Distributed Machine Learning

Application Driven

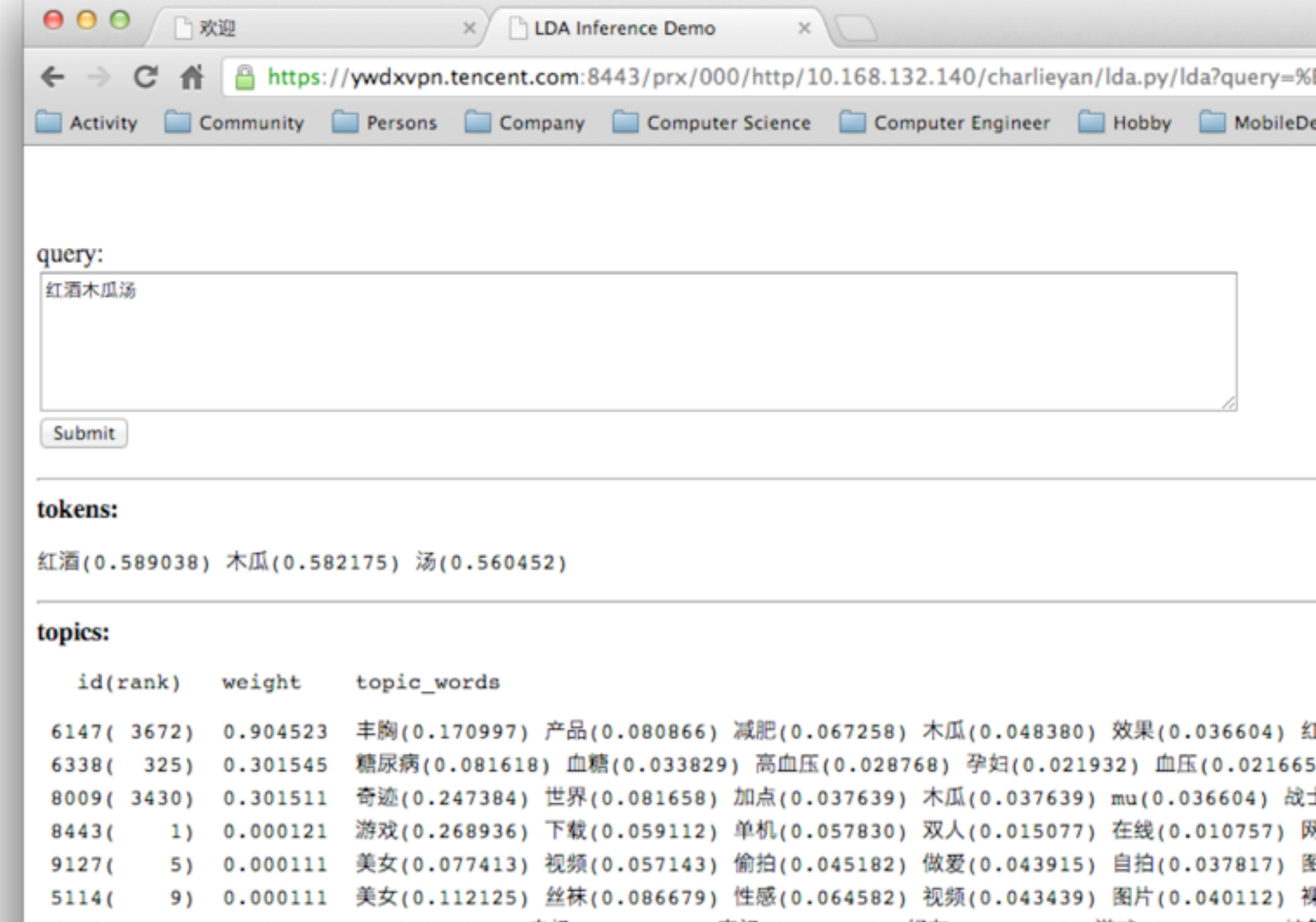
Yi Wang

Goal

- Aim before fire!
  - Recommender systems
  - Search engine
  - Online advertising

Problem

- What is the real problem behind applications?
    - semantic understanding



query:

红酒木瓜汤

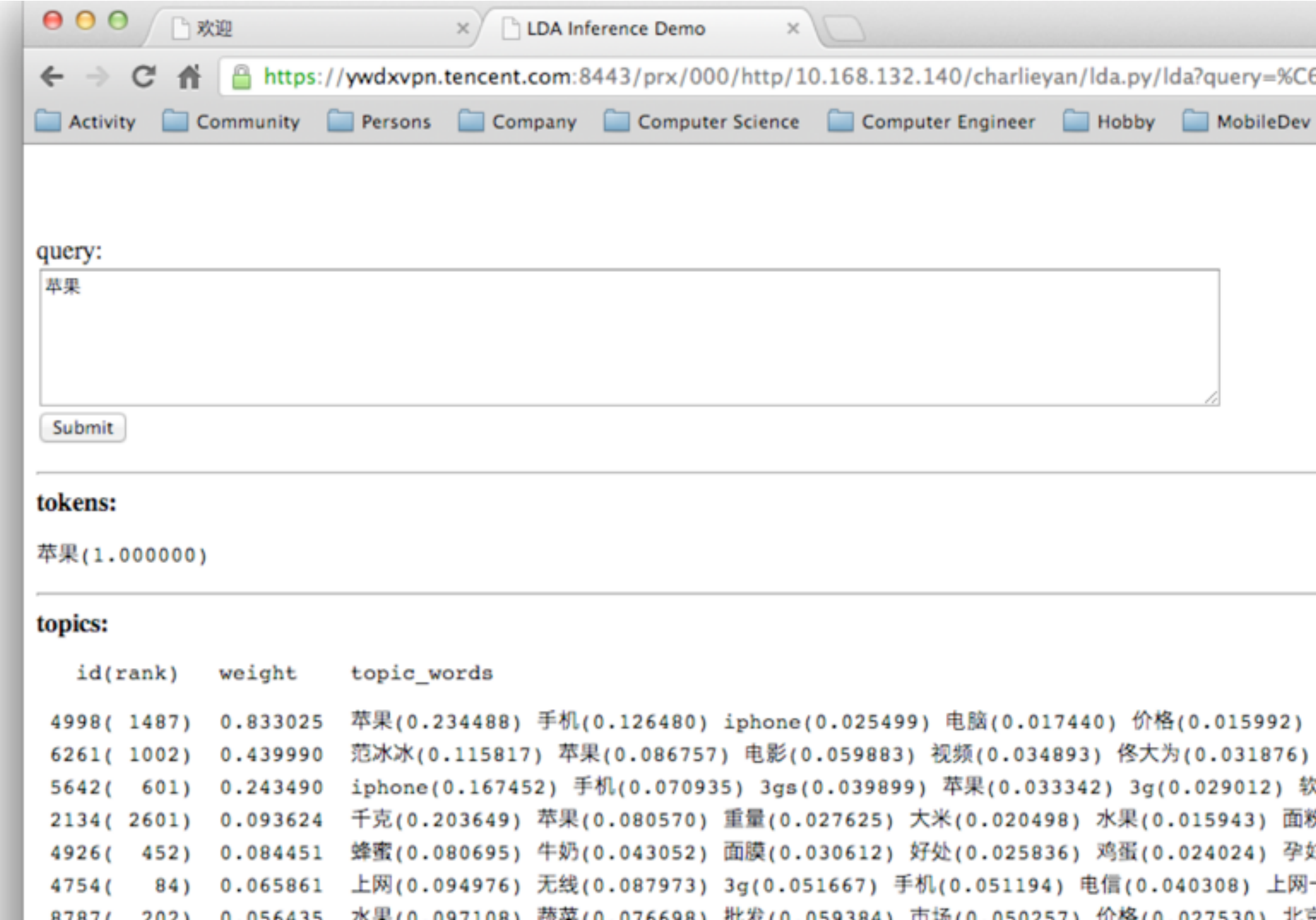
tokens:

```
红酒(0.589038) 木瓜(0.582175) 汤(0.560452)
```

topics:

```
  id(rank)   weight    topic_words

 6147( 3672)  0.904523  丰胸(0.170997) 产品(0.080866) 减肥(0.067258) 木瓜(0.048380) 效果(0.036604)
 6338(  325)  0.301545  糖尿病(0.081618) 血糖(0.033829) 高血压(0.028768) 孕妇(0.021932) 血压(0.021665
 8009( 3430)  0.301511  奇迹(0.247384) 世界(0.081658) 加点(0.037639) 木瓜(0.037639) mu(0.036604)
 8443(    1)  0.000121  游戏(0.268936) 下载(0.059112) 单机(0.057830) 双人(0.015077) 在线(0.010757)
 9127(    5)  0.000111  美女(0.077413) 视频(0.057143) 偷拍(0.045182) 做爱(0.043915) 自拍(0.037817)
 5114(    9)  0.000111  美女(0.112125) 丝袜(0.086679) 性感(0.064582) 视频(0.043439) 图片(0.040112)
```



query:

苹果

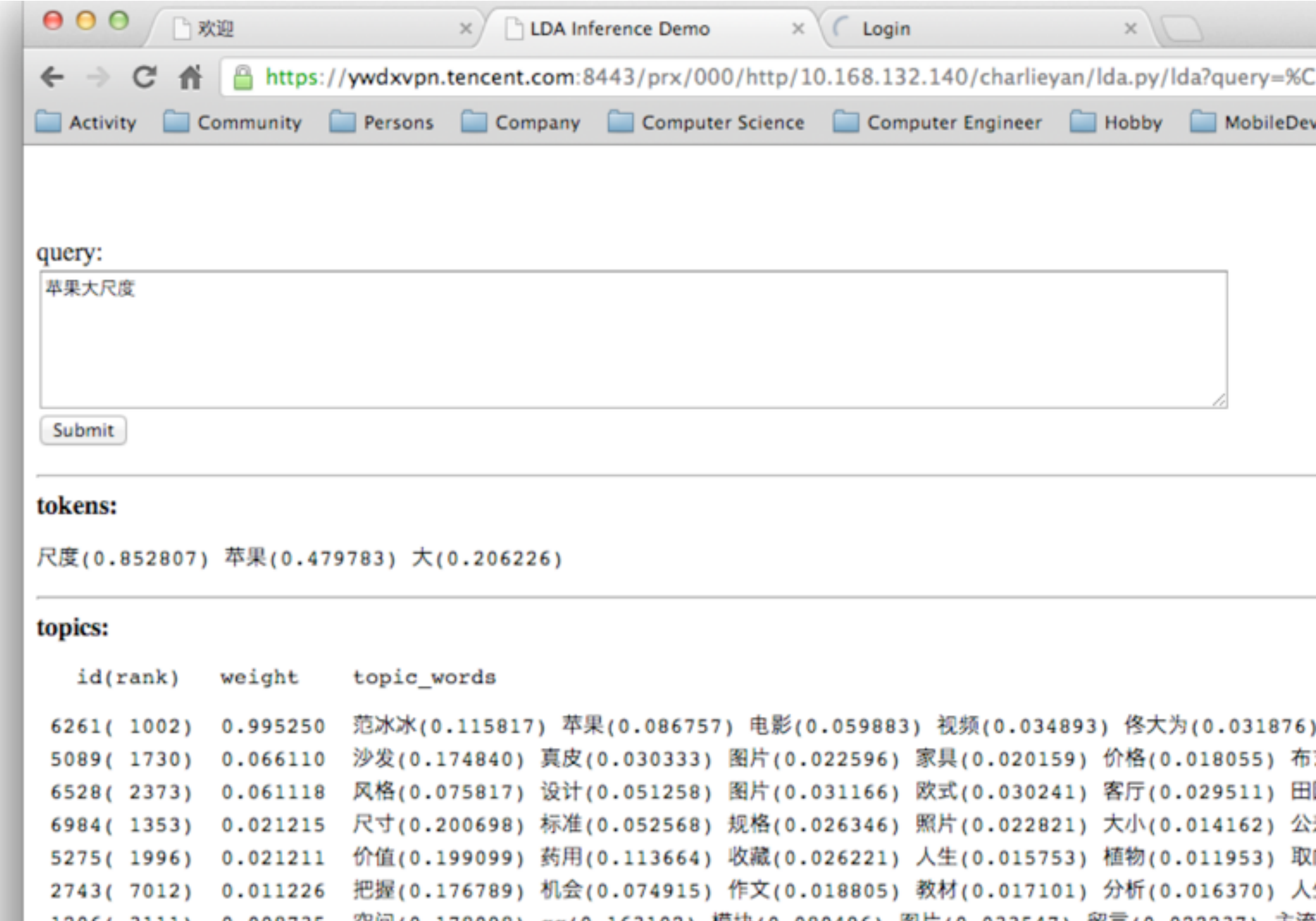
tokens:

苹果(1.000000)

topics:

```
  id(rank)   weight    topic_words

4998( 1487)  0.833025  苹果(0.234488) 手机(0.126480) iphone(0.025499) 电脑(0.017440) 价格(0.015992)
6261( 1002)  0.439990  范冰冰(0.115817) 苹果(0.086757) 电影(0.059883) 视频(0.034893) 佟大为(0.031876)
5642(  601)  0.243490  iphone(0.167452) 手机(0.070935) 3gs(0.039899) 苹果(0.033342) 3g(0.029012) 软
2134( 2601)  0.093624  千克(0.203649) 苹果(0.080570) 重量(0.027625) 大米(0.020498) 水果(0.015943) 面
4926(  452)  0.084451  蜂蜜(0.080695) 牛奶(0.043052) 面膜(0.030612) 好处(0.025836) 鸡蛋(0.024024) 孕
4754(   84)  0.065861  上网(0.094976) 无线(0.087973) 3g(0.051667) 手机(0.051194) 电信(0.040308) 上网
8787(  202)  0.056435  水果(0.097108) 蔬菜(0.076698) 批发(0.059384) 市场(0.050257) 价格(0.027530) 北京
```



https://ywdxvpn.tencent.com:8443/prx/000/http/10.168.132.140/charlieyan/lda.py/lda?query=%C

query:

苹果大尺度

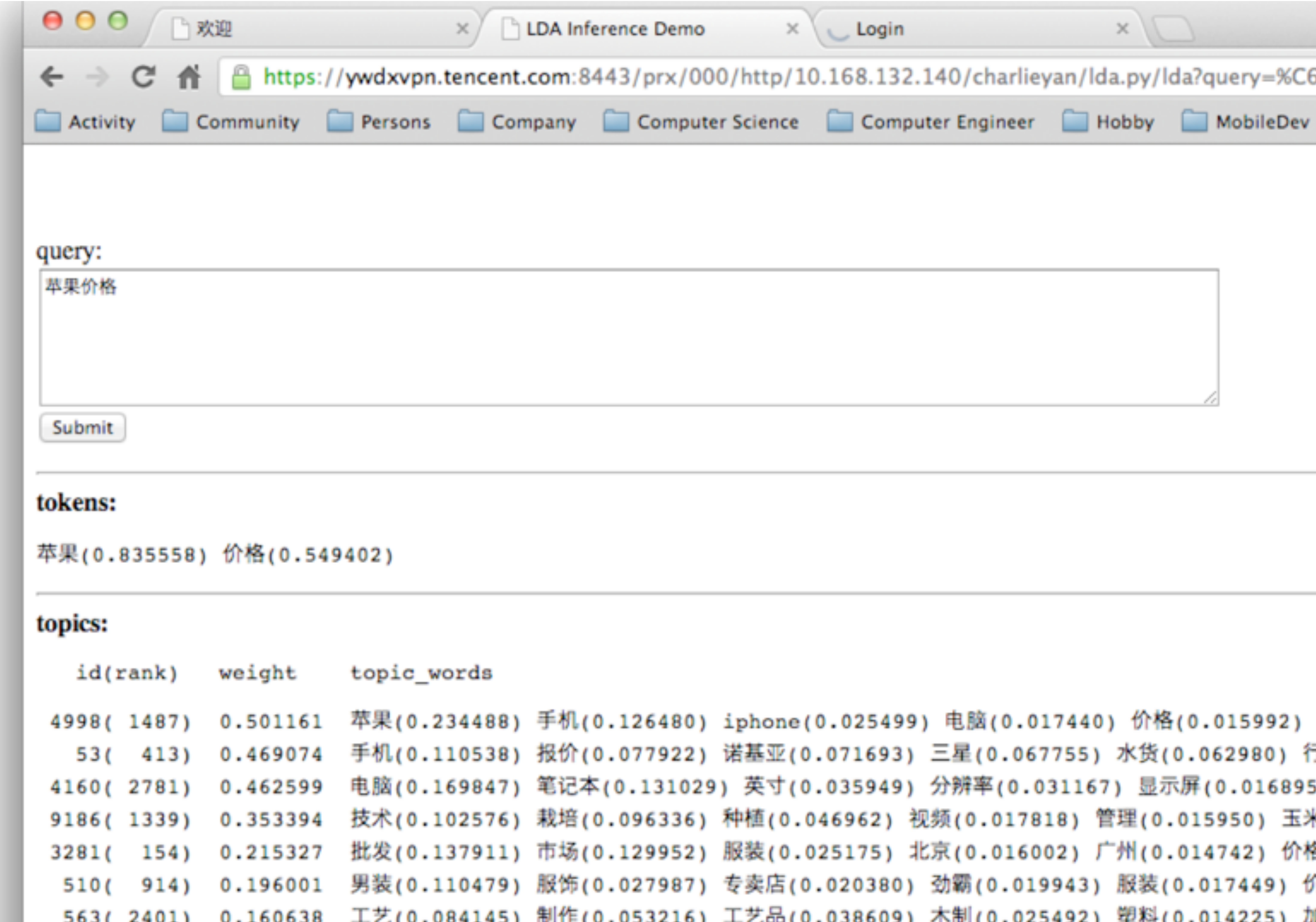
tokens:

```
尺度(0.852807) 苹果(0.479783) 大(0.206226)
```

topics:

```
  id(rank)   weight    topic_words

 6261( 1002)  0.995250  范冰冰(0.115817) 苹果(0.086757) 电影(0.059883) 视频(0.034893) 佟大为(0.031876)
 5089( 1730)  0.066110  沙发(0.174840) 真皮(0.030333) 图片(0.022596) 家具(0.020159) 价格(0.018055)
 6528( 2373)  0.061118  风格(0.075817) 设计(0.051258) 图片(0.031166) 欧式(0.030241) 客厅(0.029511)
 6984( 1353)  0.021215  尺寸(0.200698) 标准(0.052568) 规格(0.026346) 照片(0.022821) 大小(0.014162)
 5275( 1996)  0.021211  价值(0.199099) 药用(0.113664) 收藏(0.026221) 人生(0.015753) 植物(0.011953)
 2743( 7012)  0.011226  把握(0.176789) 机会(0.074915) 作文(0.018805) 教材(0.017101) 分析(0.016370)
```



query:

苹果价格

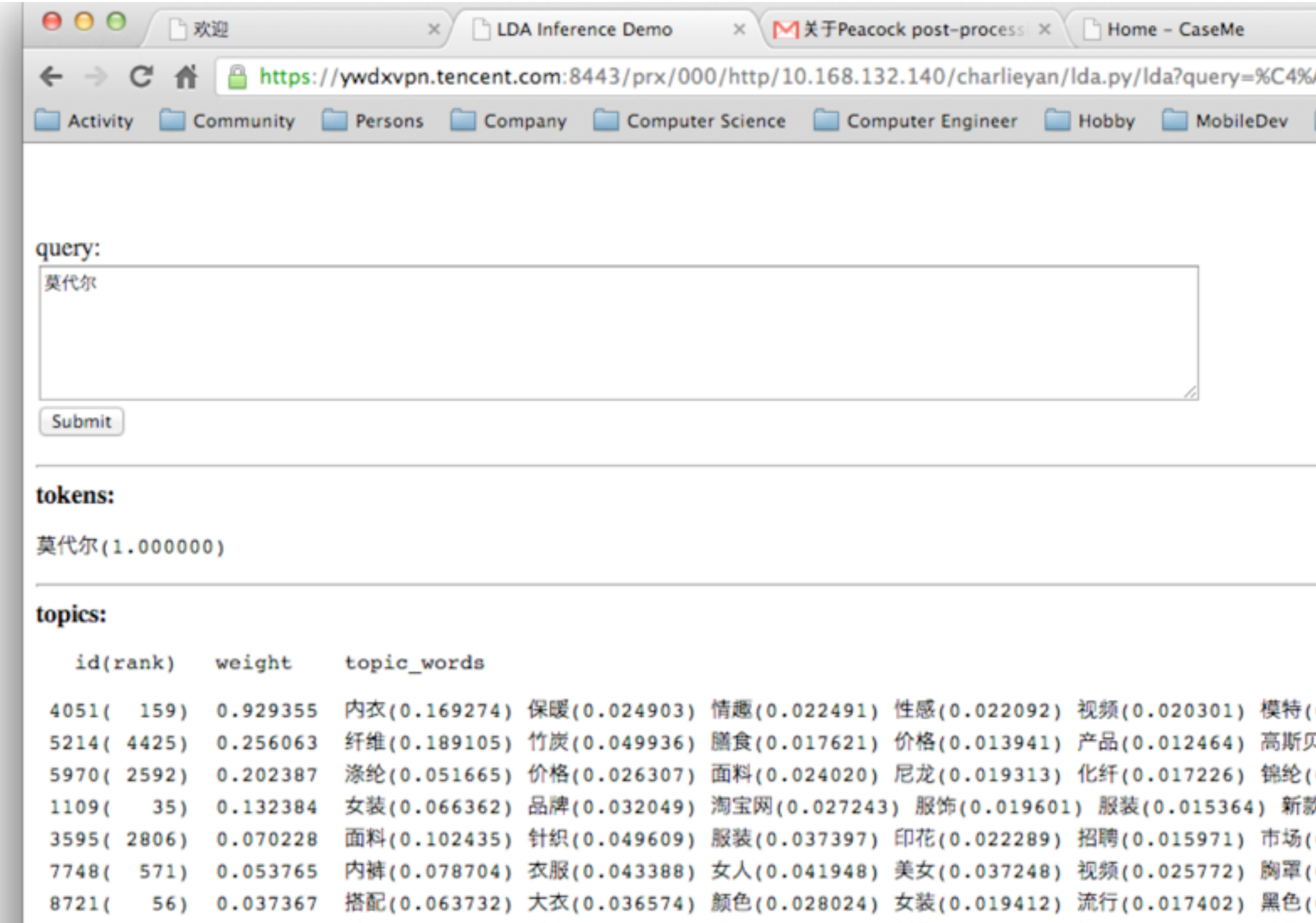
tokens:

```
苹果(0.835558) 价格(0.549402)
```

topics:

```
  id(rank)   weight    topic_words

4998( 1487)  0.501161  苹果(0.234488) 手机(0.126480) iphone(0.025499) 电脑(0.017440) 价格(0.015992)
  53(  413)  0.469074  手机(0.110538) 报价(0.077922) 诺基亚(0.071693) 三星(0.067755) 水货(0.062980)
4160( 2781)  0.462599  电脑(0.169847) 笔记本(0.131029) 英寸(0.035949) 分辨率(0.031167) 显示屏(0.01689
9186( 1339)  0.353394  技术(0.102576) 栽培(0.096336) 种植(0.046962) 视频(0.017818) 管理(0.015950)
3281(  154)  0.215327  批发(0.137911) 市场(0.129952) 服装(0.025175) 北京(0.016002) 广州(0.014742)
 510(  914)  0.196001  男装(0.110479) 服饰(0.027987) 专卖店(0.020380) 劲霸(0.019943) 服装(0.017449)
 563( 2401)  0.160638  工艺(0.084145) 制作(0.053216) 工艺品(0.038609) 木制(0.025492) 塑料(0.014225)
```



query:

莫代尔

tokens:

莫代尔(1.000000)

topics:

```
  id(rank)   weight    topic_words

 4051(  159)  0.929355  内衣(0.169274) 保暖(0.024903) 情趣(0.022491) 性感(0.022092) 视频(0.020301) 模特(
 5214( 4425)  0.256063  纤维(0.189105) 竹炭(0.049936) 膳食(0.017621) 价格(0.013941) 产品(0.012464) 高斯贝
 5970( 2592)  0.202387  涤纶(0.051665) 价格(0.026307) 面料(0.024020) 尼龙(0.019313) 化纤(0.017226) 锦纶(
 1109(   35)  0.132384  女装(0.066362) 品牌(0.032049) 淘宝网(0.027243) 服饰(0.019601) 服装(0.015364) 新
 3595( 2806)  0.070228  面料(0.102435) 针织(0.049609) 服装(0.037397) 印花(0.022289) 招聘(0.015971) 市场(
 7748(  571)  0.053765  内裤(0.078704) 衣服(0.043388) 女人(0.041948) 美女(0.037248) 视频(0.025772) 胸罩(
 8721(   56)  0.037367  搭配(0.063732) 大衣(0.036574) 颜色(0.028024) 女装(0.019412) 流行(0.017402) 黑色(
```

Applications

- Recommendation
- Search engines
- Online advertising
- Business analytics

Recommendation

- Collaborative filtering

items

	1	1		1	1
	1	1	1		1
users		1	1	1	
				1	1
				1	1

Recommendation

- Uniqueness

			items		
	1				
		1			
users			1		
				1	
					1

Recommendation

- Commonality

	items					topics	
	1	1	1			1	
	1	1	1			1	
users		1	1	1		2/3	1/3
				1	1		1
				1	1		1
topics	1	1	1			8/13	
				1	1		5/13

Search Engine

- Text similarity

words

	Bill	Gates							
	Bill	Gates	Microsoft						
			Microsoft	Windows					
text					Steve	Jobs			
					Steve	Jobs	Apple		
							Apple	iPhone	
							Apple		iPad

Search Engine

- Text similarity

words

	Bill	Gates							
	Bill	Gates	Microsoft						
			Microsoft	Windows					
text					Steve	Jobs			
					Steve	Jobs	Apple		
							Apple	iPhone	
							Apple		iPad

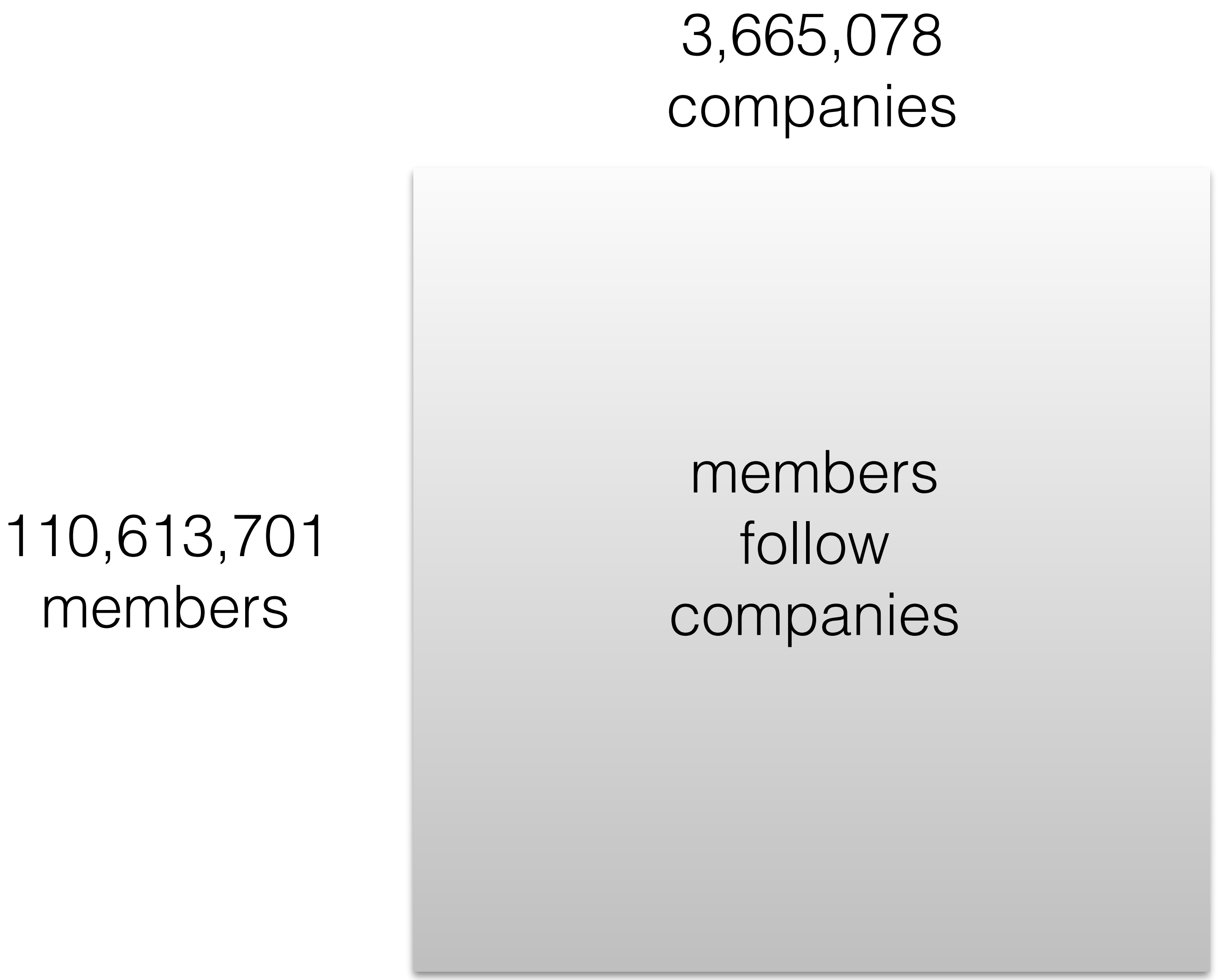
Search Engine

- Text similarity

words

	Bill	Gates							
	Bill	Gates	Microsoft						
			Microsoft	Windows					
text					Steve	Jobs			
					Steve	Jobs	Apple		
							Apple	iPhone	
							Apple		iPad

Business Analytics

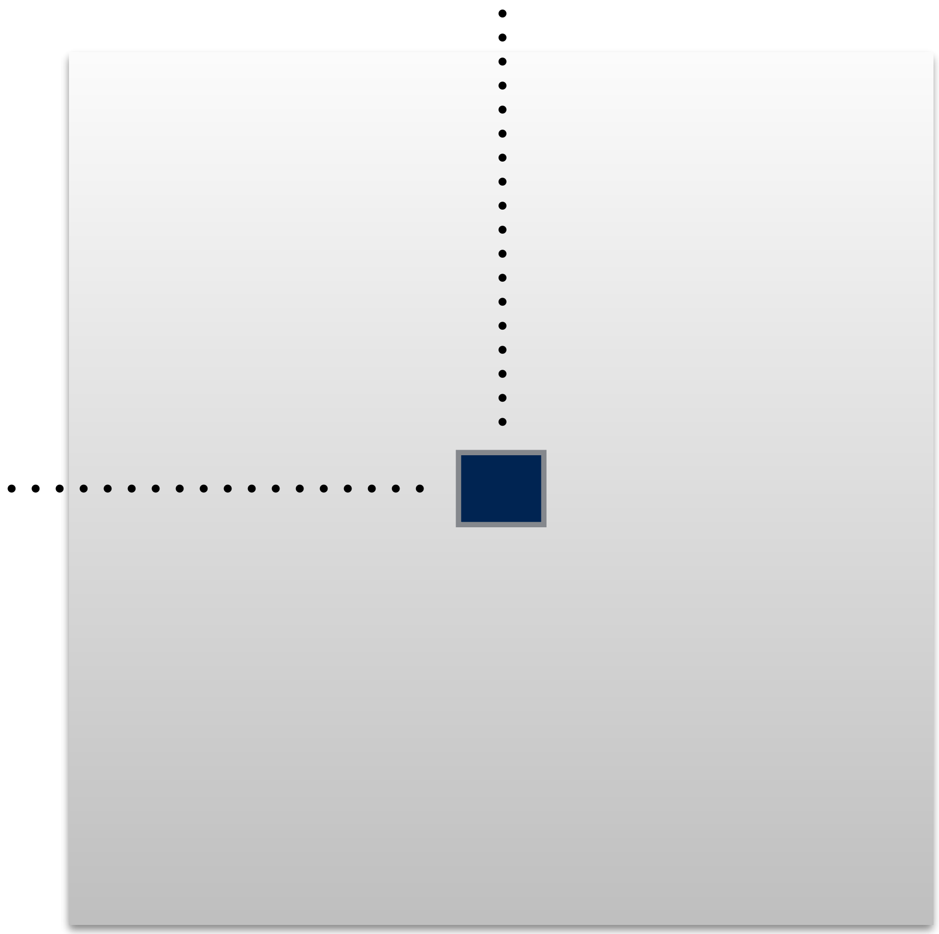


3,665,078
companies

110,613,701
members

members
follow
companies

Business Analytics

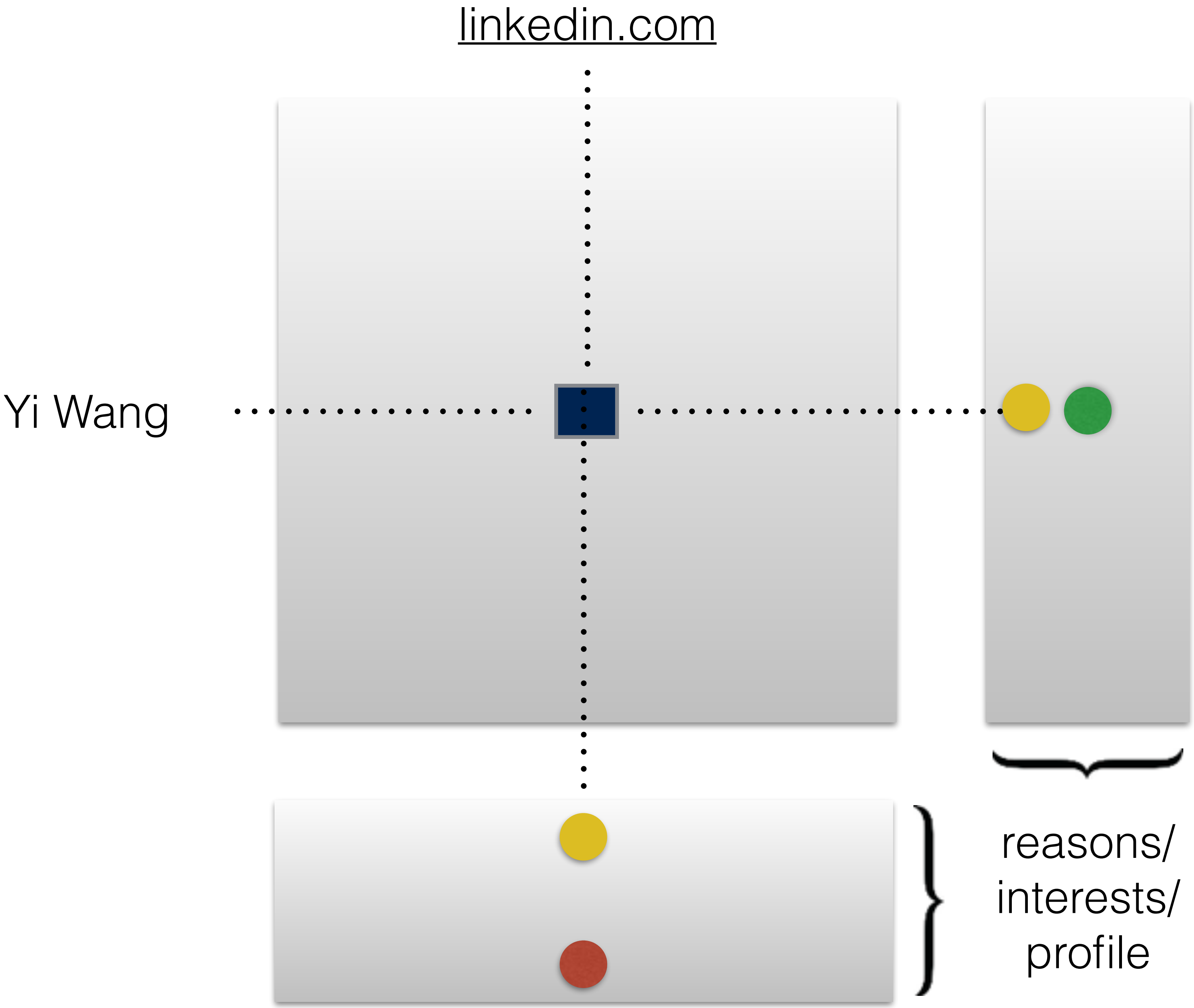


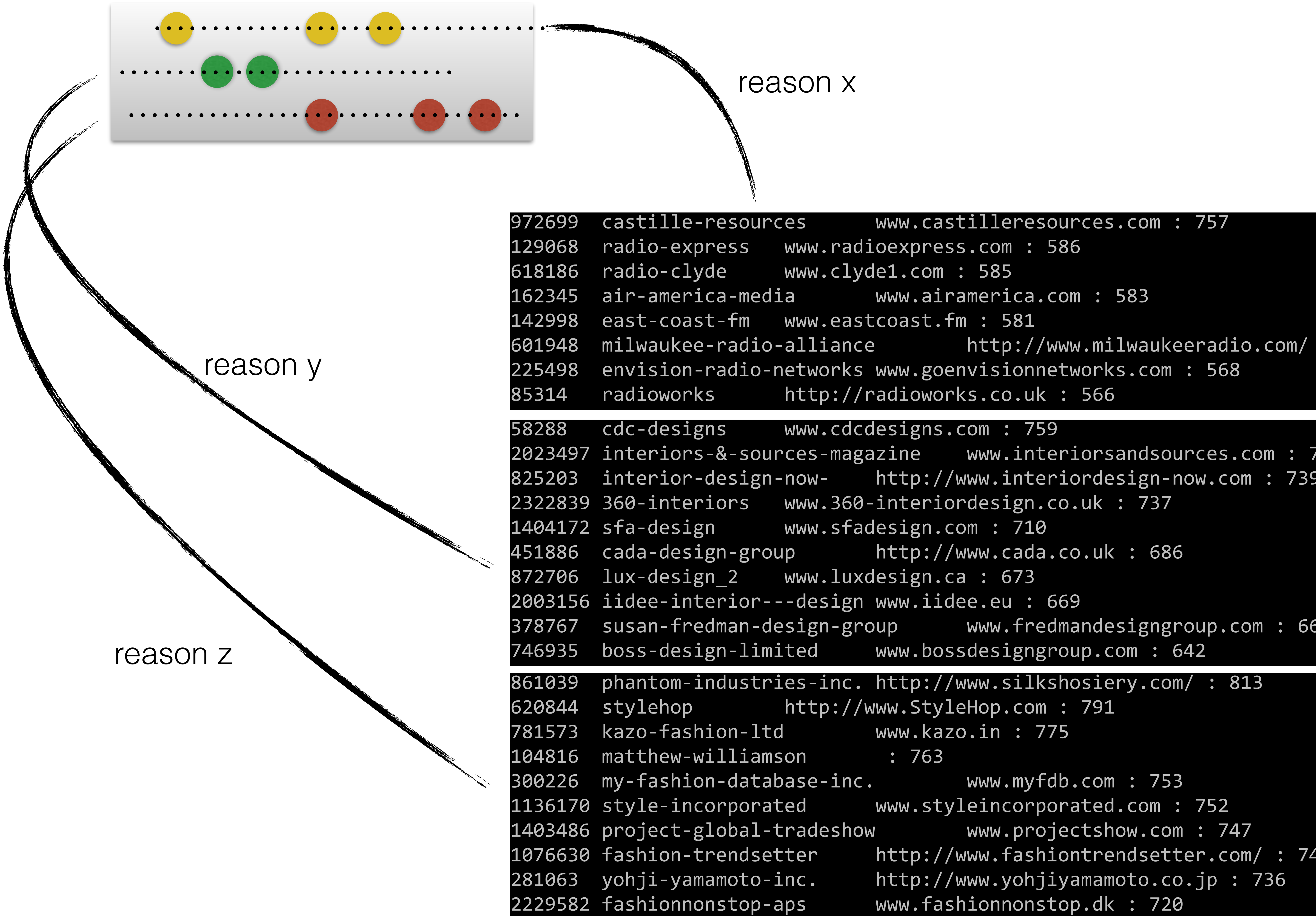
linkedin.com

Yi Wang

what is linkedin.com?
why this guy follows it?

Business Analytics



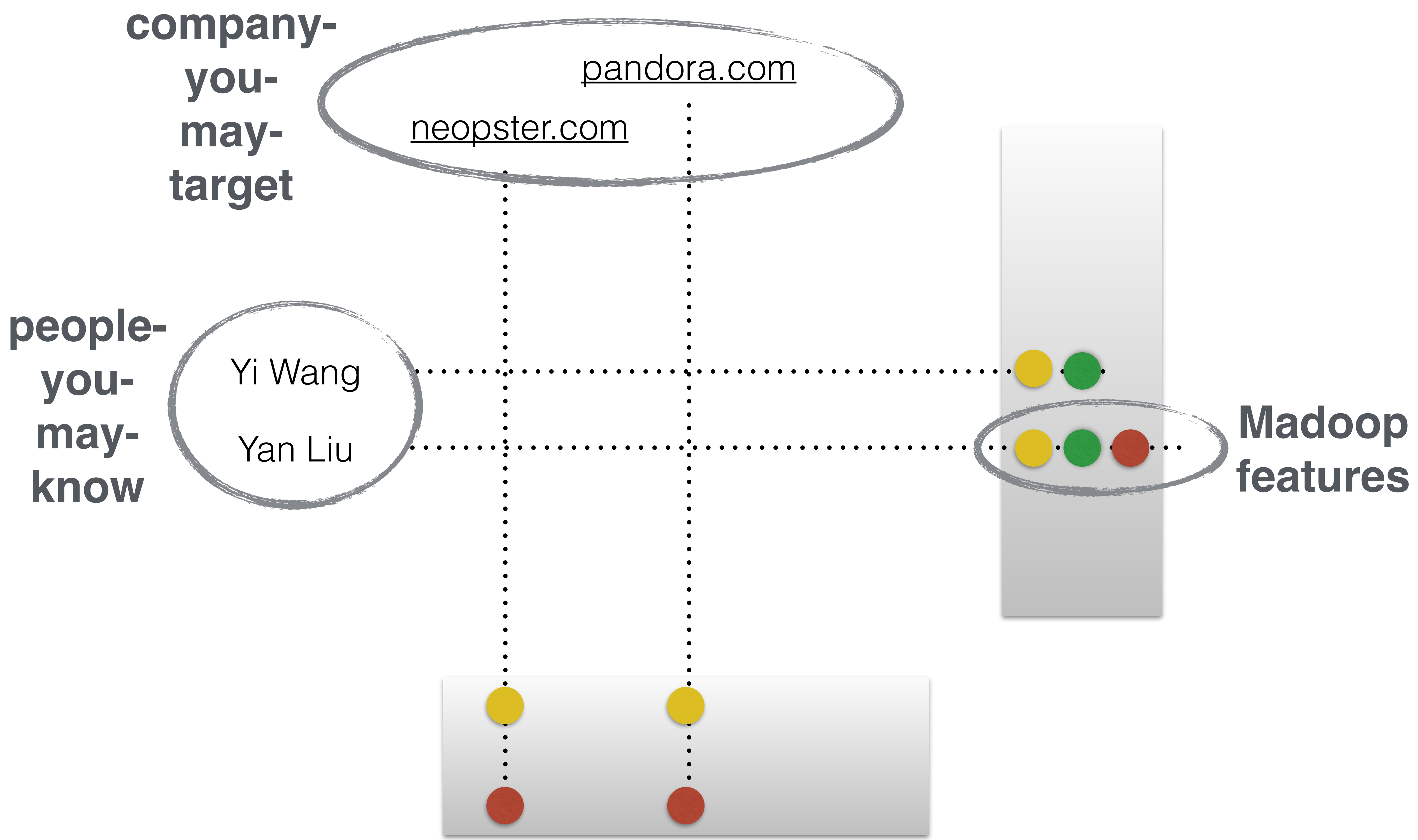


```
972699  castille-resources      www.castilleresources.com : 757
129068  radio-express   www.radioexpress.com : 586
618186  radio-clyde     www.clyde1.com : 585
162345  air-america-media       www.airamerica.com : 583
142998  east-coast-fm   www.eastcoast.fm : 581
601948  milwaukee-radio-alliance        http://www.milwaukeeradio.com/
225498  envision-radio-networks www.goenvisionnetworks.com : 568
85314   radioworks      http://radioworks.co.uk : 566
```

```
58288   cdc-designs     www.cdcdesigns.com : 759
2023497 interiors-&-sources-magazine    www.interiorsandsources.com : 7
825203  interior-design-now-    http://www.interiordesign-now.com : 739
2322839 360-interiors   www.360-interiordesign.co.uk : 737
1404172 sfa-design      www.sfadesign.com : 710
451886  cada-design-group       http://www.cada.co.uk : 686
872706  lux-design_2    www.luxdesign.ca : 673
2003156 iidee-interior---design www.iidee.eu : 669
378767  susan-fredman-design-group      www.fredmandesigngroup.com : 66
746935  boss-design-limited     www.bossdesigngroup.com : 642
```

```
861039  phantom-industries-inc. http://www.silkshosiery.com/ : 813
620844  stylehop        http://www.StyleHop.com : 791
781573  kazo-fashion-ltd        www.kazo.in : 775
104816  matthew-williamson       : 763
300226  my-fashion-database-inc.        www.myfdb.com : 753
1136170 style-incorporated      www.styleincorporated.com : 752
1403486 project-global-tradeshow        www.projectshow.com : 747
1076630 fashion-trendsetter     http://www.fashiontrendsetter.com/ : 74
281063  yohji-yamamoto-inc.     http://www.yohjiyamamoto.co.jp : 736
2229582 fashionnonstop-aps      www.fashionnonstop.dk : 720
```

Business Analytics



company-you-may-target

pandora.com

neopster.com

people-you-may-know

Yi Wang

Yan Liu

Madoop features

All Problems are Same!

- Semantics = commonalities = co-occurrences
- All methods in the history are finding co-occurrences

Prior Work

- What does prior work do?
  - unsupervised: collaborative filtering, matrix factorization, probabilistic latent semantic analysis
  - supervised: categorization and classification
  - human labour: tags

Unsupervised

- Frequent itemset mining
- collaborative filtering
- LSA - SVD decomposition of text matrix
- NMF - constraint SVD
- pLSA - probabilistic version of LSA
- LDA - smoothed pLSA
- GaP - A re-modeling of LDA
- RBM - A re-modeling of LDA
- HDP - extending LDA to infinite #semantics

All Methods are Same!

- Methods are equivalent to each other under conditions.
  - On the equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing
    http://users.cis.fiu.edu/~taoli/pub/NMFpLSIequiv.pdf
  - On an Equivalence between PLSI and LDA
    http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.6893
  - Replicated Softmax: an Undirected Topic Model papers.
    http://nips.cc/paper/3856-replicated-softmax-an-undirected-topic-model.pdf

Use It!

- Behind search engine, recommender systems, and online advertising:
    - Relevance: information retrieval
    - Ranking: click-through rate prediction