電子郵件行銷
spam2ham

第 18 組

數學三 陳鵬仁、會計三 林子昕、生傳三 黃韻文、資管四 丁啟恩、資管三 陳奕兆

文獻探討

如何避免成為垃圾郵件：

https://buzzorange.com/techorange/2021/03/05/email-ads-esp/

使用 NB 做分類器：

1. Detecting Spam by Weighting Message Words (2009)

    - M.Abdoh, M.Musa, N.Salman

2. Machine learning for email spam filtering: review, approaches and open
    research problems (2019)
    - E.Dada, J.Bassi, H.Chiroma, S. Abdulhamid, A.Adetunmbi, O.Ajibuwa

垃圾郵件分類法 - 貝氏過濾器：

https://www.ithome.com.tw/tech/28999

垃圾郵件檢測器與 trigger_word 資料集：

https://mailmeteor.com/spam-checker

研究背景

行銷效益

- 快速且大量
- 成本低廉
- 開發新客群

現況限制

垃圾郵件篩選器日新月異，多面相檢測郵件，不易繞過檢查機制

缺口

- 研究集中於更好的分類器
- 少見商業化的自動調整服務

垃圾郵件篩選器

判定標準

發送方式

- 大量發送
- 沒有標題
- 含亂碼

寄件人

- 匿名
- 網站聲譽
- 目標對象

內容

- 顏色、字體
- **spam 詞彙**
- **大量連結**
- 大量圖片

附件

- 檔案太大
- 大量使用

專案
預期目標

1 寫一個分類器進行郵件分類

2 以分類高 precision 為目標

3 調整 testing data 文本資料

4 目標為混淆模型判斷

5 降低 recall

資料集前處理

資料簡介

來源

kaggle dict

內容

ham: 2,551
spam: 500

型態

text/plain
text/html
multipart/signed
multipart/alternative

資料集前處理

1

import email.policy

完全符合 RFC 文件

2

email.parser. BytesParser (policy)

policy=
email.policy.default
default 為 compat32

3

payload = email_file. get_payload()

可以取得payload

去除 RFC 格式的標頭，得到內容


```
        bwmam05.mailsvc.email.bigpond.com(MailRouter V3.0n 44/32989362);
        22 Aug 2002 23:37:51
Received: (from tony@localhost) by hobbit.linuxworks.com.au
        (8.11.6/8.11.6) id g7MDaWX26868; Thu, 22 Aug 2002 23:36:32 +1000
Message-Id: <200208221336.g7MDaWX26868@hobbit.linuxworks.com.au.nospam>
To: Exmh Users Mailing List <exmh-users@example.com>
From: Tony Nugent <tony@linuxworks.com.au>
X-Face: ]IrGs{LrofDtGfsrG!As5=G'2HRr2zt:H>djXb5@v|Dr!jOelxzAZ`!}("]}]
    Q!)1w#X;)nLlb'XhSu,QL>;)L/l06wsI?rv-xy6%Y1e"BUiV%)mU;]f-5<#U6
    UthZ0QrF7\_p#q}*Cn}jd|XT~7P7ik]Q!2u%aTtvc;)zfH\:3f<[a:)M
Organization: Linux Works for network
X-Mailer: nmh-1.0.4 exmh-2.4
X-Os: Linux-2.4 RedHat 7.2
In-Reply-To: message-id <200208212046.g7LKkqf15798@mail.banirh.com> of Wed,
    Aug 21 15:46:52 2002
Subject: Re: Insert signature
X-Loop: exmh-users@example.com
Sender: exmh-users-admin@example.com
Errors-To: exmh-users-admin@example.com
X-Beenthere: exmh-users@example.com
X-Mailman-Version: 2.0.1
Precedence: bulk
Reply-To: exmh-users@example.com
List-Help: <mailto:exmh-users-request@example.com?subject=help>
List-Post: <mailto:exmh-users@example.com>
List-Subscribe: <https://listman.example.com/mailman/listinfo/exmh-users>,
    <mailto:exmh-users-request@redhat.com?subject=subscribe>
List-Id: Discussion list for EXMH users <exmh-users.example.com>
List-Unsubscribe: <https://listman.example.com/mailman/listinfo/exmh-users>,
    <mailto:exmh-users-request@redhat.com?subject=unsubscribe>
List-Archive: <https://listman.example.com/mailman/private/exmh-users/>
Date: Thu, 22 Aug 2002 23:36:32 +1000

On Wed Aug 21 2002 at 15:46, Ulises Ponce wrote:

> Hi!
>
> Is there a command to insert the signature using a combination of keys and not
> to have sent the mail to insert it then?

I simply put it (them) into my (nmh) component files (components,
replcomps, forwcomps and so on).  That way you get them when you are
editing your message.  Also, by using comps files for specific
folders you can alter your .sig per folder (and other tricks).  See
the docs for (n)mh for all the details.
```

```
Re: Insert signature
On Wed Aug 21 2002 at 15:46, Ulises Ponce wrote:

> Hi!
>
> Is there a command to insert the signature using a combination of keys and not
> to have sent the mail to insert it then?

I simply put it (them) into my (nmh) component files (components,
replcomps, forwcomps and so on).  That way you get them when you are
editing your message.  Also, by using comps files for specific
folders you can alter your .sig per folder (and other tricks).  See
the docs for (n)mh for all the details.

There might (must?) also be a way to get sedit to do it, but I've
been using gvim as my exmh message editor for a long time now.  I
load it with a command that loads some email-specific settings, eg,
to "syntax" colour-highlight the headers and quoted parts of an
email)... it would be possible to map some (vim) keys that would add
a sig (or even give a selection of sigs to choose from).

And there are all sorts of ways to have randomly-chosen sigs...
somewhere at rtfm.mit.edu... ok, here we go:
rtfm.mit.edu/pub/usenet-by-group/news.answers/signature_finger_faq.
(Warning... it's old, May 1995).

> Regards,
> Ulises

Hope this helps.

Cheers
Tony


_____
Exmh-users mailing list
Exmh-users@redhat.com
https://listman.redhat.com/mailman/listinfo/exmh-users
```

訓練
模型

1 training data: 400 筆的 ham 和 400 筆的 spam

2 以訓練好的模型測試 testing data

3 k-fold validation 以高 precision 為目標

4 共測試四種模型

訓練結果

4 種模型訓練結果

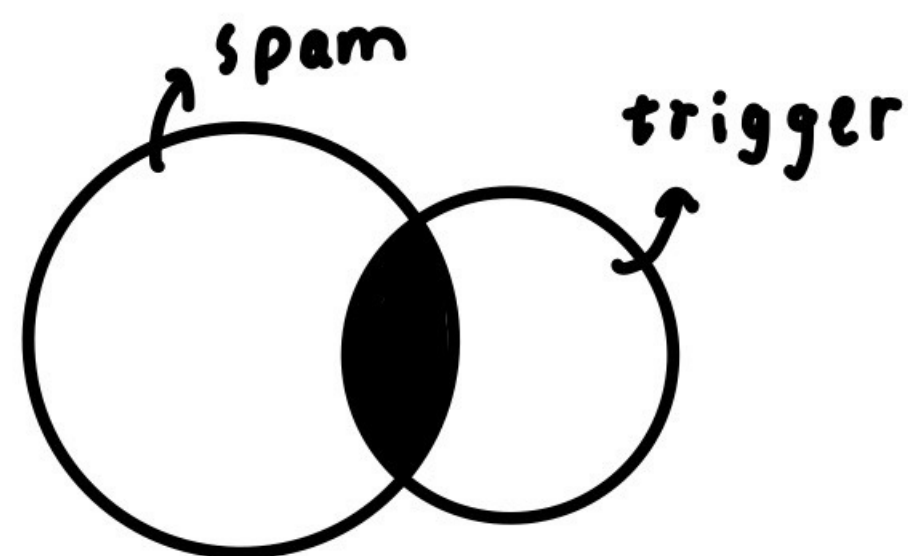| Multinomial NB	Bernoulli NB	TFIDF to SVM	SVD to SVM
`fit_prior=False`	`fit_prior=False`	`kernel='rbf'`	`kernel='rbf'` `C=2.0`
precision: 1.00 recall: 0.96	precision: 0.57 recall: 0.99	precision: 0.86 recall: 1.00	precision: 0.93 recall: 0.99

選用 Multinominal NB

文獻回顧

- 極高準確度 (95% - 99.95%)
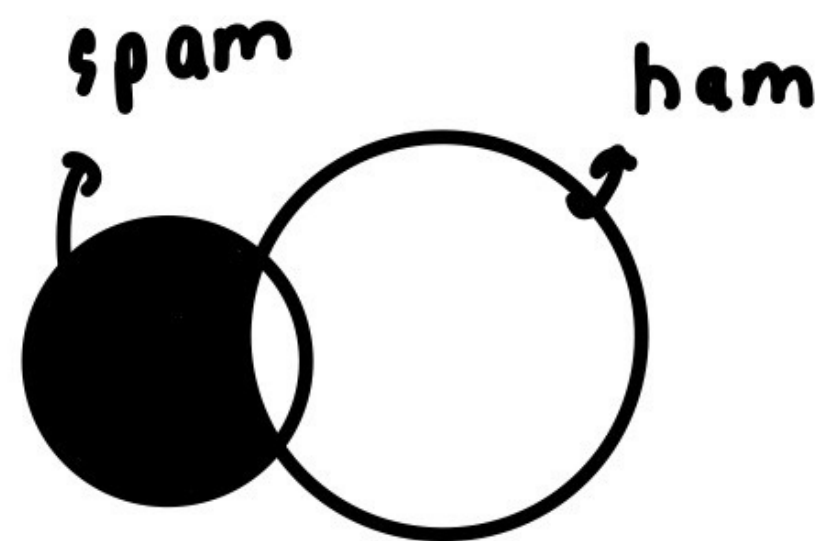- 學習方式容易、速度快
- 針對產業調整、不易被穿透

實作結果

- Testing precision 最高
- 避免將 ham 分入 spam

文本調整方法

降低與 spam 相似度

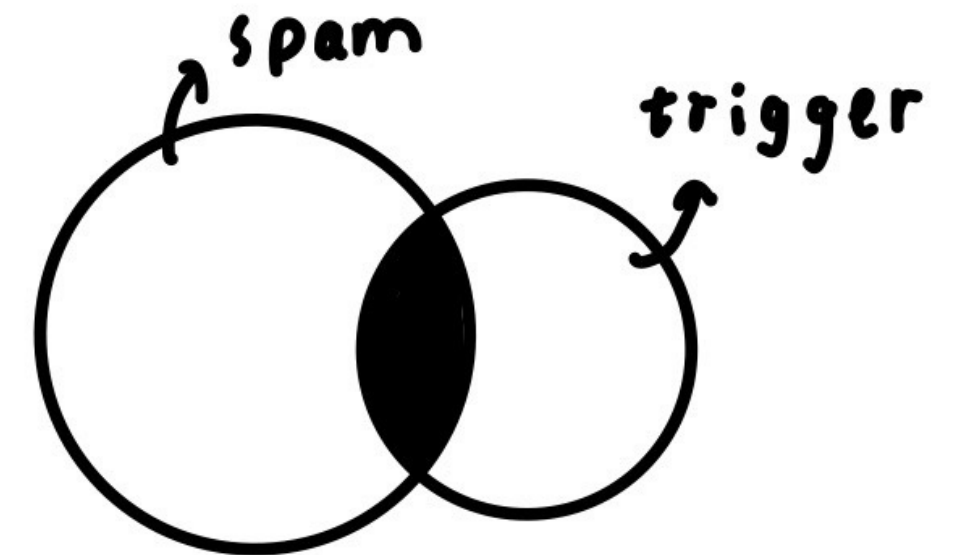spam 交集 trigger_word

降低與 spam 相似度

spam 差集 ham

提高與 ham 相似度

ham 差集 spam

調整方式 #1

1 載入 spam 常出現的字詞 (trigger word)

2 對 spam 斷詞 collections.Counter( )

3 抽換 spam 和 trigger_word 交集

4 調整 testing spam data，重新預測結果

方法一結果

同義字抽換

- word2vec 查找
- 手動抽換
- 不在 trigger word

Δ recall: -1%


```
adjusted_word = ['Access', 'Act', 'Buy', 'Call', 'Click', 'Immediately', 'Instant', 'Now', 'Only', 'Today', 'Trial', 'All',
                 'Certified', 'Cheap', 'Claims', 'Clearance', 'Compare', 'Congratulations', 'Finance', 'Financial', 'Free',
                 'Free!', 'Get', 'Gold', 'Great', 'Guarantee', 'Human', 'Legal', 'Life', 'Limited', 'Loan', 'Lottery',
                 'Member', 'Millionaire', 'Millions', 'MLM', 'Name', 'New', 'Nigerian', 'Obligation', 'Offers', 'Offshore',
                 'Opportunity', 'Opt-in', 'Order', 'Password', 'Perfect', 'Phone', 'Please', 'Profit', 'Purchase', 'Quotes',
                 'Rate', 'Rebate', 'Request', 'Sale', 'Sales', 'Save', 'Sex', 'Spam', 'Subscribe', 'Supplies', 'Unsolicited',
                 'Unsubscribe', 'Vacation', 'Viagra', 'VIP', 'Win', 'Winning', 'XXX', '#1', '%', '0%', '100%', '99%', 'Bonus',
                 'Extra', 'FAST', 'Guaranteed', 'Million', 'Thousands', 'Unlimited', 'Wonderful', '$$$', 'Affordable', 'Bank',
                 'Bankruptcy', 'Billion', 'Cash', 'Casino', 'Check', 'Cost', 'Costs', 'Credit', 'Deal', 'Debt', 'Discount',
                 'Dollars', 'Earn', 'Income', 'Insurance', 'Investment', 'Lifetime', 'Loans', 'Money', 'Mortgage', 'Offer',
                 'Price', 'Prices', 'Profits', 'Quote', 'Rates', 'Refinance', 'Addresses', 'Form', 'Freedom', 'Here', 'Hidden',
                 'Home', 'Leave', 'Lose', 'Marketing', 'Never', 'Removal', 'Remove', 'Satisfaction', 'Stop', 'Success', 'Teen',
                 'Warranty', 'FREE', '-', '&', '#', 'address', '*', '��', ',', '.', 'U.S.', '***', '--', '>', '...',
                 '#1', '(un)subscription', '!', '/', 'NOW!', '**', 'à', '_/', '[ILUG]', '+', 'Q.', ';', '$', '<br>',
                 'AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA', '->', '·', 'Q:', '@', 'A:',
                 'FREE!', '<a', '!!!', '=', '===========================================', 'now!',
                 '-------------------------------------', '\\tab', '있는', '===', '(tm)', '$300+', 'YOUR', 'THE', 'click', 'THIS', 'FOR']

sub_word = ['test', 'ordinance', 'acquire', 'visit', 'read', '', '', '', 'nearly', '', 'trial', 'most', 'Premium', 'inexpensive', 'Suit',
            'headroom', 'contrast', 'congrats', 'banking', 'banking', 'complimentary', 'complimentary', 'obtain', 'silver', 'astonishing',
            'human being', 'authorized', 'fate', 'ltm', 'discharge', 'jackpot', 'one of the group', 'rich person', 'a big amount', 'networ
            'fresh', 'nigerians', 'job', 'brings', 'abroad', 'odds', 'decide to join', 'order', 'code', 'ideal', 'cellphone', '', 'earnings',
            'acquire', 'preview', '', 'decrease', 'need', 'vendue', 'vendues', 'saving', 'gender', 'garbage mail', 'please visit',
            'equipments', 'unwanted', 'not visit', 'trip', 'good person', 'vip', 'victory', 'victory', '', '',
            'percentage', 'o percentage', '1oo percentage', '99 percentage',
            'good point', 'another', 'quick', 'certificated', 'a big amount', 'a big amount', 'infinite', 'vey good', '',
            'inexpensive', 'bank', 'bank shut down', 'a big amount', 'coins', 'gaming field', 'make sure', 'consume', 'fees', 'credit',
            'transaction', 'discharge', 'decrease', 'coins', 'have', 'coins', 'making sure', 'investment', 'longtime', 'lending coins', 'co
            'pledge', 'give', 'coin', 'coins', 'benefit', 'reference', '', 'sponsor', 'address', 'form', 'without coin',
            '', 'unseen', 'house', 'let it go', 'put aside', 'promotion', 'do not', 'eradication', 'eradicate', 'content', 'do not', 'victory
            'teenager', 'assurance', 'complimentary', '', 'and', '', 'location', '', '', '', 'America', '', '', '', '', '',
            'not keep watching', '', '', '', 'a', '', '', 'Q', '', 'coins:', '', '', '', '', 'complimentary', '', '', '',
            '', '', '', '', '', '', '', 'your', 'the', 'visit', 'this', 'for']
```

方法一結果

同義字抽換

- word2vec 查找
- 手動抽換
- 不在 trigger word

Δ recall: -1%

符號刪除

!-+*#

從斷詞頻率查找

Δ recall: +1%

連結刪除

spam: 512

ham: 4,310

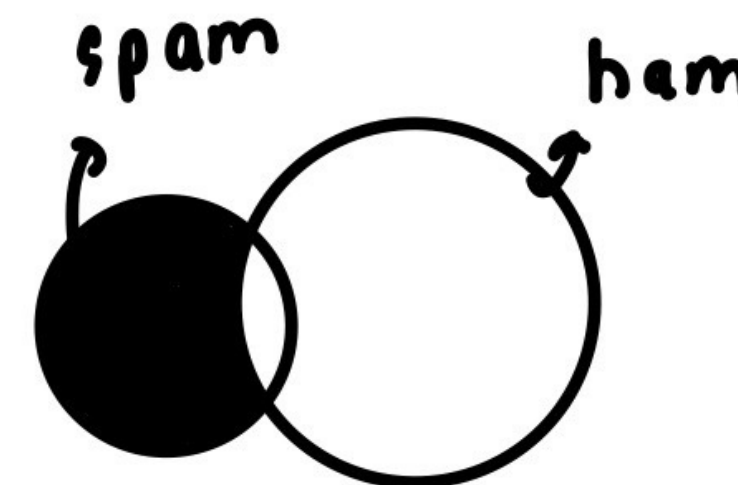**Δ recall: -0%**

推測：ham 裡面也有那些字、符號和連結，相關性同時下降

調整方式 #2

1 分別對 spam 和 ham 做斷詞

2 將 spam 及 ham 的 TF 相減，取得 TF > 0 的字

3 取出前 100 "相對" 常出現在 spam 的字

方法二結果

直接刪除	故意拼錯	word2vec 換字
將差集的字元直接刪除	將差集的字加上特殊字元，使得該字不曾出現在 vocabulary	將差集且高頻的字換上最相近的 word2vec
Δ recall: -2%	**Δ recall: -0%**	**Δ recall: -0%**

調整方式 #3

1　分別對 spam 和 ham 做斷詞

2　找出 ham 差集的字

3　調整 spam 的字
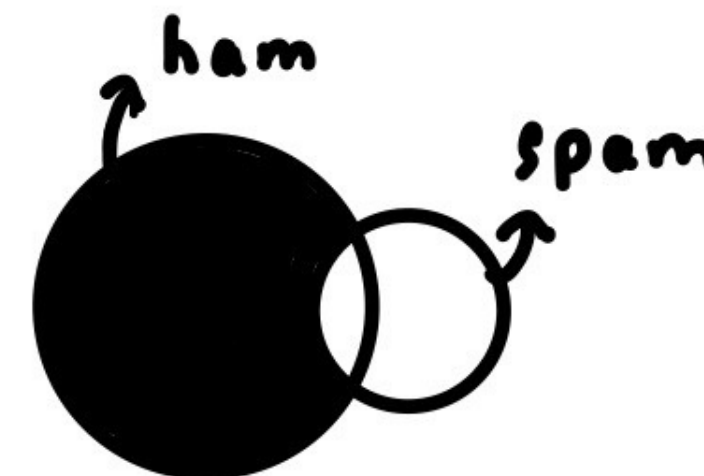
方法三結果

直接篩入-1

挑出差集字元中常出現的 20 字，加入 testing 文本

recall : -50%

直接篩入-2

從常出現的 55 字中挑出 8 個字，結合 stop words 形成兩個句子

recall : -20%

Δ recall 下降率排行榜

研究成果

1 挑出差集字元 20 字加入文本 **-50%**

2 挑 8 個字結合 stop words 形成句子 **-20%**

3 差集字元直接刪除 **-2%**

4 word2vec 抽換同義字 **-1%**

5 刪除符號、連結 **-0%**

建議作法1 - 推薦字

Dear friend,

Financial - bank

I am a Financial Consultant in control of privately owned funds placed for long term investments.

My client intends to invest these funds in projects. I am willing to finance projects at a guaranteed 5% ROI per annum for projects ranging from 2 years term and above but not exceeding 12 years.

Please answer ASAP.

Overall score: **Poor**
Words: **57**
Read time: **a few seconds**

🚨 Urgency (1)

🔞 Shady (4)

🤩 Overpromise (2)

💬 Unnatural (1)

推薦機制：同義字，但不常出現於 spam

建議作法2 - 加句子

ham - spam
高頻字
(前50個)

['wrote:', 'URL:', 'XML', 'ALB>', '[1]', '<RPM-List@freshrpms.net>', 'RPM-List', 'http://lists.freshrpms.net/mailman/listinfo/rpm-list', 'Matthias', '0.99;', 'Bush', 'Hat', 'said.', 'Groups', '[2]', 'rpm', 'Perl', 'unseen', 'DataPower', '0.000', 'install', 'exmh', 'packages', '{', 'Java', '0.01', 'msgs', 'kernel', '[Spambayes]', '@@', 'log', 'that.', 'wrote', 'Sent:', '}', 'OSDN', 'phone?', 'lists/l-k', 'Saou', 'https://www.inphonic.com/r.asp?r=sourceforge1&refcode1=vs3390', 'Wed,', 'Message-----', 'writes:', 'heaven.', '-----Original', '==', 'looks', 'by:ThinkGeek', 'http://thinkgeek.com/sf', 'SA', '-------', 'systems', 'Exmh-users', 'Exmh-users@redhat.com', 'https://listman.redhat.com/mailman/listinfo/exmh-users']

製作句子

**I wrote an unseen message for you,
you said that the hat looks original ==**

客製化選擇

根據不同主題製作多種句子，讓業主依信件內容
挑選合適的句子加入文本

研究限制與貢獻

研究限制

- w2v 精準度太低，且有些字詞並無相似字，需手動排除
- 未能顯著降低 spam recall rate，結果可信度待加強
- 文本量不夠，可能有偏差

研究貢獻

- 排除掉較無用的文本調整方式：刪除特殊符號、刪除連結等
- 提出更好的文本調整方式，讓業主更有機會將訊息傳遞給特定受眾

文字探勘期末報告

謝謝大家

第 18 組

數學三 陳鵬仁、會計三 林子昕、生傳三 黃韻文、資管四 丁啟恩、資管三 陳奕兆