Data Structure--Final Project Proposal

Group 2-林貞妮、林聖典、彭琮鈺、傅俊益
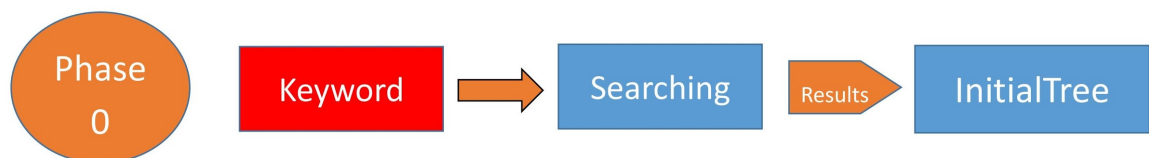
<I>Find what? How to beat google?

Discount restaurant. The more discount information is better.
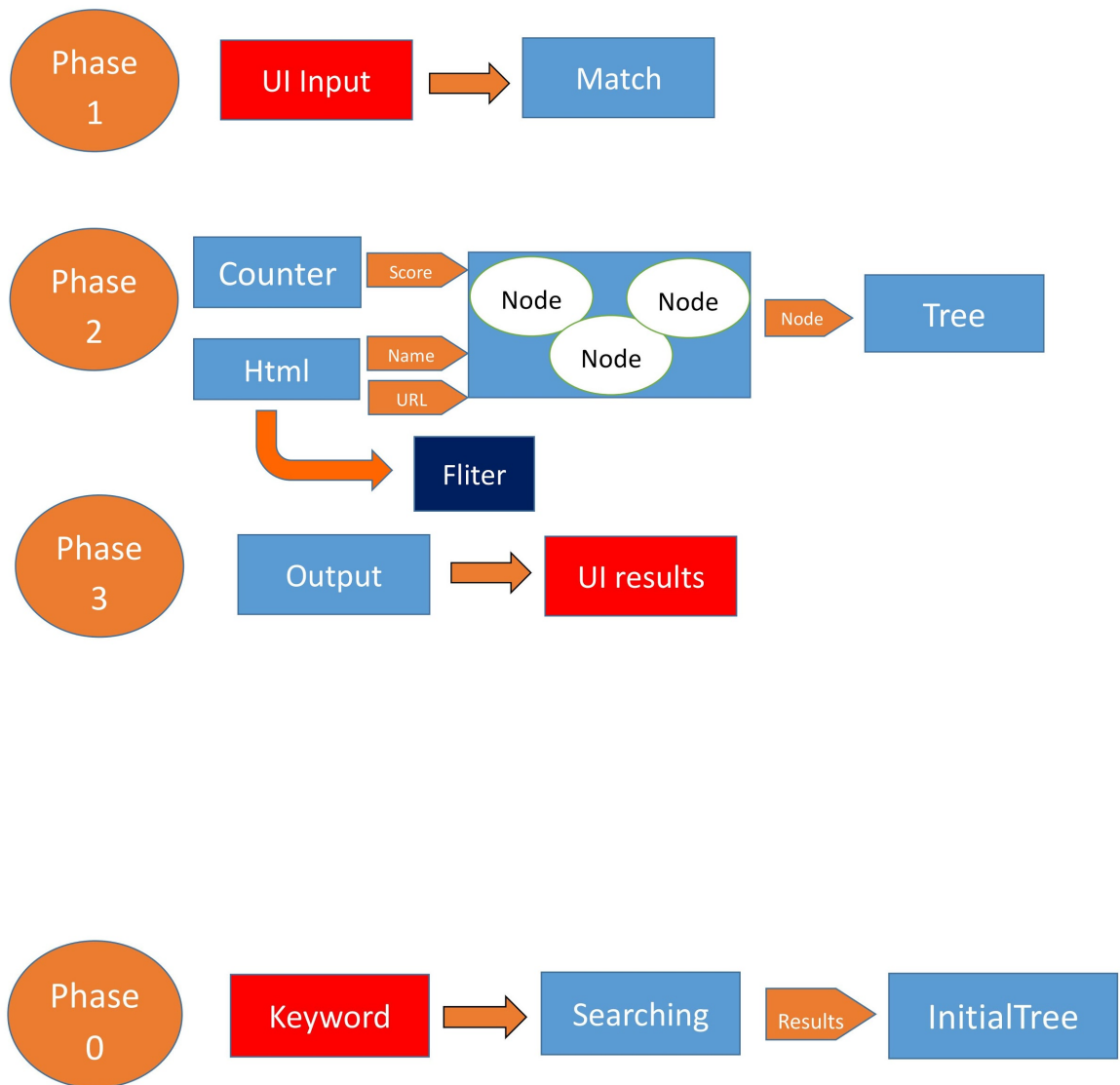
<II>Search tricks(Keyword)

服務、環境、整潔、份量…

- 折扣：優惠、X人同行、好康、生日、起、學生證、折價、餐廳、折

  抵、打折、壽星、學生、特約

<III>System design

Phase 1
UI Input → Match

Phase 2
Counter — Score
Html — Name, URL
Node, Node, Node → Tree
Html → Fliter

Phase 3
Output → UI results

Phase 0
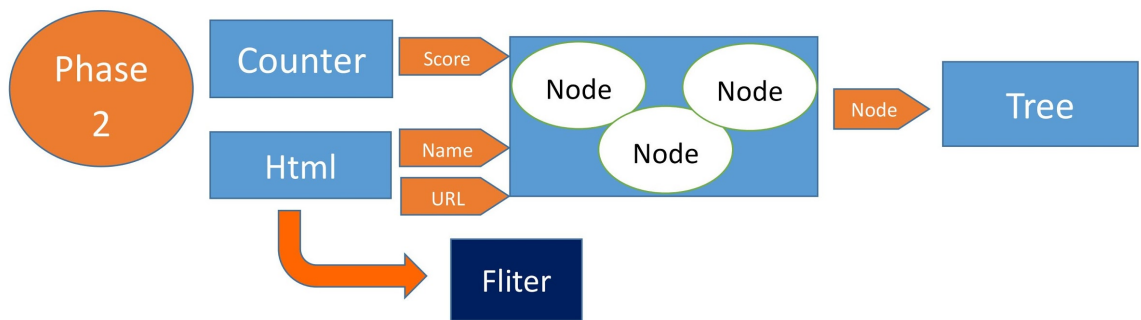Keyword → Searching — Results → InitialTree

Phase 0: First, we set the keyword search engine to do the first search, the search results made into nodes and organized into InitialTree
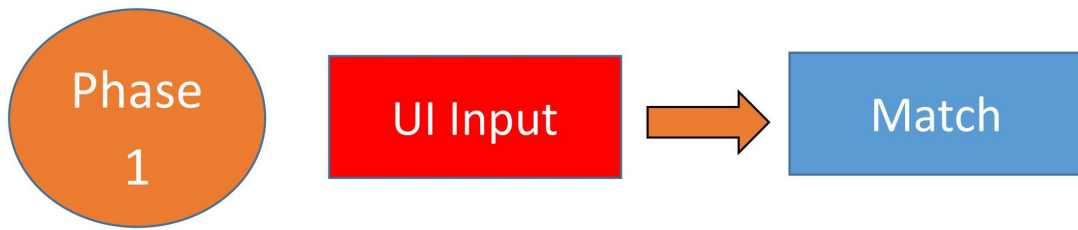
Phase 1: create a user interface (UI), the user input keywords sent to InitialTree to compare whether there is a user input keyword

Phase 2: Tree1 matching search results made into Nodes, Node contains name, count, score, url and other attributes, where the result of the url by the HTMLTag method to catch the original code <a href=.....> Web site return derived. The result of the score is returned by the Keyword Counter operation. Finally, many nodes combined into a Tree, and use tree to help each node ranking.
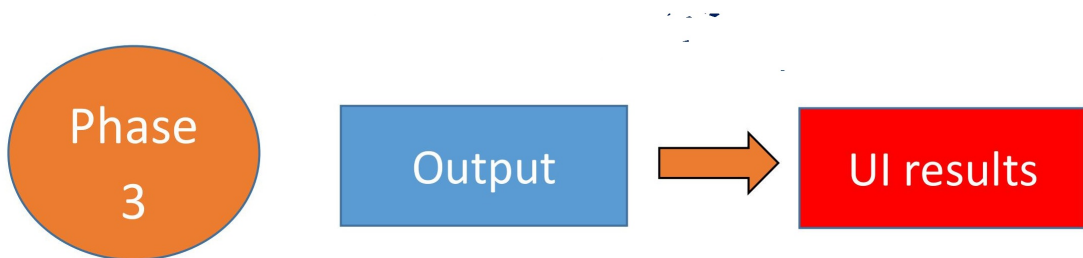
Phase 3: pass the result of rank back to the page, and then output the top 20 scores of the search results

<IV>Method interface

Keyword:

    public Keyword(String name, double weight)

    public String toString()

WordCounter :

    public WordCounter(String urlStr)

    private String fetchContent() throws IOException

    public int countKeyword(String keyword) throws IOException

WebPage:

    public WebPage(String URL, String name)

    public double setScore(ArrayList<Keyword> keywords) throws IOException

WebNode:

    public WebNode(WebPage webpage)

    public double setNodeScore(ArrayList<Keyword> keywords) throws IOException

    public void addChild(WebNode child)

WebTree:

    public WebTree(WebPage rootPage)

    public double setPostOrderScore(ArrayList<Keyword> keywords) throws IOException

    private double setPostOrderScore(WebNode startNode, ArrayList<Keyword> keywords)throws IOException

    public void eularPrintTree()

    private void eularPrintTree(WebNode startNode)

GoogleQuery:

    public GoogleQuery(String searchKeyword)

    public ArrayList<Keyword> disjointKeyword(String connectKeyword)

    public HashMap<String, String> query() throws IOException

    public ArrayList<Rerank> quickSort(ArrayList<Rerank> original)

Rerank:

    public Rerank(String siteTitle, String url,double finalScore

    public String toString()

Main:

    public static void main(String[] args) throws IOException

<V>Schedule(stage)

PERT

Activity	Optimistic Time	Realistic Time	Pessimistic Time	Expected Time
(A)Requirements Collection	1	2	3	3
(B)Anlysis	2	4	6	4
(C)Keyword Counting	1	2	3	2
(D)Page Ranking	1	2	3	2
(E)Site Ranking	2	3	4	3
(F)Refine the rank of Google	2	4	6	4
(G)Semantics Analysis	1	2	3	2
(H)UI Design	4	6	8	6
(I)Testing	2	4	6	4

| (J)Publish Your Work Online | 2 | 2 | 2 | 2 |
| (K)Export Your Work to App | 3 | 6 | 9 | 6 |

Activity	TE	TL	Slack Time	On Critical Path
A	3	3	0	v
B	7	7	0	v
C	9	9	0	v
D	11	11	0	v
E	14	14	0	v
F	17	17	0	v
G	19	19	0	v
H	10	19	9	
I	21	21	0	v
J	22	22	0	v

K	28	28	0	v

TE=3 TE=7 TE=9 TE=11 TE=14 TE=17 TE=19 TE=21 TE=22 TE=28
TL=3 TL=7 TL=9 TL=11 TL=14 TL=17 TL=19 TL=21 TL=22 TL=28

A → B → C → D → E → F → G → I → J → K

TE=10
TL=19

H
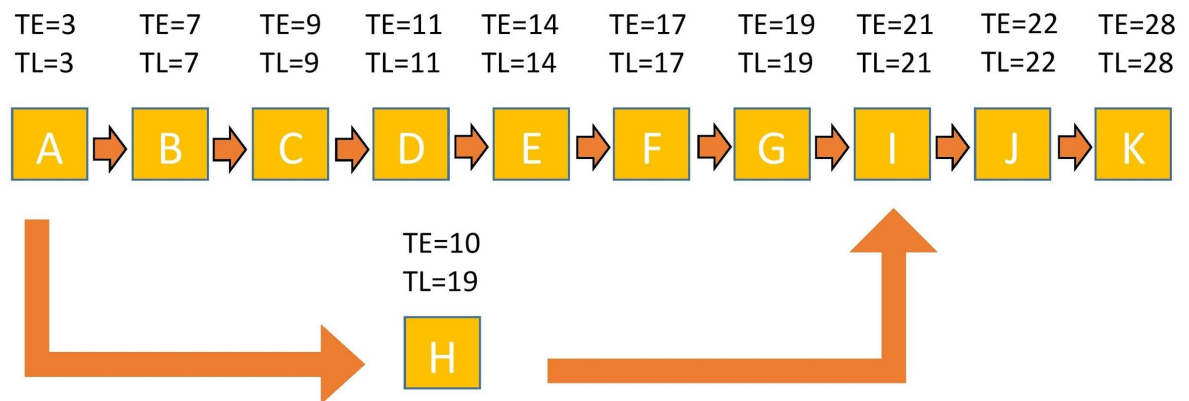
\<VI\>Challenge

1. How to webrize our search engine?

2. How to make our results more precise?

3. How to make program run faster?

Final presentation

Group 2

Member: Henry Jason Jimmy Jennie

105306046,105306016,105306087,
105306002

SeeFood

To find what you can not only eat delicious but
also can be cheap

Contents

1. What does seefood do?(our idea)

2. Procedures in our project

    A. Combine our homework
    B. Set our default keywords
    C. Call google search
    D. Catch subsites
    E. Create trees
    F. Calculate scores of supersite and all of its subsites
    G. Re-rank our supersites
    H. Print out results
    I. Webrize our search engine

3. What difficulties do we face?

4. Demo our search engine—seefood!

1 What does seefood do?(our idea)

- Seefood can find restaurants which have discounts for students.
- Surely, our goal is beating google with our better results.

2 Procedures in our project

A. Combine our homework
B. Set our default keywords
C. Call google search
D. Catch subsites
E. Create trees
F. Calculate scores of supersite and all of its subsites
G. Re-rank our supersites
H. Print out results
I. Webrize our search engine

A. Combine our homework

- We create a new class to help print out our results.
  --Rerank
- We reuse those classes in our homework.
  --Keyword, WordCounter, WebPage, WebNode, WebTree,
    GoogleQuery

B. Set our default keywords

- We define some default keywords to make our results more precisely and meet our needs.

(以搜尋結果的網頁，去比對以下關鍵字是否有出現在

那些網頁內，若有，則乘上權重10)

E.g. 折價、餐廳、優惠、折抵、折扣、打折、生日、壽

星

C. Call google search

- Call google to search keywords(user input)
- Special: we use google built-in searching method to make our results more precisely
- We add "intitle:" in our call google search

- 結果雖然不到 100 個，但那是因為我們抓的是

  intitle，而且不會印出分數為 0 的網頁，所以結果

  較少，但更精準。

D. Catch subsites

- We use a for loop to catch subsites in the for loop which we catch supersites.

```
for (Element u : manysubcite){

    String subciteUrl = u.attr("href");

    Document subdoc = Jsoup.parse(subciteUrl);
    Elements subtitle_HTML = subdoc.getElementsByTag("title");
    String subtitle = subtitle_HTML.text();

    WebPage childPage;
    if(subciteUrl.startsWith("/") || subciteUrl.startsWith("#")){
        childPage = new WebPage(citeUrl+subciteUrl, subtitle);
    }
    else{
      childPage = new WebPage("Invalid url", "no title");
    }

    tree.root.addChild(new WebNode(childPage));

}
```

-

E. Create trees(一個母網頁一個)

- We use a for loop to catch subsites in the for loop which we catch supersites.
- Detail codes at right side.

F. Calculate scores of supersite and all of its subsites

- Before we calculate scores, we create a new method, disjointKeyword (String connectKeyword) to separate keywords if users input many keywords in a long string.

```
public ArrayList<Keyword> disjointKeyword(String connectKeyword){
    ArrayList<Keyword> keywords = new ArrayList<Keyword>();
    String[] input = connectKeyword.split(" ");

    for (int i = 0; i < input.length; i++){

            Keyword k = new Keyword(input[i], 2);
            keywords.add(k);
    }
    return keywords;
}
```

- Then, we call setPostOrderScore() method in webtree class to calculate the scores of our tree, that is, scores of supersite and all of its subsites.

```
double treeScore = tree.setPostOrderScore(disjointKeyword(searchKeyword));
Rerank superRank = new Rerank(title, citeUrl ,treeScore);
```

G. Re-rank our supersites

- We use QuickSort to re-rank our results(supersites).

```java
public ArrayList<Rerank> quickSort(ArrayList<Rerank> original){
    if(original.size() <= 1 ){
        return original;
    }else if(original.size() == 2){
        if(original.get(0).finalScore < original.get(1).finalScore){
            Rerank temp = original.get(0);
            original.add(temp);
        }
        return original;
    }else{
        ArrayList<Rerank> result = new ArrayList<Rerank>();
        int pivotIndex = original.size() / 2;
        Rerank pivotKeyword = original.get(pivotIndex);

        ArrayList<Rerank> lessList = new ArrayList<Rerank>();
        ArrayList<Rerank> equalList = new ArrayList<Rerank>();
        ArrayList<Rerank> greaterList = new ArrayList<Rerank>();

        for(int i = 0;i < original.size(); i++){
            Rerank k = original.get(i);
            if(k.finalScore > pivotKeyword.finalScore){
                greaterList.add(k);
            }else if(k.finalScore < pivotKeyword.finalScore){
                lessList.add(k);
            }else{
                equalList.add(k);
            }
        }

        result.addAll(quickSort(greaterList));
        result.addAll(equalList);
        result.addAll(quickSort(lessList));

        return result;
    }
}
```

H. Print out results

- We create an object, called Rerank, to help us store three attributes of tree.
- siteTitle, url, finalScore
- Use an arraylist to store Reranks, and print it out.

```java
1
2  public class Rerank {
3
4      protected String siteTitle;
5      protected String url;
6      protected double finalScore;
7
8      public Rerank(String siteTitle, String url,double finalScore){
9          this.siteTitle = siteTitle;
10         this.url = url;
11         this.finalScore = finalScore;
12     }
13     @Override
14     public String toString(){
15         return "[" + siteTitle + "," + url + "," + finalScore +"]";
16     }
17
18 }
```

火鍋
(【精選】20間台北最好吃的火鍋:涮涮鍋、麻辣鍋、酸菜 ... - 下午茶女王Venus,376.0,https://venuslin.tw/2014-12-17-2878/
)
(台中火鍋台中鍋物推薦。2018.1月更新。台中美食餐廳推薦懶人包總整理...,192.0,https://sunnylife.tw/2017-04-02-1107/
)
(台北火鍋餐廳餐券-GOMAJI最大吃喝玩樂券平台,0.0,www.gomaji.com/Taipei_p182349.html
)
(火鍋- ?基百科,自由的百科全?,0.0,https://zh.wikipedia.org/zh-tw/火鍋
)
([懶人包]台北火鍋總整理(2017.07.18更新) 台北火鍋推薦火鍋、麻辣鍋...,0.0,piliapp-mapping.softc.tw/.../335089922-%5B懶人包%5D台北火鍋總整理-%282017....
)
(【台北火鍋】台北新北必吃火鍋私藏名單32家餐廳捷運沿線整理推薦！涮涮...,428.0,https://zineblog.com.tw/blog/post/10231
)
(【台北火鍋懶人包】台北新北火鍋推薦/涮涮鍋、麻辣、海鮮、溫體牛、大盤...,0.0,z78625.pixnet.net/.../204323167-【台北火鍋懶人包】台北%2B新北火鍋推薦-日式
)
(台北火鍋總整理～推薦一覽表(2017.12.5更新) - 愛吃鬼芸芸,532.0,https://aniseblog.tw/4318
)
((懶人包)台北火鍋、鍋物總整理,41家北部火鍋吃到飽、單品、麻辣鍋...,0.0,may1215may.pixnet.net/.../401019653-%28懶人包%29台北火鍋、鍋物總整理,41...
)
[【精選】20間台北最好吃的火鍋:涮涮鍋、麻辣鍋、酸菜 ... - 下午茶女王Venus,https://venuslin.tw/2014-12-17-2878/,376.0]
[台中火鍋台中鍋物推薦。2018.1月更新。台中美食餐廳推薦懶人包總整理...,https://sunnylife.tw/2017-04-02-1107/,192.0]
[台北火鍋餐廳餐券-GOMAJI最大吃喝玩樂券平台,www.gomaji.com/Taipei_p182349.html,0.0]
[火鍋- ?基百科,自由的百科全?,https://zh.wikipedia.org/zh-tw/火鍋,0.0]
[[懶人包]台北火鍋總整理(2017.07.18更新) 台北火鍋推薦火鍋、麻辣鍋...,piliapp-mapping.softc.tw/.../335089922-%5B懶人包%5D台北火鍋總整理-%282017....,0.0]
[【台北火鍋】台北新北必吃火鍋私藏名單32家餐廳捷運沿線整理推薦！涮涮...,https://zineblog.com.tw/blog/post/10231,428.0]
[【台北火鍋懶人包】台北新北火鍋推薦/涮涮鍋、麻辣、海鮮、溫體牛、大盤...,z78625.pixnet.net/.../204323167-【台北火鍋懶人包】台北%2B新北火鍋推薦-日式,0.0]
[台北火鍋總整理～推薦一覽表(2017.12.5更新) - 愛吃鬼芸芸,https://aniseblog.tw/4318,532.0]
[(懶人包)台北火鍋、鍋物總整理,41家北部火鍋吃到飽、單品、麻辣鍋...,may1215may.pixnet.net/.../401019653-%28懶人包%29台北火鍋、鍋物總整理,41...,0.0]

I. Webrize our search engine

- Use Apache-Tomcat to webrize our search engine.

3 What difficulties do we face?

- MalformedURLException

    difficulty: subsites' url has to start with "http://", while the value in attribute "href" is not this case.

    solution: we find that the value in attribute "href" should be attached after supersite's url.

```
if(subciteUrl.startsWith("/") || subciteUrl.startsWith("#")){
    childPage = new WebPage(citeUrl+subciteUrl, subtitle);
}
else{
    childPage = new WebPage("Invalid url", "no title");
}

tree.root.addChild(new WebNode(childPage));
```

因為之前噴錯時，都是URL url = new URL(this.urlStr);這行出錯，推測應該是子網頁的網址出現問題，讓我們的程式跑不動，

我們認為可能是網址後面是中文，讀進去後會是一堆亂碼，所以我們用**catch**內的**for loop**做，從子網頁的網址最後面往前推，，去抓**"/"**，找到較正確的子網頁網址去跑

•NullPointerException

difficulty: the url in attribute "href" isn't complete, so java can't load it.

solution: we use a try-catch to solve this problem.

```java
try
{
    URL url = new URL(this.urlStr);

    HttpURLConnection conn = (HttpURLConnection)url.openConnection();//通道
    conn.connect();

    InputStream in = conn.getInputStream();
    BufferedReader br = new BufferedReader(new InputStreamReader(in,"UTF-8"));

    String line = null;

    while((line = br.readLine()) != null)
    {
        retVal = retVal + line + "\n";
    }
}//取得網頁內容
catch (Exception e)
{
    for (int i = urlStr.length() - 1; i >= 0; i--)
    {
        char u = '/';
        if (urlStr.charAt(i) == u)
        {
            urlStr = urlStr.substring(0, i);
            break;
        }
    }
}
```

發現子網頁網址必須以http:// 開頭，所以我們用citeUrl+subciteUrl，也就是把母網頁的網址加上子網頁的部分，合成最後完整的網址

4.Demo our search engine

火鍋 吃到飽

[肉多多火鍋-火鍋第一品牌- 首頁│ Facebook, https://zh-tw.facebook.com/rododohotpot/, 2550.0]
[【台北火鍋】台北新北必吃火鍋私藏名單32家餐廳捷運沿線整理推薦！涮涮 ..., https://zineblog.com.tw/blog/post/10231, 1630.0]
[【精選】20間台北最好吃的火鍋：涮涮鍋、麻辣鍋、酸菜 ... - 下午茶女王Venus, https://venuslin.tw/2014-12-17-2878/, 1440.0]
[台北火鍋總整理～推薦一覽表(2017.12.5更新）- 愛吃鬼芸芸, https://aniseblog.tw/4318, 1430.0]
[台中火鍋台中鍋物推薦。2018.1月更新.台中美食餐廳推薦懶人包總整理 ..., https://sunnylife.tw/2017-04-02-1107/, 1160.0]
[冬天必吃！不退燒的老字號石頭火鍋│ 食尚玩家 - TVBS新聞, https://supertaste.tvbs.com.tw/topic/article/306729, 420.0]
[懶人救星！用冷水煮的自煮火鍋開箱試吃啦！！【謝秉鈞Attila的小開箱 ..., https://www.youtube.com/watch?v=ypYrM8BtxxQ, 200.0]
[「火鍋」相關食譜共416 道- iCook 愛料理, https://icook.tw/recipes/search?q=火鍋, 20.0]
[這一鍋、火鍋、鴛鴦鍋、麻辣鍋│ 這一鍋皇室祕藏鍋物, https://www.toponepot.com/, 10.0]
[【台北吃到飽懶人包】65間台北燒烤、火鍋、日本料理、泰式、韓式、自助餐 ..., https://www.tony60533.com/taipeiallyoucaneat/, 6420.0]
[吃到飽排行榜│ 美食│ 百大口碑-DailyView 網路溫度計, https://dailyview.tw/Top100/topic/61, 2555.0]
[台中吃到飽│ 漢來海港百匯~海鮮.異國料理吃到飽總整理.台中廣三sogo店, https://niniandblue.com/blog/post/196263444, 1970.0]
[全台吃到飽大盤點。，17Life，團購、宅配24小時出貨、優惠券、即買即用, https://www.17life.com/event/brandevent.aspx?u=2017_cd8656ba&cpa=17..., 1725.0]
[【超實用懶人包42間】台北吃到飽餐廳推薦火鍋燒烤、buffet自助餐、日本 ..., https://tenjo.tw/allyoucaneat-taipei/, 1450.0]
[台北15家CP值超高吃到飽Buffet大集合！│ GQ瀟灑男人網 - GQ Taiwan, https://www.gq.com.tw/life/food/content-32125.html, 1200.0]
[全台吃到飽餐廳推薦，訂位享優惠│ 24小時線上餐廳訂位│ EZTABLE 簡單 ..., https://tw.eztable.com/channel/209, 900.0]
[台中小火鍋吃到飽。大樂鍋。個人獨享鴛鴦鍋、鍋喜燒、溫補鍋不用300元 ..., https://sunnylife.tw/dalepotgood/, 900.0]
[4G 上網吃到飽│ 網路門市│ 台灣大哥大, https://www.taiwanmobile.com/content/event/estore-4g-499-599/index.html, 740.0]
[吃到飽餐廳海鮮不斷消失跑進阿嬤「神奇百寶袋」│ 網搜熱話題│ 社會│ 聯合 ..., https://udn.com/news/story/8864/2874213, 590.0]
[網路吃到飽戰不停！2017年度電信業大事件回顧！│ 每日調查 ..., https://dailyview.tw/Daily/2018/01/08, 430.0]
[亞太電信249元21Mbps上網吃到飽又回來了！- 小丰子3C俱樂部, https://tel3c.tw/blog/post/19716, 360.0]
[日本WiFi SoftBank 501HW 4G LTE 上網吃到飽 - iVideo, https://www.ivideo.com.tw/wifi/49610, 330.0]
[行家嚴選美食優惠券-吃到飽-雄獅旅遊, https://www.liontravel.com/promotion/taiwan/gourmet/buffet.html, 290.0]
[遠傳推快閃498元吃到飽衝業績挑戰電信三雄最低│ 鉅亨網- 股股新聞, https://news.cnyes.com/news/id/3994969, 280.0]
[看準小資族！星級飯店吃到飽「300有找」- YouTube, https://www.youtube.com/watch?v=GAf5WH1VD94, 260.0]
[千人吃到飽399元！火烤兩吃泰國蝦秒殺│ 食尚玩家, https://supertaste.tvbs.com.tw/topic/article/306398, 240.0]
[吃到飽(@allcaneat) ? Instagram photos and videos, https://www.instagram.com/allcaneat/?hl=zh-tw, 70.0]
[上網吃到飽買30天送30天易付卡首購上網計日型，買幾天就送幾天 - 遠傳, https://promotion.fetnet.net/pmt/if/201704_3/, 60.0]