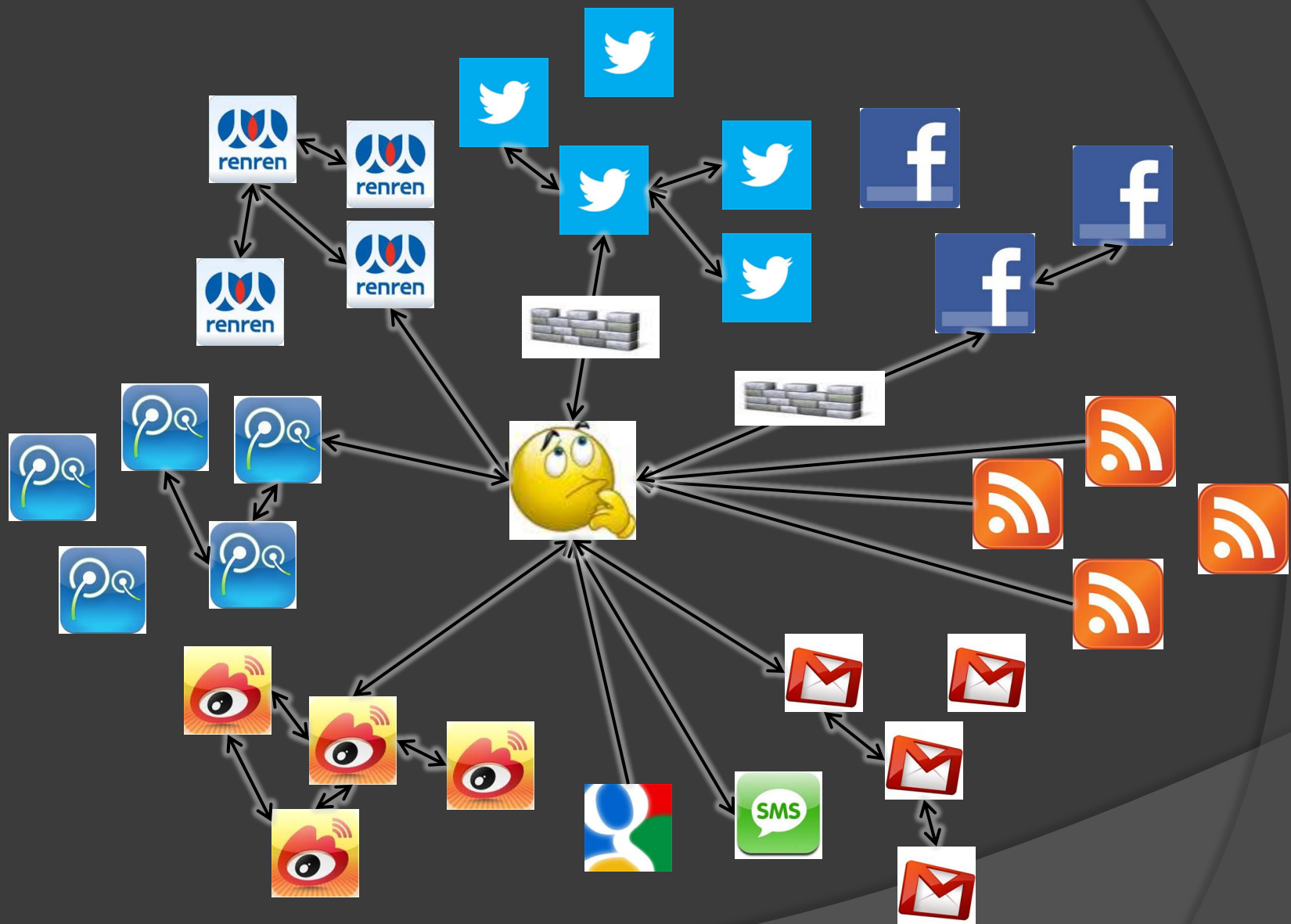


# A FRAMEWORK FOR INTELLIGENT MESSAGE ROUTING ON SNS

HU Pili

Dec 4, 2012



# Motivation -- SNSAPI

- Too Many Platforms
- Heterogeneous Interface

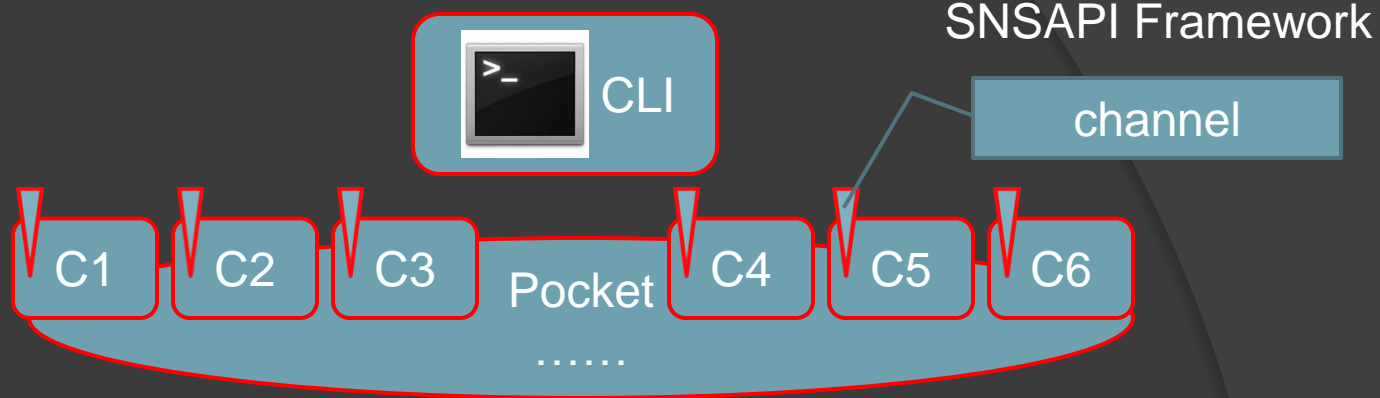


- Data Safety!! (they may block your account one day!)

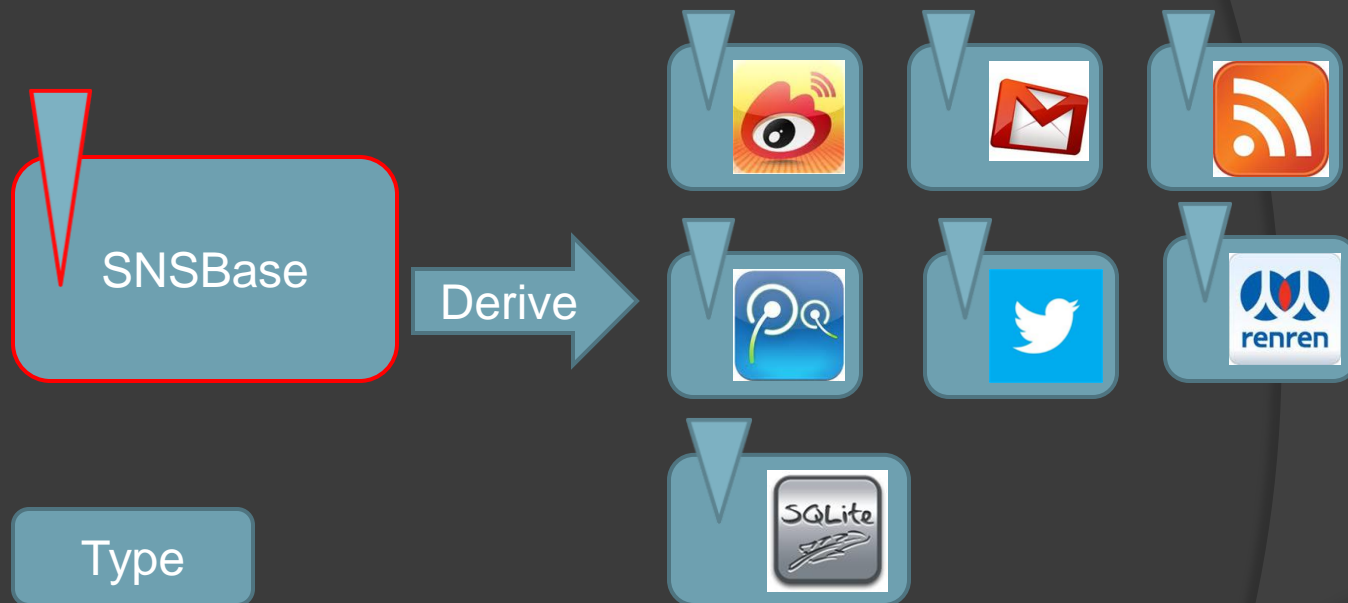


Msg = Read(Renren)  
Write(msg, SQLite)

Application



Interface



Physical



## 1. Python Functions

## 2. STDIN STDOUT

list current  
loaded channels

```
In [2]: list_channel()
```

Current channels:

```
* sina_account_1: SinaWeiboStatus yes
* twitter_account_1: TwitterStatus yes
* email_1: Email yes
* test_feed_2: RSS2RW yes
* qq_account_1: TencentWeiboStatus no
* feed_hpl_renren_zhan: RSS yes
* test_sqlite: SQLite yes
* renren_account_2: RenrenShare yes
* renren_account_1: RenrenStatus yes
```

```
In [3]: print home_timeline(1, channel = "sina_account_1")
```

```
[INFO][20121202-230036][snssocket.py][home_timeline][254]Read 1 statuses
<0>
[纯银V] at Sun, 02 Dec 2012 22:58:14 HKT
    最近又进入密集测试期，不断找到bug的我眼泪掉下来.....
```

```
In [4]: print home_timeline(1, channel = "twitter_account_1")
```

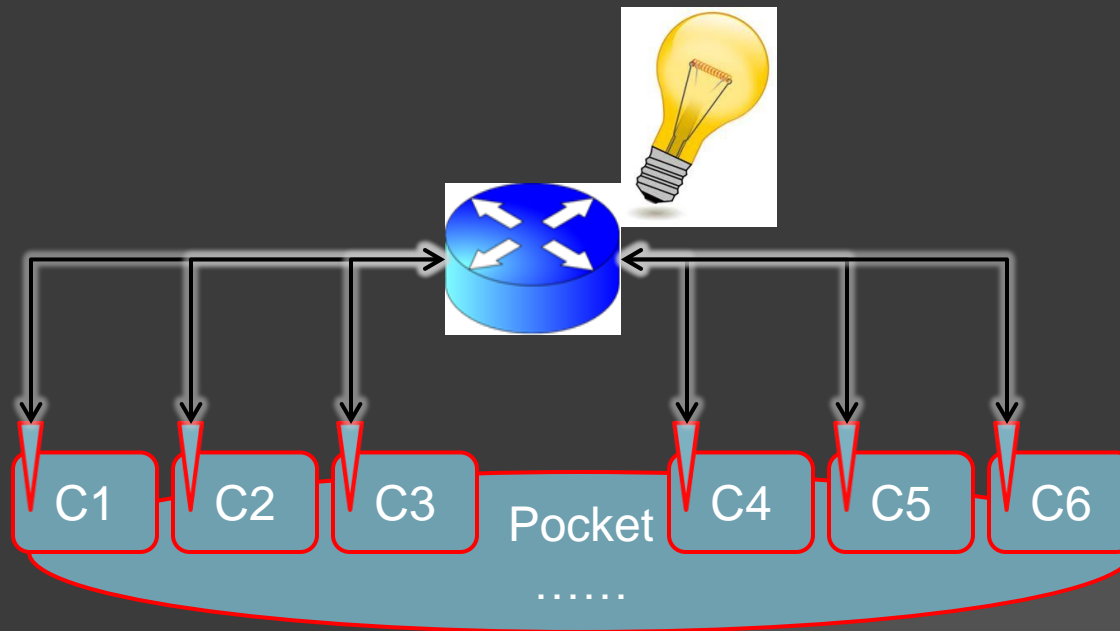
```
[INFO][20121202-230042][snssocket.py][home_timeline][254]Read 1 statuses
<0>
[IDF] at Sun, 02 Dec 2012 22:46:41 HKT
    How does the #Israel Air Force learn from crashes? To find out, we visited the
    IAF Accident Investigations Branch: http://t.co/IxtggqSJ
```

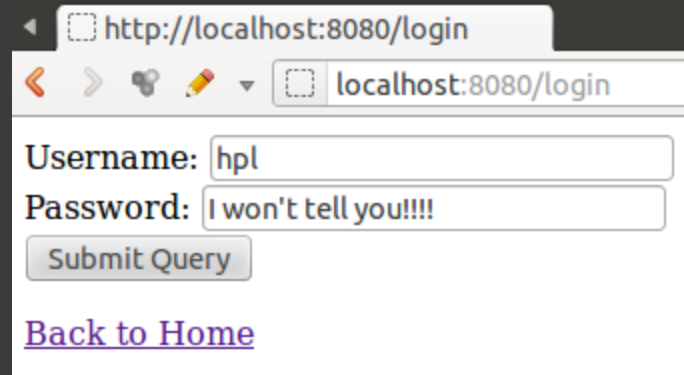
```
In [5]: update("a test from snscli...")
```

update a status

# Motivation -- SNSRouter

- Too many messages
- Different quality.
- Noise.





- Use “bottle” as micro-framework
- FE is all Python; run everywhere


# home\_timeline

Unseen Messages: 26610

[Back to Home](#)

Not informative  
for me

Original


[raw]  百度网盘官方微博 @ Mon, 03 Dec 2012 11:21:36 HKT [Mark as Seen] 0.0894503656755 why?

#度小盘说事#【地铁10号线12月8日停运一天】为配合10号线一期、二期贯通调试，地铁10号线一期将于12月5日采取末班车提前、12月8日采取停运措施。各位亲请提前安排出行计划哦！[挖鼻屎]大家禁得起这次停运一天吗？

[{ mark }](#) [{ gold }](#) [{ silver }](#) [{ bronze }](#) [{ news }](#) [{ interesting }](#) [{ nonsense }](#) [{ tech }](#) [{ data }](#) [{ echo }](#) [{ case }](#)

Sounds  
interesting


[raw]  新浪科技 @ Mon, 03 Dec 2012 11:20:46 HKT [Mark as Seen] 0.25082600543 why?

【IE10广告：让软黑变软饭】微软近日发布一则互联网广告视频，主角是一位软黑，在社交网站不断黑IE10，称IE10烂透了。当看到IE10的各种进步和主流科技媒体对IE10的赞誉后，他终于承认，IE10其实一点也不烂。http://t.cn/zjfbXvN

[{ mark }](#) [{ gold }](#) [{ silver }](#) [{ bronze }](#) [{ new }](#) [{ nonsense }](#) [{ tech }](#) [{ data }](#) [{ echo }](#) [{ case }](#)

Not informative  
for me

[raw]  陈怀临 @ Mon, 03 Dec 2012 11:19:43 HKT [Mark as Seen] 0.0104689387376 why?

[NetScreen的往事（56）] 我真正的强项是million critical system，做数通是玩票。所以对有军工项目背景的人特喜欢。觉得一定是个负责的人。Ning半信半疑的点了点头，又接着去看@读图年代 的美女图了。。。后来告诉我：招了。而且非常不错。Ying，你一定能看到这个weibo，饭票好了没有？。。。




# home\_timeline

Unseen Messages: 26610

[Back to Home](#)

Ranked !!!!!

Recsys fits my recent interest!


[raw]  研究者July @ Mon, 03 Dec 2012 09:49 HKT [Mark as Seen] 0.465108095879 [why?](#)

昨日读书会，陈运文博士从常用的推荐算法(Item/user-based，及LFM等)，到实践中的关键点(数据预处理/冷启动等问题)，相当精彩，PPT下载：<http://t.cn/zjf4vDg>；郝培强则从书的历史讲到未来，不乏启发性和思考性，PPT：<http://t.cn/zjf4vDd>。感谢诸位，来年元旦后，第5期见！Ye || @研究者July：「读书会第4期正式报名通知」时隔半年，久违的读书会第4期终于来了！时间：本周日12月2日下午2点-5点，地点：上海交大徐汇校区工程馆107教室(近包兆龙图书馆)，主题：1、数据挖掘技术在推荐系统中的应用@清风运文；2、做书的历史与未来@tinyfool。转发本微薄报名，详见下图，或<http://t.cn/zO7kGTI>。

[{ mark }](#) [{ gold }](#) [{ silver }](#) [{ bronze }](#) [{ news }](#) [{ interesting }](#) [{ nonsense }](#) [{ tech }](#) [{ data }](#) [{ echo }](#) [{ case }](#)

Submit Query


Industrial news, I may want to follow the link and read further

[raw]  新浪科技 @ Mon, 03 Dec 2012 09:35:02 HKT [Mark as Seen] 0.34421366784 [why?](#)

什么才是不妥协的平板？苹果与微软理念完全不同。苹果擅减法，iPad强化内容消费的定位体验，谷歌等也遵循这一理念；微软则做加法，工作娱乐都想要。RT版Surface意在挑战iPad，应用不兼容却是硬伤；而Pro(电池4小时、售价900美元、重量900克)则完全不像平板，更像超极本。(郑峻) <http://t.cn/zjf2ibE>

[{ mark }](#) [{ gold }](#) [{ silver }](#) [{ bronze }](#) [{ news }](#) [{ interesting }](#) [{ nonsense }](#) [{ tech }](#) [{ data }](#) [{ echo }](#) [{ case }](#)

Tweets from a big shot social network researcher

[raw]  Jure Leskovec @ Mon, 03 Dec 2012 10:23:21 HKT [Mark as Seen] 0.309014101218 [why?](#)

Model brain with 2.5 million neurons configures itself to solve problems. <http://t.co/gARDMlZR>  
<http://t.co/3CCbtWFE>

# Formulation

- Extracted k-D features for N messages:

$$X^T = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix}$$

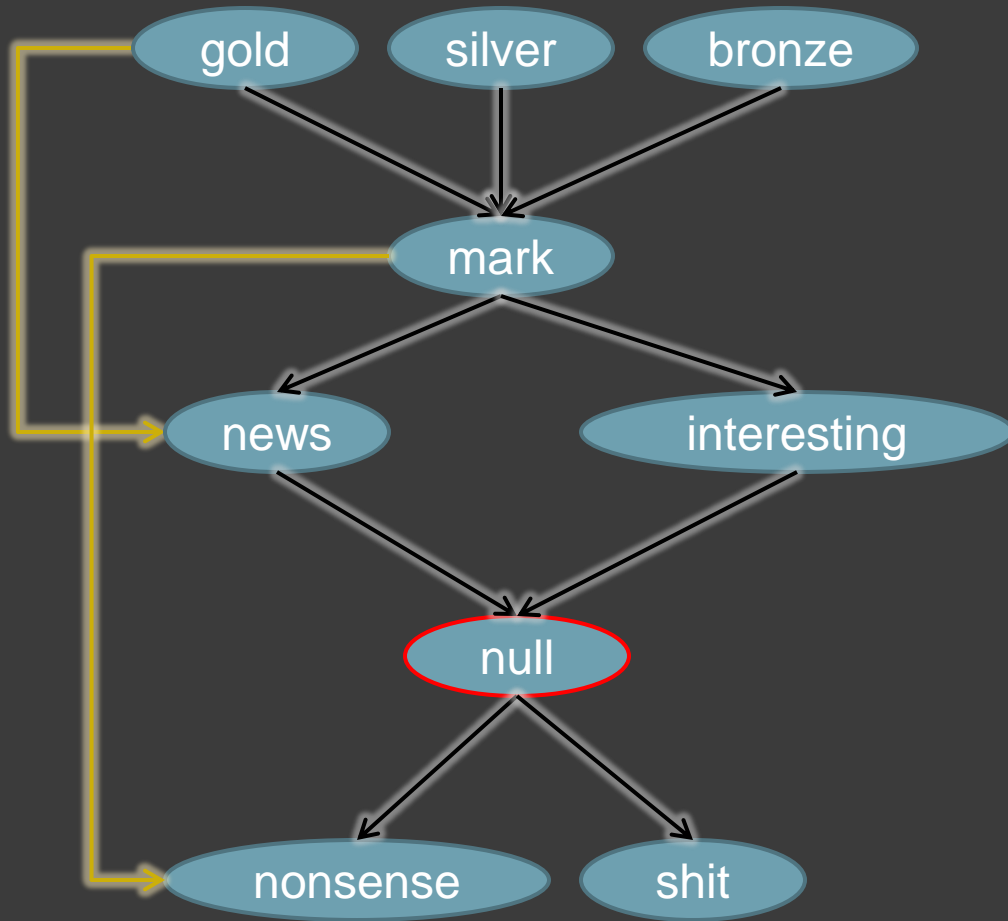
- Linear combination yields a score:

$$y = Xw$$

- The score  $y$  should capture user preference.
- Sort messages by  $y$ .

HOW?

# Graph Induction



User Specified



Graph Induced



e.g.  
Floyd Algorithm

# Formulation

- Induced preference graph:

$$G = \langle V, E \rangle$$

- Linear regression with preference constraint:

$$\begin{aligned} \min_{y, w} \quad & \|y - Xw\|_2^2 \\ \text{s.t.} \quad & y_i > y_j, \forall (i, j) \in E \end{aligned}$$

- Rank Preserving Regression (RPR)

# Re-formulation

- ⊙ Existing solvers?
  - Ordinal regression?
  - Isotonic regression?
- ⊙ Constraint as objective: (indicator function)

$$\min_{y=Xw} \sum_{(i,j) \in E} 1 - \mathbb{I}[y_i > y_j]$$

- ⊙ Approximation by Sigmoid:

$$\min_{y=Xw} f(w) \equiv \sum_{(i,j) \in E} 1 - \text{Sigmoid}[y_i - y_j]$$

# Training

- Gradient Descent: (S short for Sigmoid)

$$\nabla f(w) = \sum_{(i,j) \in E} \nabla f_{ij}(w)$$

$$\nabla f_{ij}(w) = (1 - S[y_i - y_j])S[y_i - y_j](x_j - x_i)$$

- Observation: summation of per pair partial gradient.
- Stochastic Gradient Descent.

# Evaluation

- Kendall's tau correlation coefficient:  
(modified version for our problem)

$$K = \frac{\sum_{(i,j) \in E} \mathbb{I}[y_i > y_j] - \sum_{(i,j) \in E} \mathbb{I}[y_j > y_i]}{|E|}$$

# of correct pairs

# of total pairs

# of incorrect pairs

# Result – Basic Statistics

Item	Value
# of total messages	32533
# of seen messages	7553
# of tagged messages	924
# of forwarded messages	167
# of derived pairs (training)	231540
# of derived pairs (testing)	229009
# of features (+1 noise)	15

Data source: HU Pili personal deployment. Oct 2012 – Dec 2012



# Result – Training with SGD

Item	1.	2.	3.
# of rounds of SGD	200,000	400,000	1,000,000
Wall clock time	32.63s	60.81s	159.57s
Kendall's score (training)	0.8178	0.8349	0.8414
Kendall's score (testing)	0.7598	0.7758	0.7865

Straight SGD implemented in SNSRouter project. Code has not been optimized.

- ◎ Scale linearly
- ◎ Online learning is possible
- ◎ Easy configuration
- ◎ Easy to add new features

# Project Output

- ◎ SNSAPI (5000+ lines)

- A middleware for different SNS



- ...more to come.



- ◎ SNSRouter (2800+ lines)

- A portable web frontend
- Real data collection (1+ month)
- A flexible algorithm framework (RPR-SGD)
- Sample feature extraction modules

# Reference

- ◎ SNSAPI Website: <https://snsapi.ie.cuhk.edu.hk/>
- ◎ SNSAPI Github: <https://github.com/hupili/snsapi/>
- ◎ SNSRouter Github: <https://github.com/hupili/sns-router>

# Related Work

- ◎ IFTTT: <https://ifttt.com/>
- ◎ Yahoo Pipe: <http://pipes.yahoo.com/pipes/>

# Acknowledgements

- ◎ LI Junbo @ BUPT: Cofounder of SNSAPI project.

Q/A?

Thanks

# Add a new feature

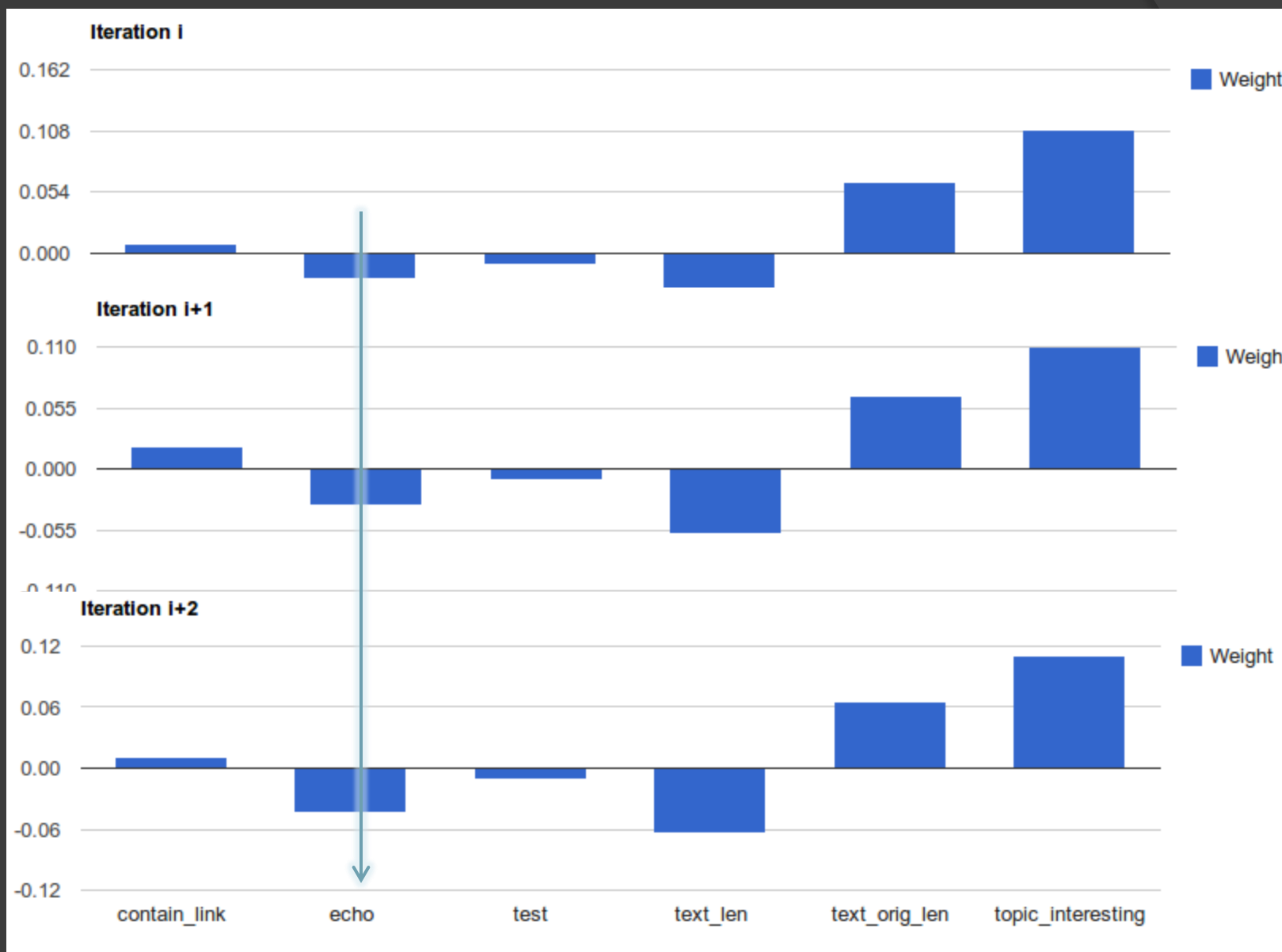
- Avoid echo:
- 1. Add tag “echo”
- 2. Specify preference
- 3. Add feature
- 4. AUTO train we

The screenshot shows a web browser window with the address bar displaying 'http://localhost:8080/config'. The page title is 'Tags'. The main content area contains a table with the following data:

id	name	visible	parent	Toggle
1	null	0	None	<a href="#">Toggle</a>
2	mark	1	None	<a href="#">Toggle</a>
3	gold	1	None	<a href="#">Toggle</a>
4	silver	1	None	<a href="#">Toggle</a>
5	bronze	1	None	<a href="#">Toggle</a>
6	news	1	None	<a href="#">Toggle</a>
7	interesting	1	None	<a href="#">Toggle</a>
8	shit	0	None	<a href="#">Toggle</a>
9	nonsense	1	None	<a href="#">Toggle</a>
10	text	0	None	<a href="#">Toggle</a>
11	tech	1	None	<a href="#">Toggle</a>
12	data	1	None	<a href="#">Toggle</a>
13	echo	1	None	<a href="#">Toggle</a>
14	case	1	None	<a href="#">Toggle</a>

Below the table, there is a form to 'Add new tag:' with an input field containing the text 'echo' and a 'Submit Query' button. On the left side of the browser window, a sidebar is visible with a search bar, a 'Back to Home' link, and a list of tags including 'mark', 'gold', 'silver', 'bronze', 'news', 'interesting', 'shit', 'nonsense', 'text', 'tech', 'data', 'echo', and 'case'.

# Auto Weight Learning



# Features

Name	Description
noise	Random variable [0,1]
echo	Whether the message is from myself
contain_link	Whether the message contain text link
topic_interesting	TF*IDF for {interesting}
topic_tech	TF*IDF for {mark}{gold}{silver}{bronze}
topic_news	TF*IDF for {news}
topic_nonsense	TF*IDF for {nonsense}{shit}
user_interesting	As above; regard “user” as “term”
user_tech	As above
user_news	As above
user_nonsense	As above
text_len	Length of all message (original + retweet)
text_len_clean	Length without face icon, link, @xxx, and punctuation
text_orig_len	Length of original message

# Future Works -- System

- ◎ RESTful interface for all components
  - e.g. one can outsource computationally intensive training to other servers
- ◎ SNSRouter as a platform
  - e.g. can be used to aggregate multiple channels



# Future Works -- Algorithm

- ⦿ Add regularization to alleviate overfitting
- ⦿ Advanced feature extraction.
- ⦿ SGD can do online training.
  - e.g. one sample in, derive some pairs, do SGD on those pairs.
  - Naturally time sliding.

# Why not classification?

- ⦿ Less competitive result (logit) or hard to interpret rules (J48)

A sample branch for mark (3.0/1.0):

```
topic_news <= 0.00603 && topic_tech <= 0.041455  
&& topic_interesting <= 0.042225 && topic_nonsense  
<= 0.010593 && text_len > 0.12 && id <= 30634 &&  
user_tech <= 0.010894 && text_len_clean <= 0.0575  
&& user_tech > 0.001621
```

- ⦿ Hard cut
  - Do not output a “likelihood”
- ⦿ Human can only process sequentially
  - Accurate classification is not needed.