基于多社交媒体的个体信息融合

关键技术研究

学生姓名：詹　辉

指导老师：崔晓晖　教授

内容提要

一、论文背景：

1. 社交媒体的使用现状；
2. 本阶段的社交媒体选择。

二、系统需求与设计：

1. 本系统的需求分析；
2. 本系统的模块设计。

三、系统实现：

介绍本系统关键部分的实现：

1. 数据采集；2. 实体相关词抽取；3. 信息融合。

四、处理结果：

显示本系统的处理结果。

一、论文背景

1. 融合的信息是什么？

（P9. 1.1.1选题的背景）

2. 为什么选择 脸谱网 与 推特 作为现阶段的目标社交媒体？

（P16. 2.3数据来源）

二、系统需求与设计

1. 总体需求

获取一个个体在多个社交媒体中补全后的较为完整的信息。

2. 功能需求

1）数据采集；

2）文本预处理；

3）文本实体抽取；

4）实体相关词抽取；

5）信息融合；

6）处理结果读取；

7）处理结果显示。

（P18. 3.1系统总体功能需求）

3. 模块设计

（P23. 4.1.2系统模块设计）

三、系统实现

- 数据采集
- 文本预处理
- 文本实体识别
- 实体相关词抽取
- 信息融合
- 处理结果读取
- 处理结果显示

1. 推特数据采集

采集地址：mobile.twitter.com

使用工具：requests

2. 脸谱网数据采集

采集地址：m.facebook.com

使用工具：selenium

（P35. 5.2数据采集模块）

三、系统实现

文本预处理

1. 文本切词

使用工具：Noah's ARK 切词工具

2. 单词标准化

使用工具：NLTK 词形还原工具（Lemmatization）

3. 词性标注

使用工具：NLTK 词性标注工具

PR VB RB NN JJ NN

（P38. 5.3.1预处理文本）

文本实体识别

- 实体相关词抽取
- 信息融合
- 处理结果读取
- 处理结果显示

文本实体识别

使用工具：MITIE 命名实体识别工具、NLTK 实体识别工具

PR VB RB NN JJ NN

（P39. 5.3.2.1文本实体识别）

三、系统实现

1. 基于文本词性句型的相关词抽取方法

举例说明：以（句型）NN NN VB JJ NN NN NN NN NN 为例

已训练 My daughter @lucy1995 won the special prize of this year's annual painting competition.

(NN NN VB JJ NN NN NN NN NN)

待抽取 Jonathon Fury sent me a cute ted from Las Vegas Roling-Madam souvenir shop!

(NN NN VB JJ NN NN NN NN NN NN)

Fury 相关词：sent, cute, ted, souvenir, shop.

（P39. 5. 3. 2. 2实体相关词抽取（1）句型方法）

三、系统实现

1. 基于文本词性句型的相关词抽取方法

对500个常见的词形句型进行手动标记，得到每一种句型对应的实体词抽取办法。（使用这500个常见句型的句子占整个语料库的10%）

人工标记：

```
STAT: Start marking 23 pattern.
STAT: Pattern: NN NN NN VB NN NN NN NN JJ.
STAT: Mark 7 text of the pattern.
        0               1       2       3       4
        NN              NN      NN      VB      NN
A       Glenview        man     wa      found   shot    to
5               6                       7               8
NN              NN                      NN              JJ
death   in      Chicago over    the     weekend :       trib.in/1GZv0rW
The indice of relevance: 1 3 5 7
```

（P39. 5. 3. 2. 2实体相关词抽取（1）句型方法）

三、系统实现

2. 基于卡方统计量的相关词抽取方法

卡方统计量常运用在检验两个事件的相互独立性。利用卡方统计量计算实体词与实体词所在文本中的每一个词的相互独立性，得到的结果中与实体词不相互独立的词，即为该实体的相关词。

举例说明：

在整个语料库中统计“实体词”与“词A”一起出现、一起不出现的频率，如果“词A”在语料库中明显呈现出与“实体词”同时出现以及同时不出现，则可断定“词A”与“实体词”不是相互独立的。

实体词与词A同时出现	56	实体词出现、词A不出现	7
实体词不出现、词A出现	21	实体词与词A同时不出现	19450

实体词与词B同时出现	1	实体词出现、词B不出现	58
实体词不出现、词B出现	41	实体词与词B同时不出现	19436

（P41. 5. 3. 2. 2实体相关词抽取（2）卡方方法）

三、系统实现

2. 基于卡方统计量的相关词抽取方法

卡方统计量则是计算两个词之间是否相互独立的量化数值，若卡方统计量超过10.83，则该二词相互独立的概率将低于0.001，近似与小概率事件。公式如下：

e_{11}计算公式（e_{10}、e_{01}、e_{00}计算公式以此类推）

$$e_{11} = n * \frac{n_{11} + n_{10}}{n} * \frac{n_{11} + n_{01}}{n}, \quad n = n_{11} + n_{10} + n_{01} + n_{00}$$

X^2计算公式

$$\chi^2 = \sum_{e_t \epsilon \{0,1\}} \sum_{e_c \epsilon \{0,1\}} \frac{(n_{e_t e_c} - e_{e_t e_c})^2}{e_{e_t e_c}}$$

在上例中，通过计算可得：

词A与实体词的卡方统计量为：12606.48

词B与实体词的卡方统计量为：6.04

（P41. 5.3.2.2实体相关词抽取（2）卡方方法）

三、系统实现

信息融合

处理结果读取

处理结果显示

信息融合

在两个社交媒体中补全该个体所提到的信息：

1. 在两个社交媒体中提到的相同实体的不同相关词；
2. 在两个社交媒体中提到的不同实体。

（P42. 5. 3. 3信息融合）

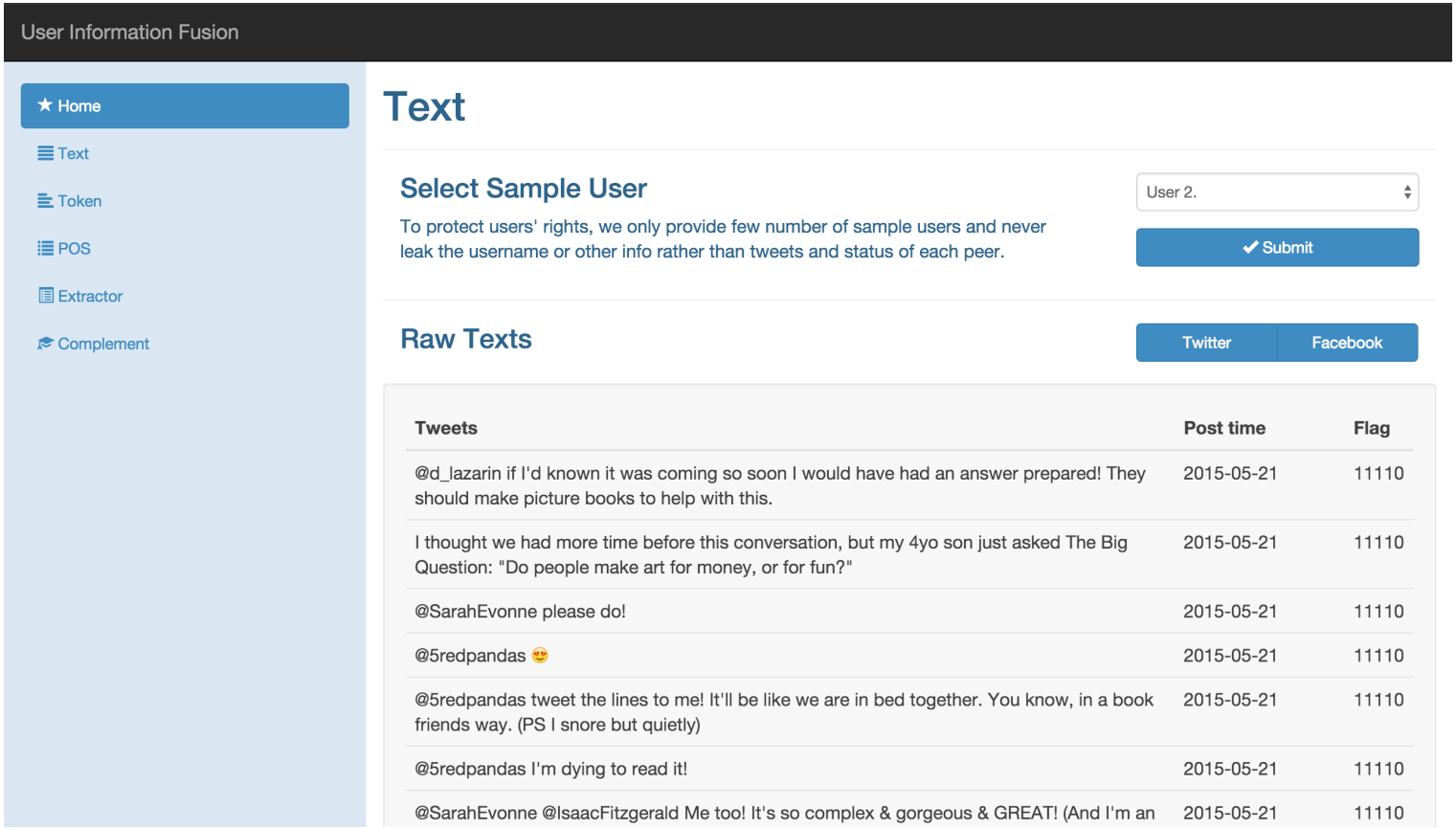
处理结果读取与显示

使用框架：flask

（P43. 5. 4处理结果显示模块）

四、处理结果

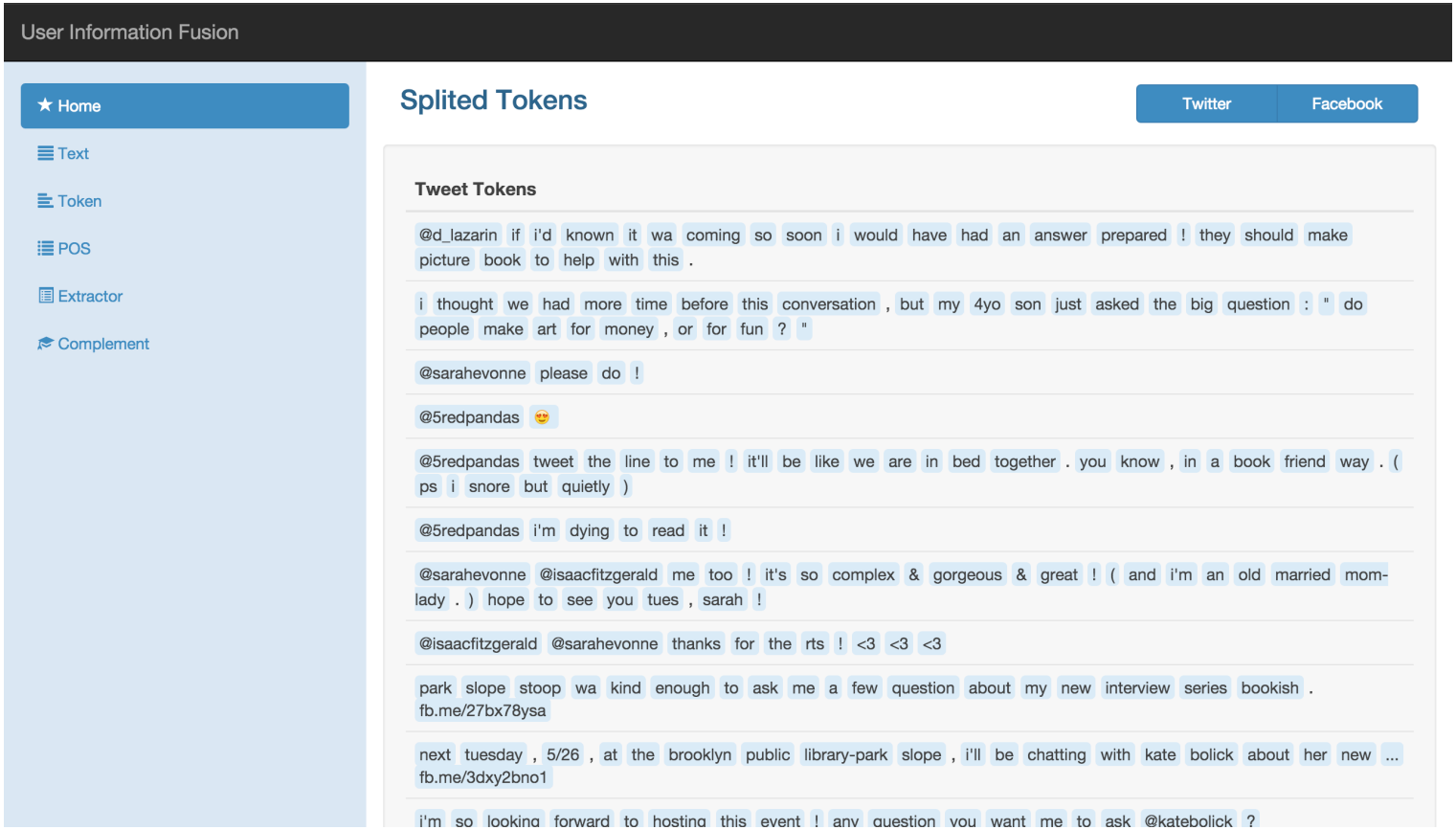
1. 原始文本



（P44. 5. 4. 4. 1原始社交文本）

四、处理结果

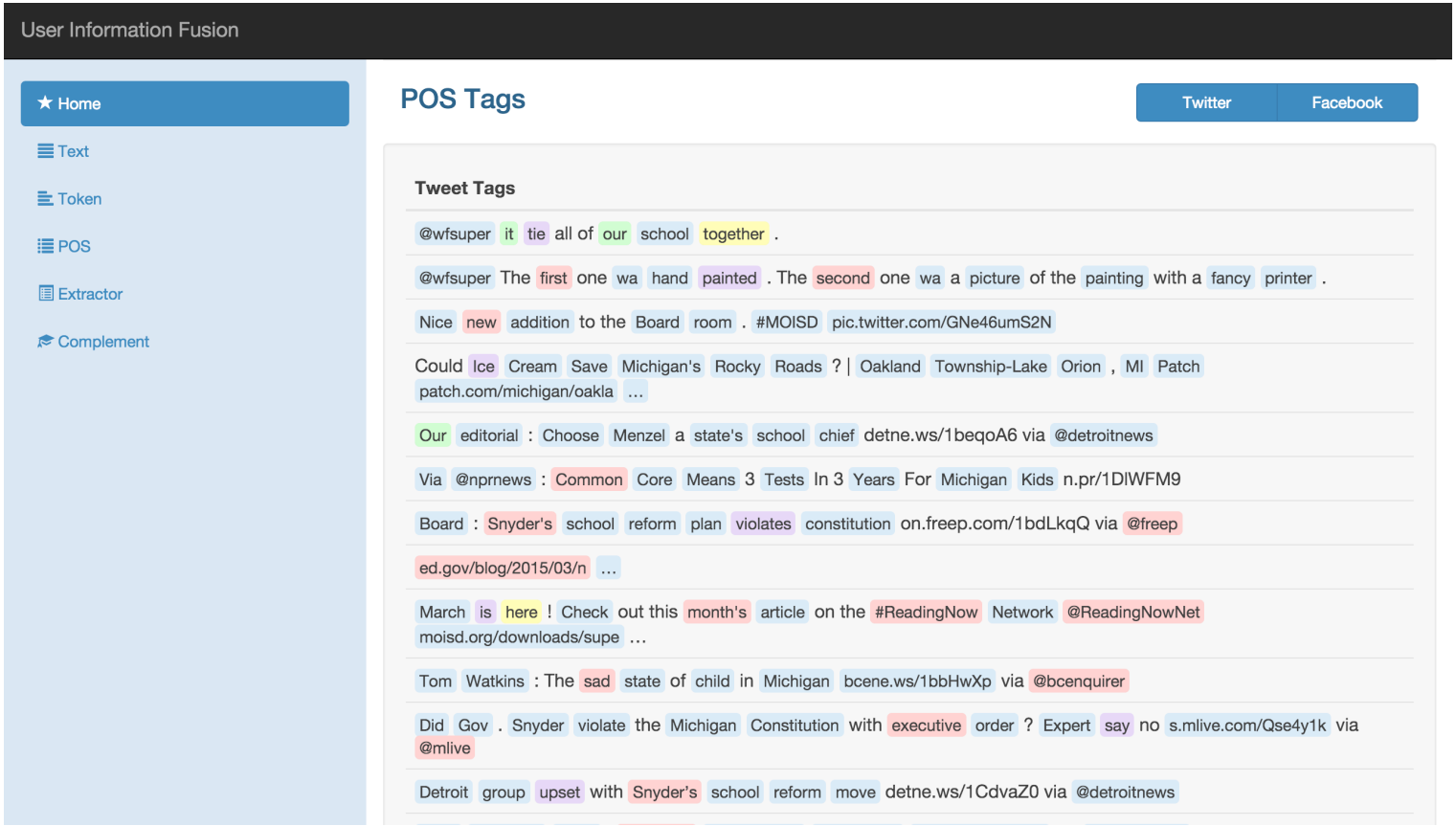
2. 中间结果：文本切词与单词标准化



（P45. 5. 4. 4. 2文本切词与单词标准化）

四、处理结果

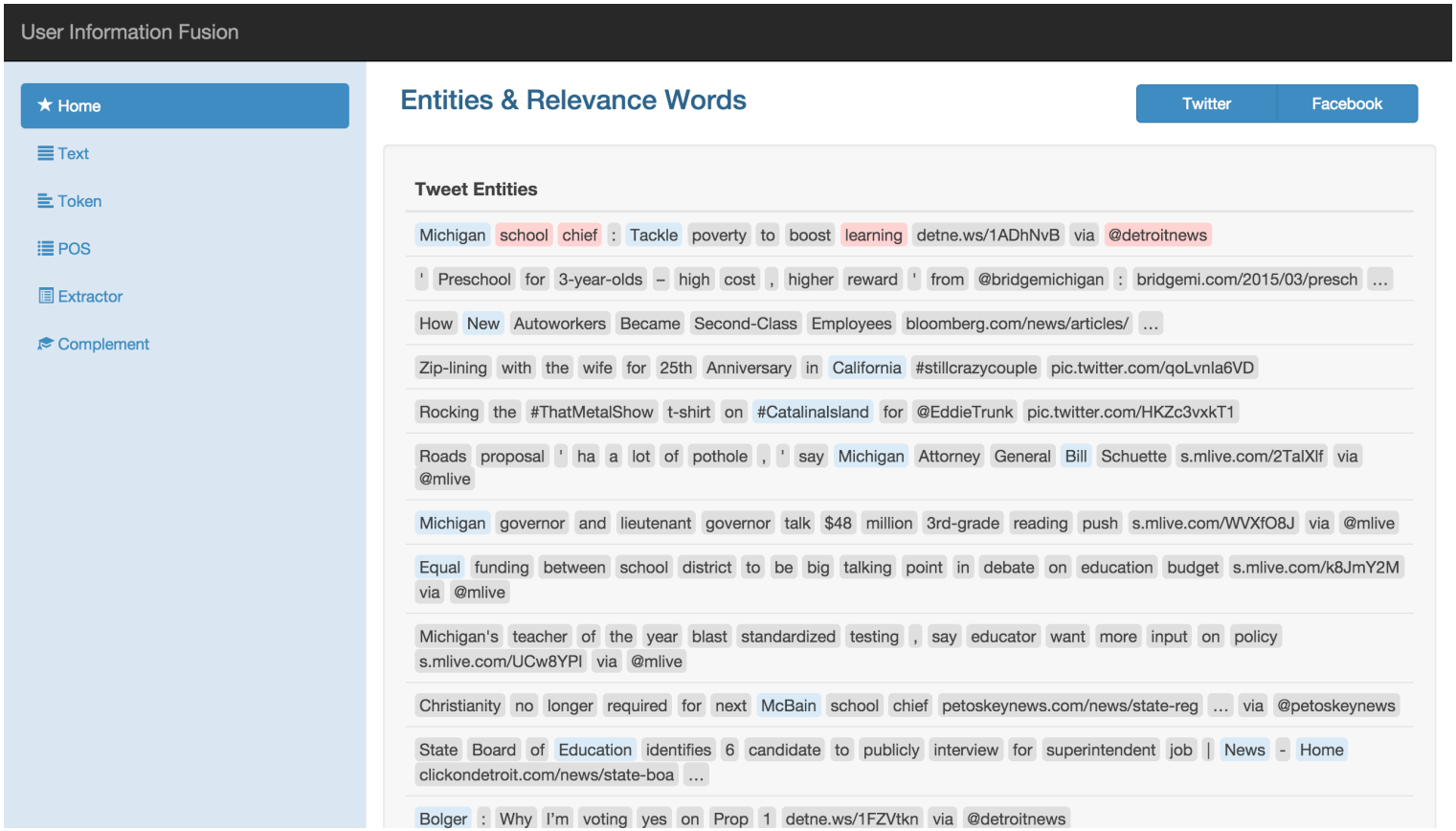
2. 中间结果：文本词性标注



（P45. 5. 4. 4. 3文本词性标注）

四、处理结果

2. 中间结果：文本实体识别与实体相关词抽取



（P46. 5. 4. 4. 4信息抽取）

四、处理结果

3. 最终结果：信息融合

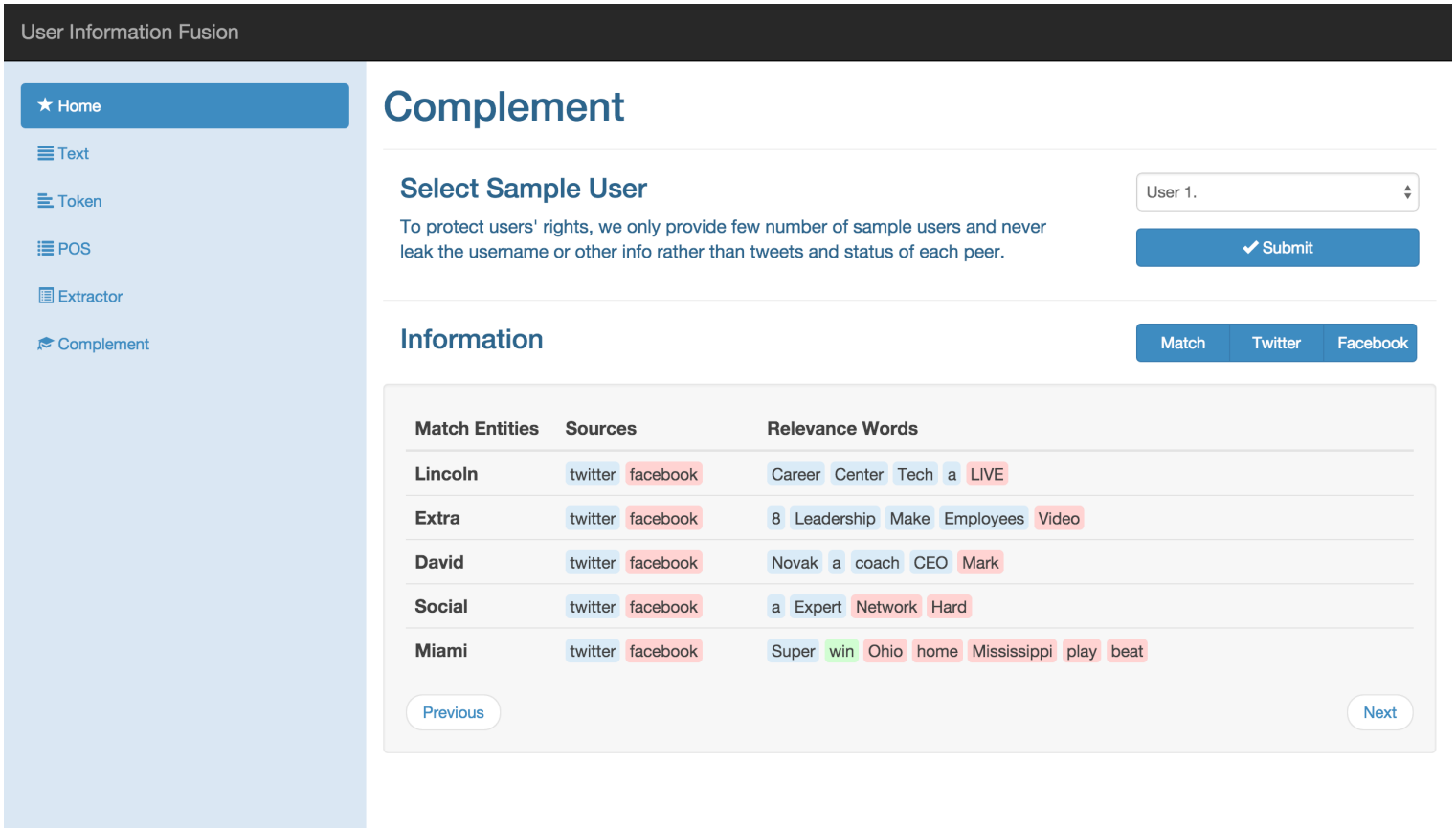
1）是否能够表现融合的两个账户属于统一个个体；

2）是否能够补全一个个体在两个社交媒体中的不同信息。

（P20. 3. 6信息融合功能）

四、处理结果

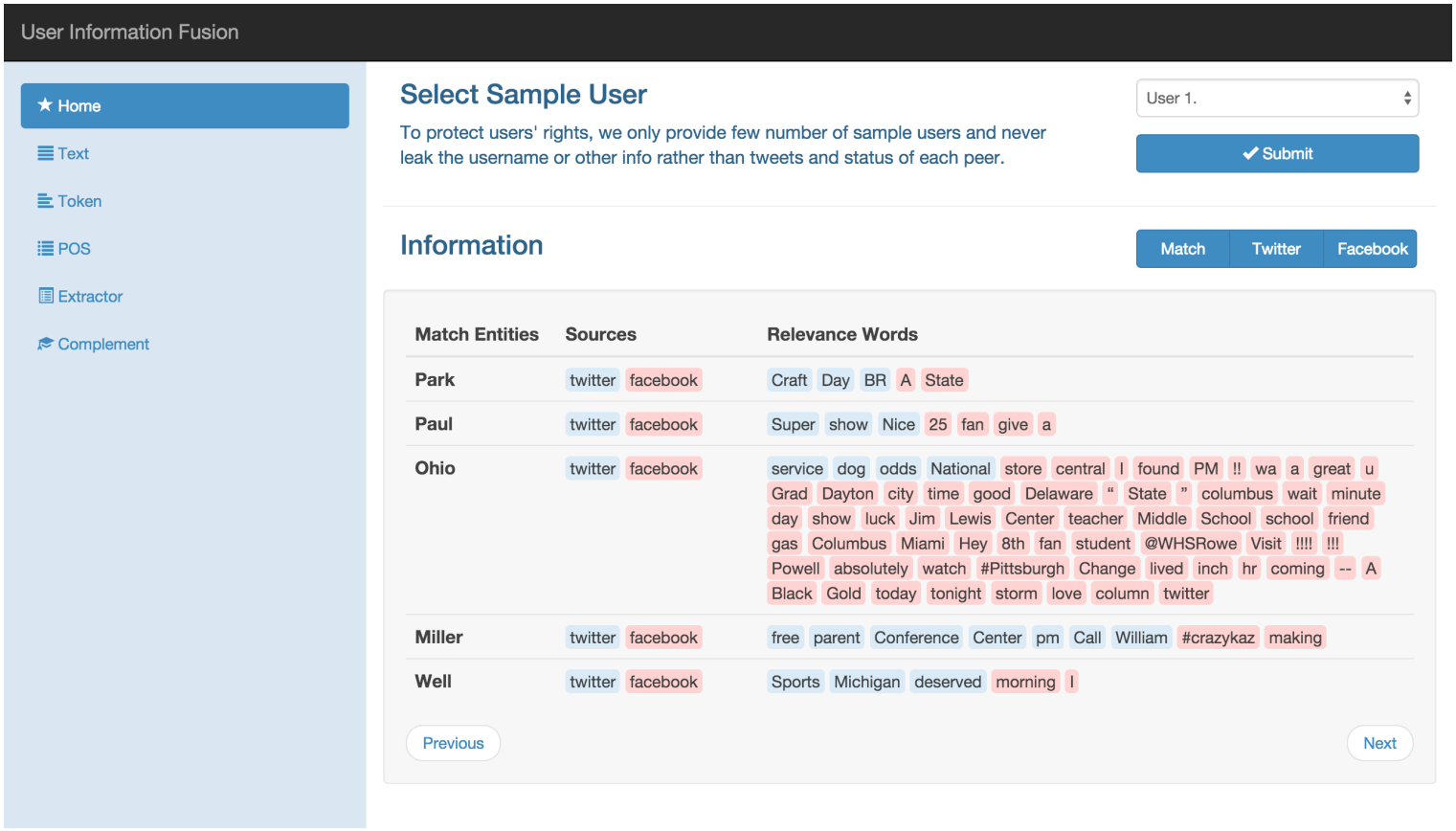
3. 最终结果：信息融合 User1（不同个体的账号）



（P46. 5. 4. 4. 5信息融合）

四、处理结果

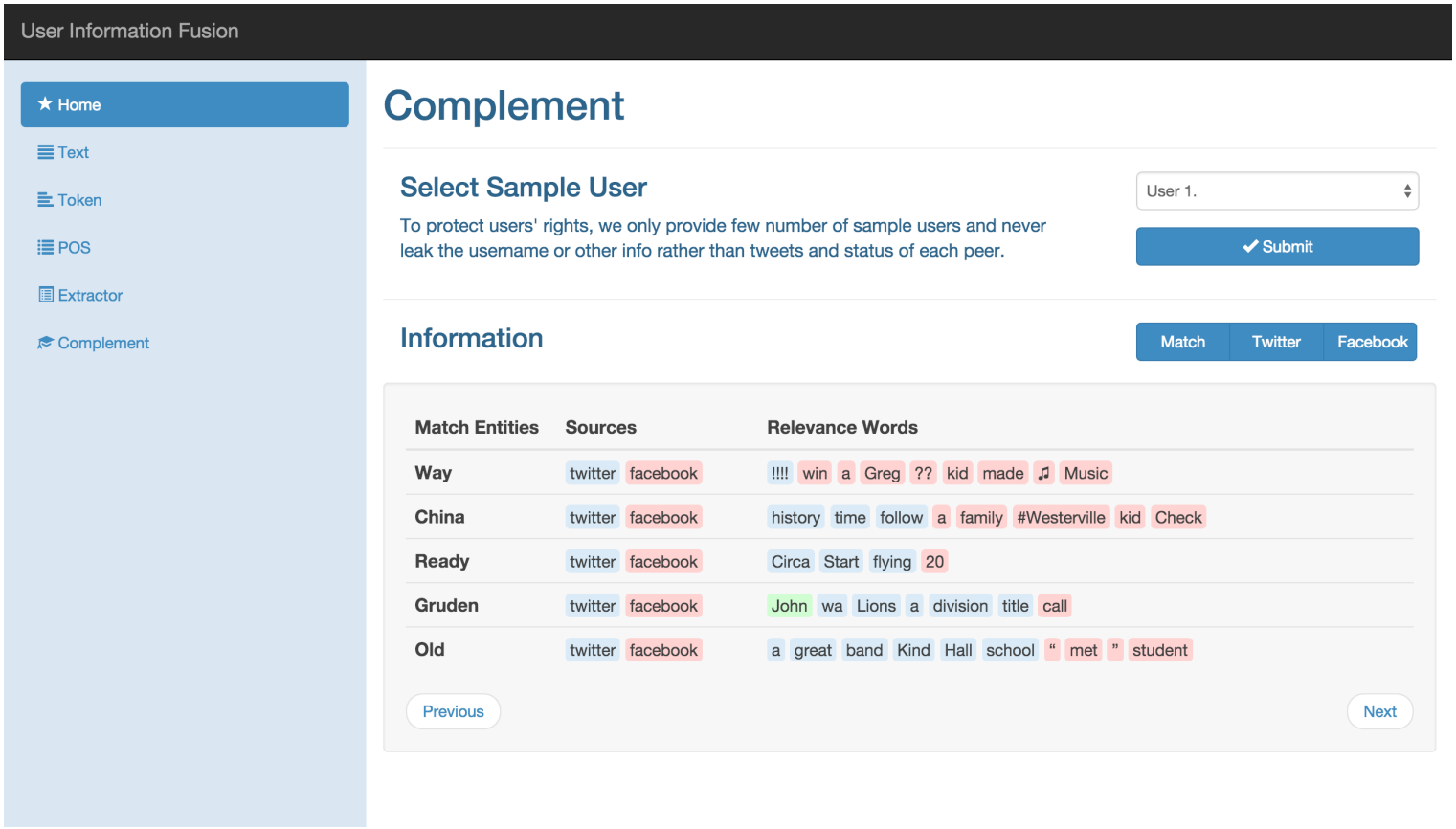
3. 最终结果：信息融合 User1（不同个体的账号）



（P46. 5. 4. 4. 5信息融合）

四、处理结果

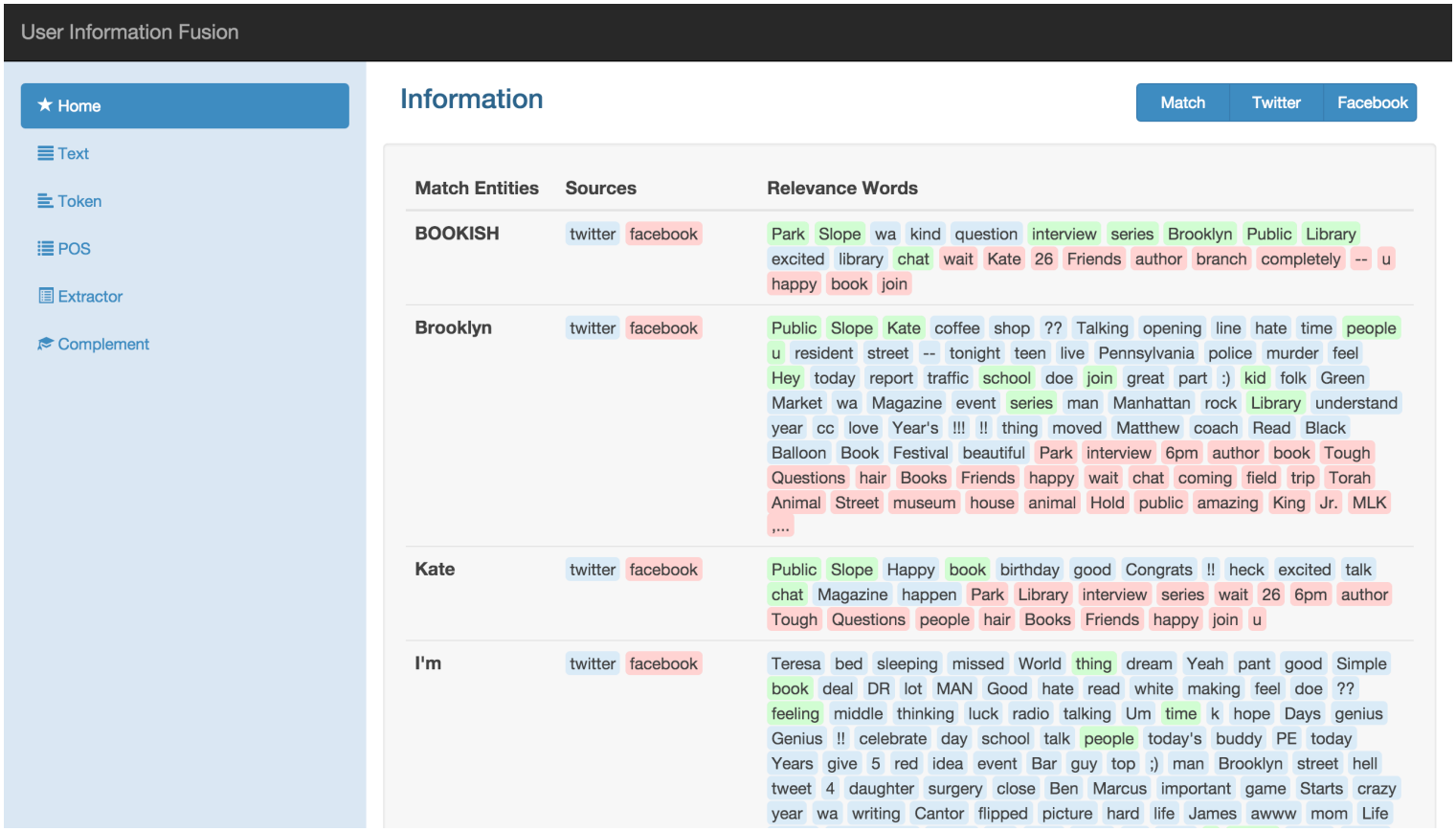
3. 最终结果：信息融合 User1（不同个体的账号）



（P46. 5. 4. 4. 5信息融合）

四、处理结果

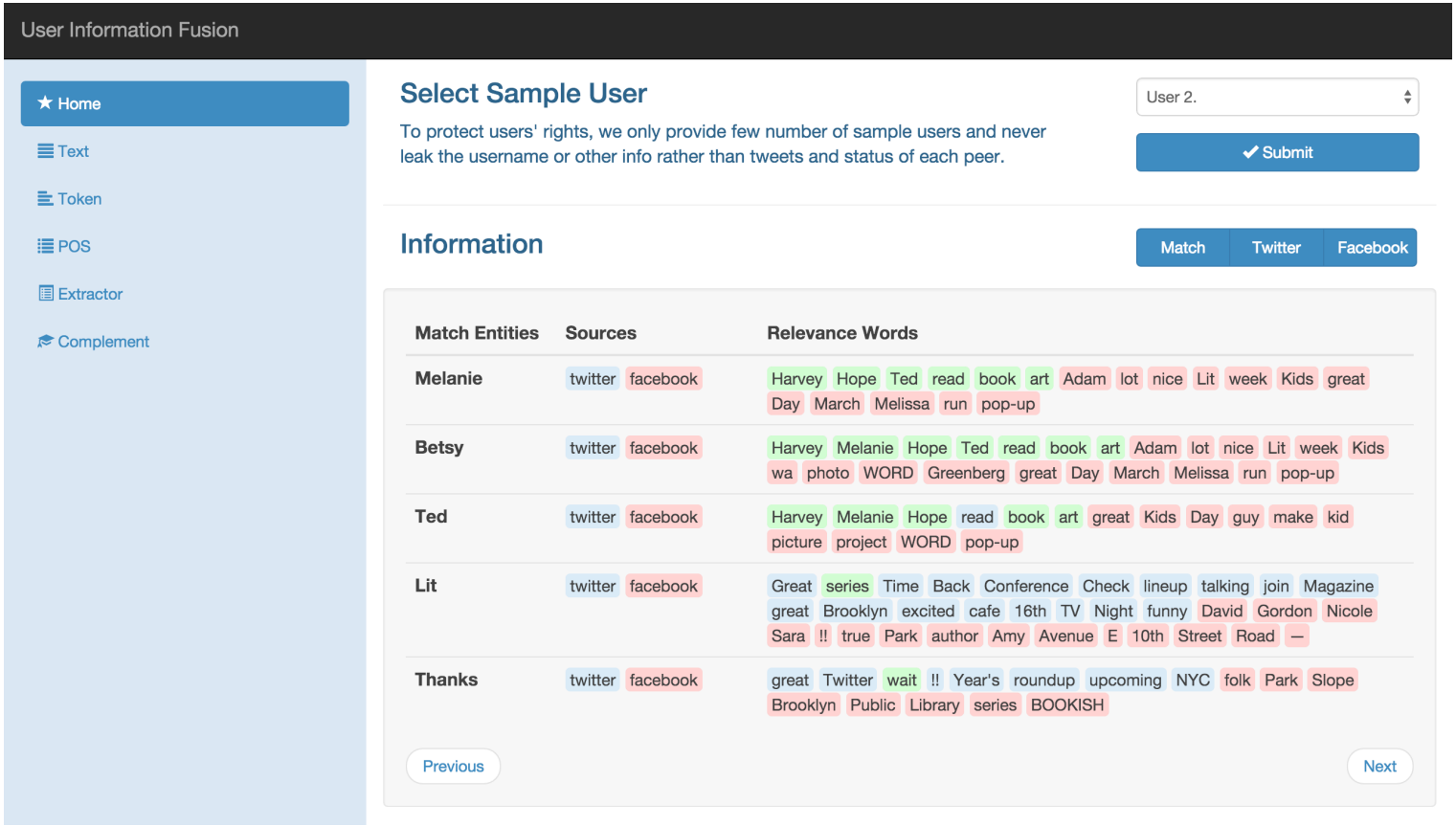
3. 最终结果：信息融合 User2（同一个个体的账号）



（P46. 5. 4. 4. 5信息融合）

四、处理结果

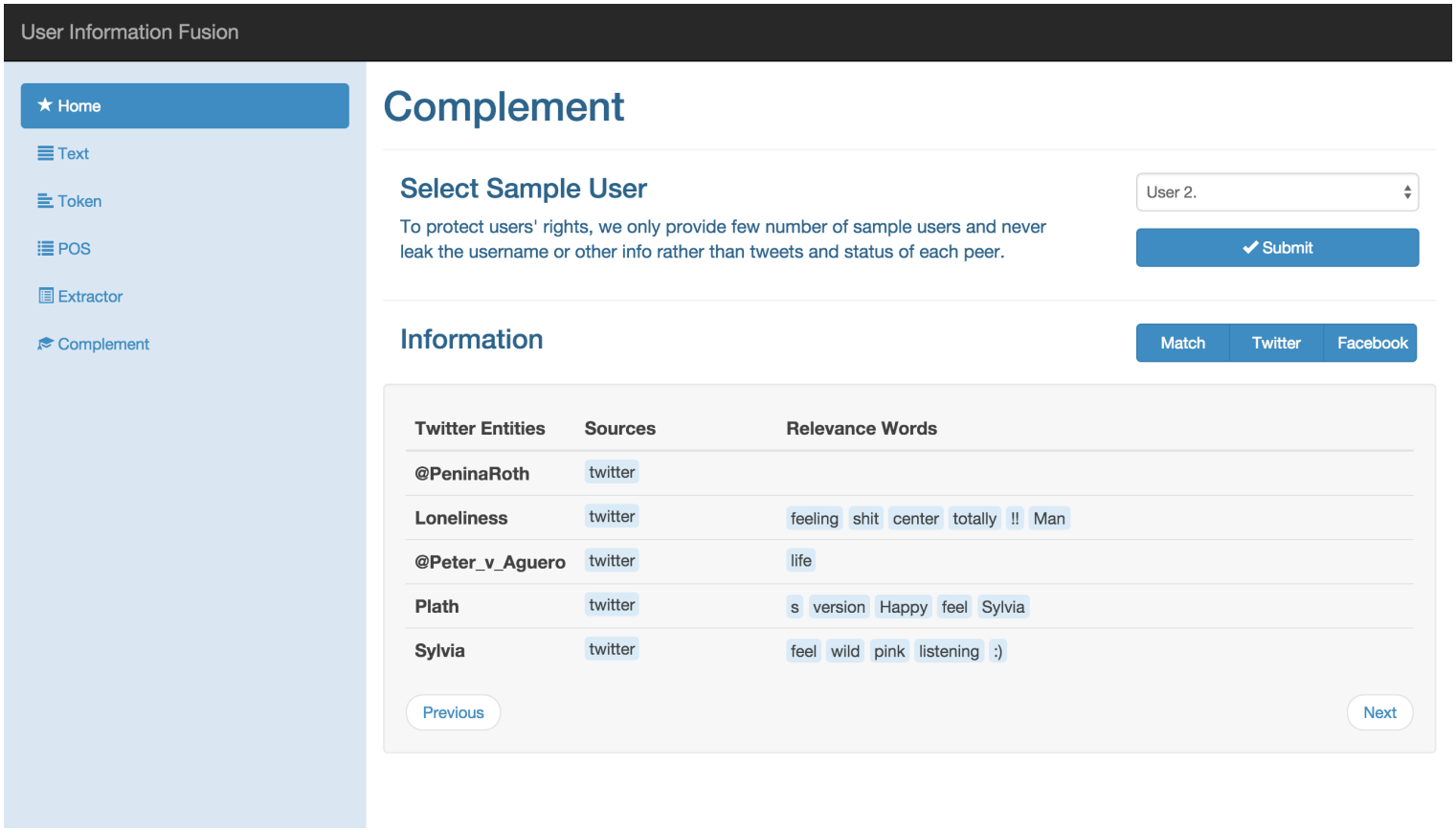
3. 最终结果：信息融合 User2（同一个个体的账号）



（P46. 5. 4. 4. 5信息融合）

四、处理结果

3. 最终结果：信息融合 User2（同一个个体的账号）



（P46. 5. 4. 4. 5信息融合）

结束

致 谢

THE END