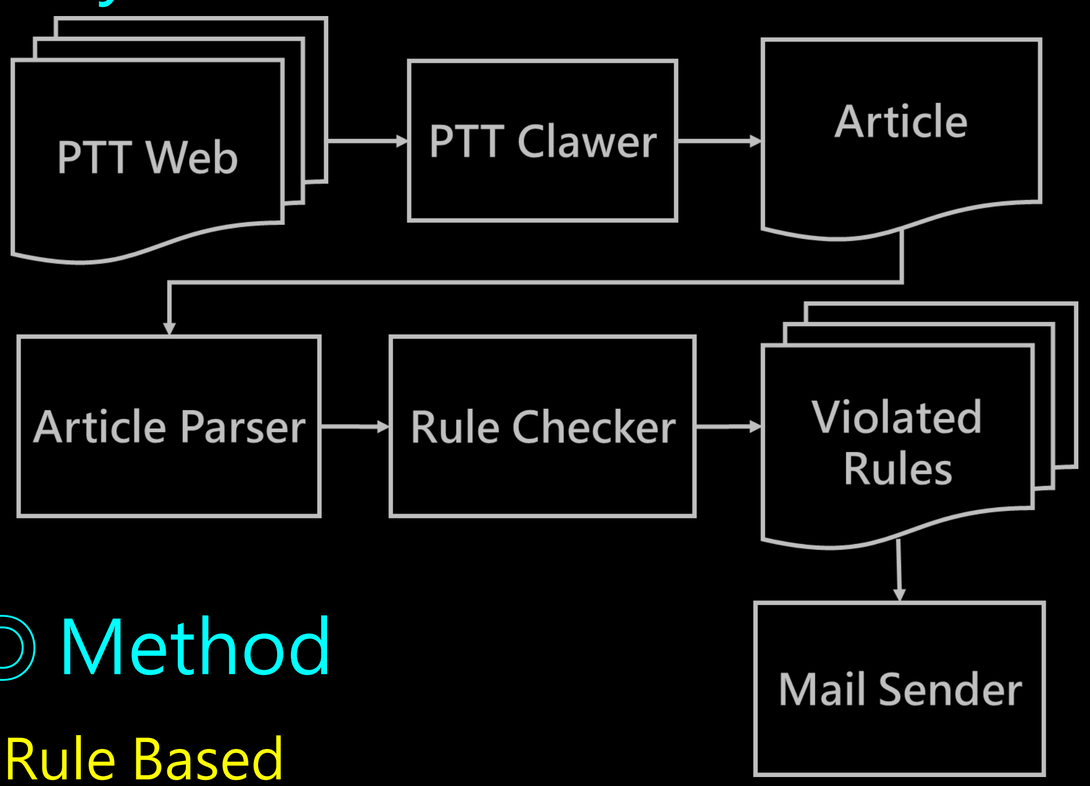
作者	Ching Lin, Yen Chao, Ming-Je, Hsin-Huei	看板	Gossiping
標題	PTT Moderator Helper		
日期	Tue Jun 23 18:30:00 2015		

◎ Motivation

PTT is the most popular BBS site in Taiwan. With over 1,000,000 users, it is a heavy burden for moderators to maintain order. We aim to create a helper program which notices the PTT forum moderator when there's any posts against the forum rules. Our target is a subset of the rules of the Gossiping forum.

◎ System Overview



◎ Method

- Rule Based

Some forum rules are based on simple statistical data such as number of words, number of posts..., etc. We collect these rule and formulate them to program.

版規九：繁體中文未滿 20 字，退回並水桶六個月

```
tradCharsCount = count.traditional_chinese_chars(article.body)
  if tradCharsCount < 20:
     violated = True
```

Input: PTT articles

"作者": "barkingdog(創世截顱南宮毅)",
"日期": "2015-5-19 19:56:46",
"標題": "[問卦]請問有正晶店的八卦嗎",
"內文": "台北市市民大道五段有一家店叫正晶店耶\n\nhttp://i.imgur.com/BvfYh49.jpg\n",

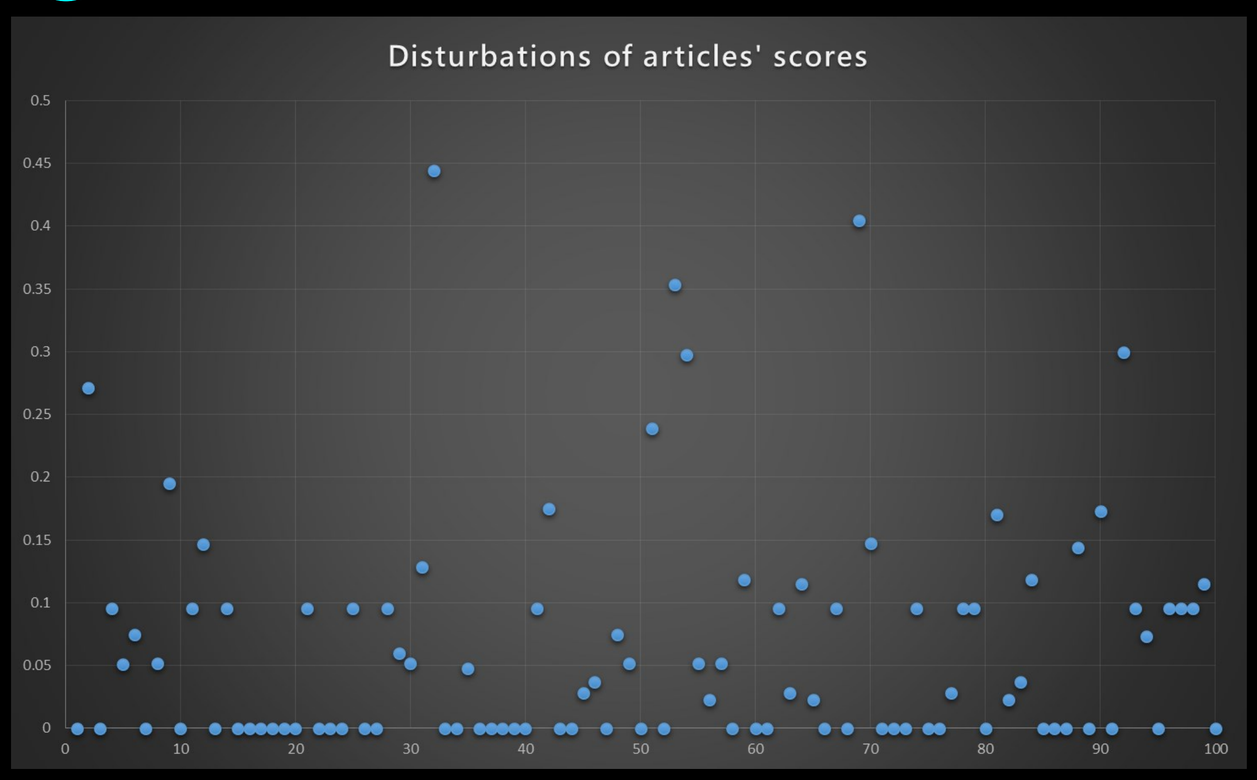
Output: violated rules

版規九：繁體中文未滿 20 字，退回並水桶六個月

- TF-IDF Keyword Extraction

Term frequency-inverse document frequency (TF-IDF) is a numerical statistic that reflects how important a word is to a document in a collection or corpus. With this method, our system can find keywords from thousands of articles which are in the same topic. We use articles from HatePolitics to extract keywords about politics.

◎ Results



		Ground Truth		
		Politics Related	Not Politics Related	
System Result	Politics Related	9	9	18
	Not Politics Related	2	80	82
		11	89	

TPR = 9/11 = 0.818 FPR = 9/80 = 0.112