend made you come here.
He says I'm depressed much of the time.	I am sorry to hear you are depressed.
It's true, I am unhappy.	Do you think coming here will help you not to be unhappy?
I need some help; that much seems certain.	What would it mean to you if you got some help?
Perhaps I could learn to get along with my Mother.	Tell me more about your family.
My mother takes care of me.	Who else in your family takes care of you?
My father.	Your father.
You are like my father in some ways.	What resemblance do you see?

1.3. Evaluating Language Understanding Systems

Keywords *Why are you looking at me?* Replace *me* with *you*

Word	Rank	Pattern	Output
alike	10	?X	In what way?
			What resemblance do you see?
are	3	?X are you ?Y	Would you prefer it if I weren't ?Y?
	3	?X are ?Y	What if they were not ?Y?
always	5	?X	Can you think of a specific example?
			When?
			Really, always?
what	2	?X	Why do you ask?
			Does that interest you?

if no keyword: *"tell me more" "go on"*

"Necessity is the mother of invention": "Tell me more about your family".

"Green the adzabak are the a ran four": "What if they were not the a ran four?"

1.3. Evaluating Language Understanding Systems

2008年图灵测试Loebner prize所设置任务：

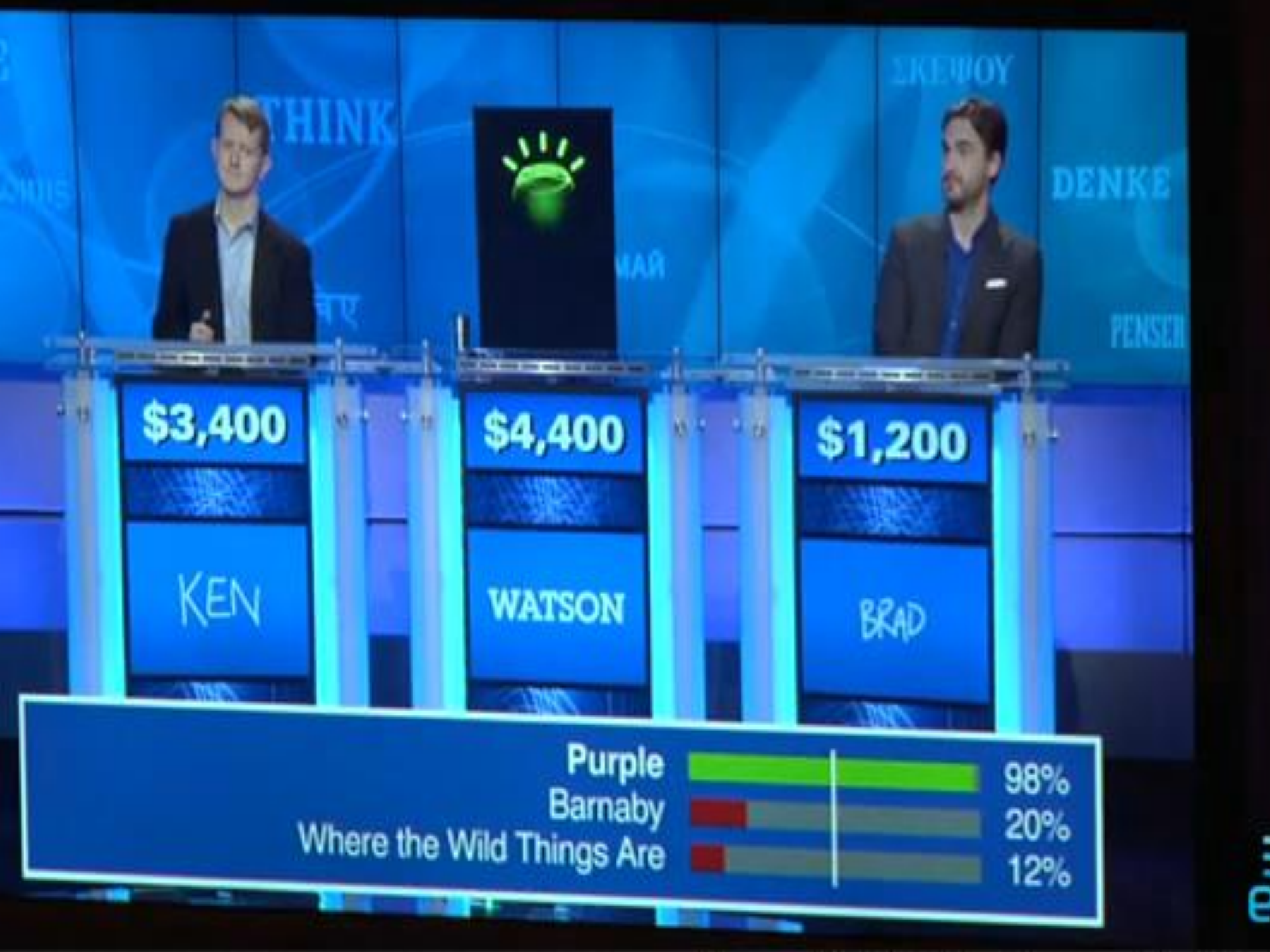
“人工交谈实体”

（Artificial conversational entity）

检验语言交谈能力

冲击图灵测试的重要努力：IBM的Watson (DeepQA)

JEOPARDY!


THINK
DENKE
PENSER
$3,400
KEN
$4,400
WATSON
$1,200
BRAD
Purple
98%
Barnaby
20%
Where the Wild Things Are
12%

* Jeopardy! 是美国著名的益智节目长青树，1964 年至今近五十年
* 为何华生要选择这个节目？华生需要能够听懂包括俚语、双关词、委婉语、譬喻等人话、也要能够说人话，并且以最快的速度抢达，还有该如何选择接下来的问题，以及何时该拒绝答题（答错问题倒扣分）
* 图灵测试（人工智能）

Question: Who was presidentially pardoned on September 8, 1974?

One of the retrieved passages is "Ford pardoned Nixon on Sept. 8, 1974."

One passage scorer counts the number of IDF-weighted terms in common between the question and the passage.
Another passage scorer based on the Smith-Waterman sequence-matching algorithm, measures the lengths of The longest similar subsequences between the question and passage (for example "on Sept. 8, 1974").
A third type of passage scoring measures the alignment of the logical forms of the question and passage.

The logical form alignment identifies Nixon as the object of the pardoning in the passage, and that the question is asking for the object of a pardoning.
Logical form alignment
gives “Nixon” a good score given this evidence.

In contrast, a candidate answer like “Ford” would receive near identical scores to “Nixon” for term matching and passage alignment with this passage, but would receive a lower logical form alignment score.

Question Analysis:

The system attempts to understand what the question is asking and performs the initial analyses that determine how the question will be processed by the rest of the system.

The DeepQA: shallow parses, deep parses, logical forms, semantic role labels, coreference, relations, named entities, and so on, as well as specific kinds of analysis for question answering.

* 云计算支持的知识处理

华生由多台IBM Power7处理器平台服务器（数千个CPU核），+ IBM专为这次问答开发的IBM DeepQA软件构成，本身是不连网的超级计算机。华生能够透过语音进行分析，并且从背后共15TB的数据库和知识库（Wordnet, Wikipedia 等）中找出或计算出答案（1T大约为1万亿个字符）。