Google ML Winter Camp
谷歌机器学习应用冬令营

*Yujie Lu*
*Muge Chen*
*Yanqiong Chen*

# Webpage Wizard

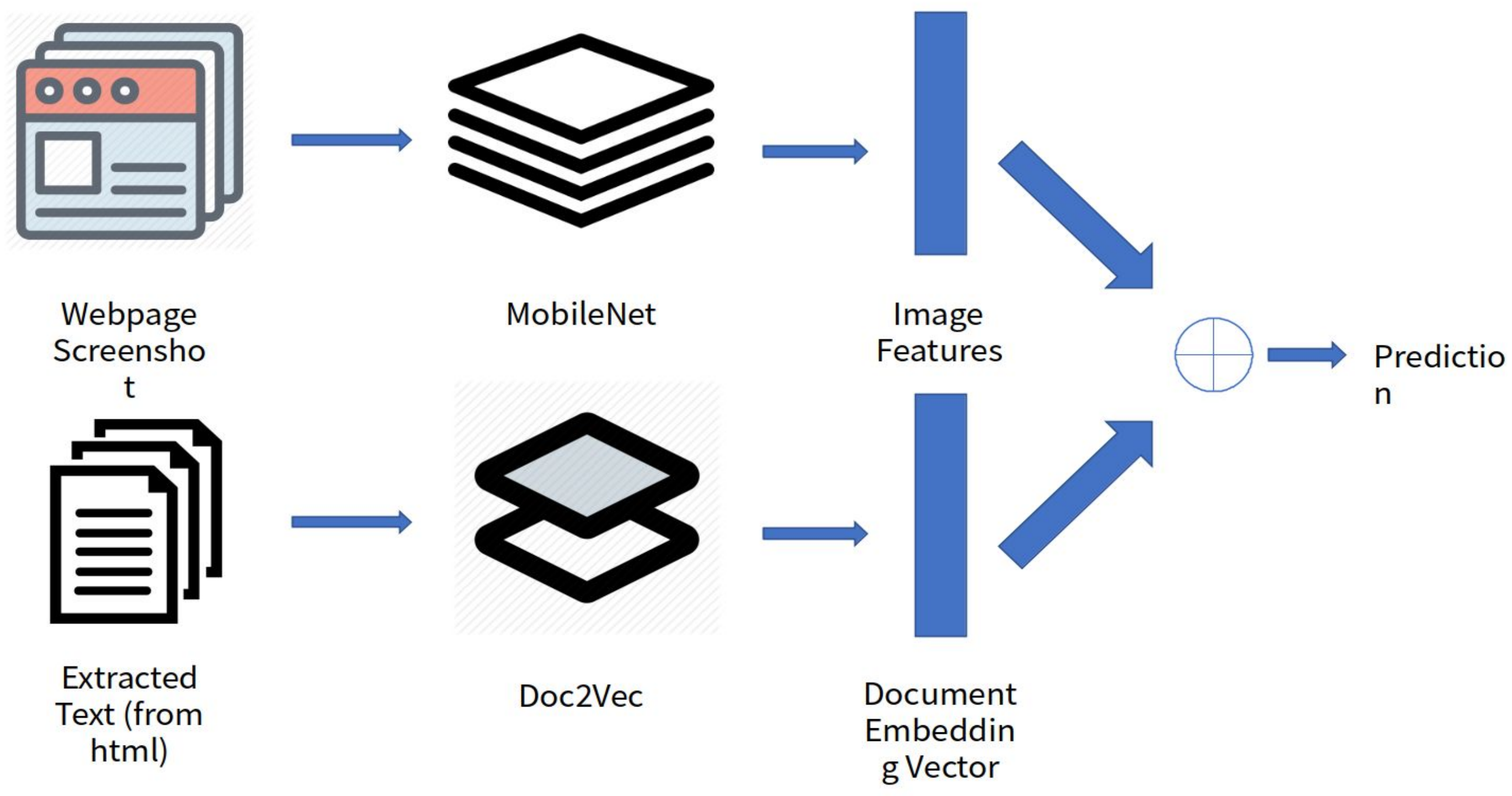
## - A Useful Model for Webpage Classification

- Introduction

In the world of machine learning, most of the application domain suffers from not having sufficient labeled data whereas unlabeled data is available cheaply.

In our project, we try to do webpage classification on a dataset that **only a small amount of data is labeled.**

- Pipeline



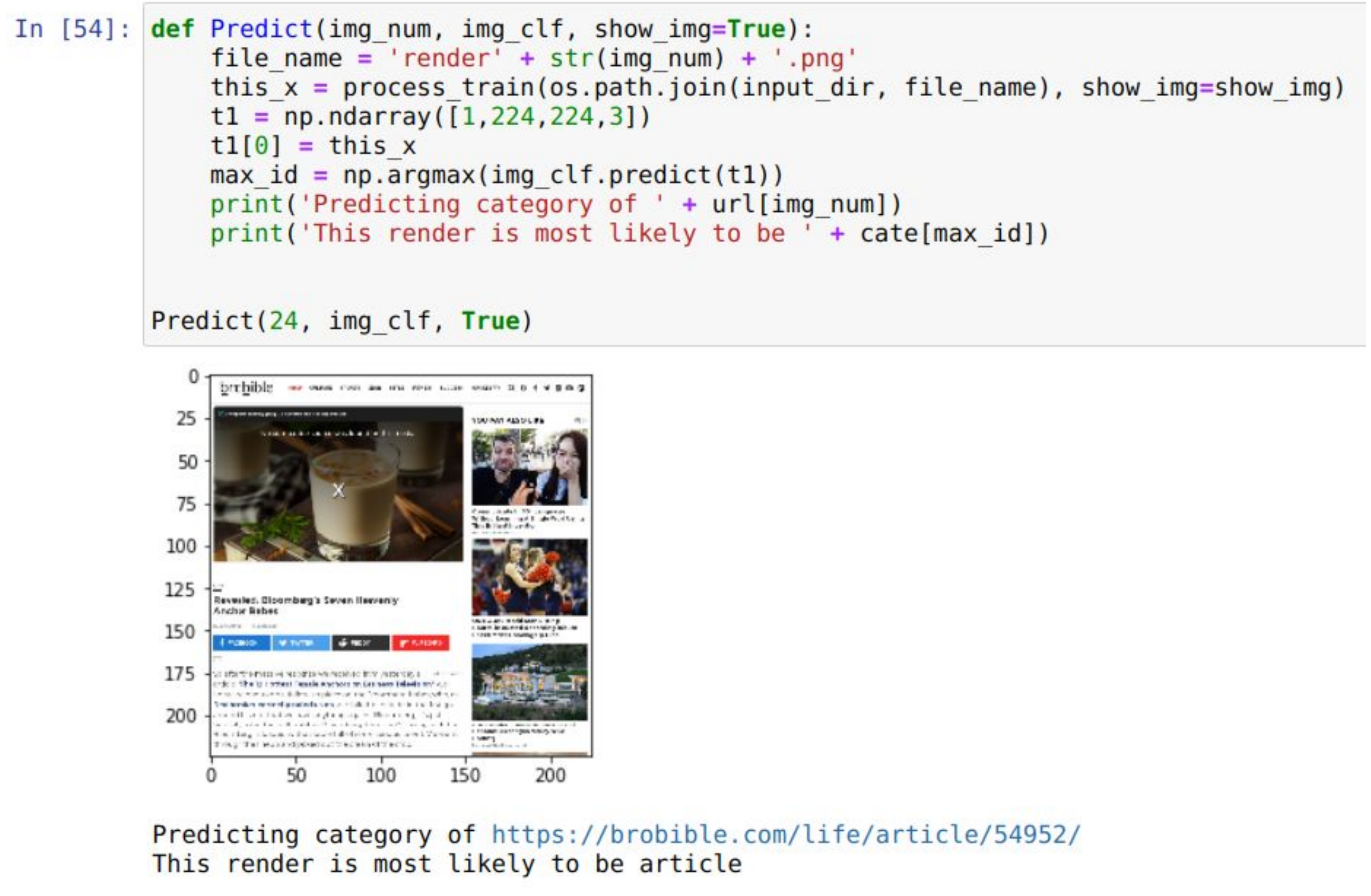
- Implementation Detail

**See github link**

https://github.com/foreseeable/Aelous

- Applications

1 Our Demo



```
In [54]: def Predict(img_num, img_clf, show_img=True):
             file_name = 'render' + str(img_num) + '.png'
             this_x = process_train(os.path.join(input_dir, file_name), show_img=show_img)
             t1 = np.ndarray([1,224,224,3])
             t1[0] = this_x
             max_id = np.argmax(img_clf.predict(t1))
             print('Predicting category of ' + url[img_num])
             print('This render is most likely to be ' + cate[max_id])


         Predict(24, img_clf, True)
```

```
Predicting category of https://brobible.com/life/article/54952/
This render is most likely to be article
```

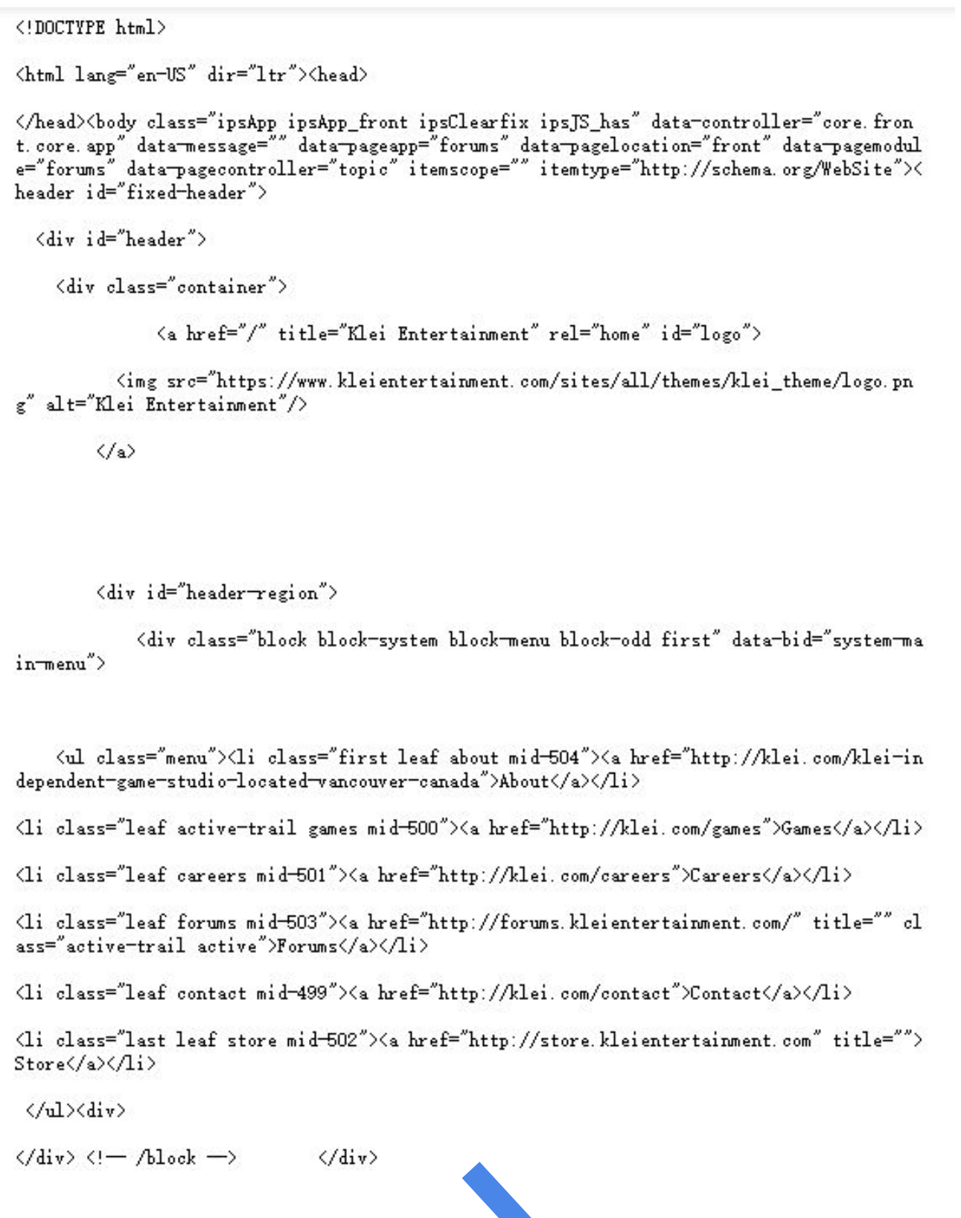
- Dataset:

10K data: url + html + screenshot
2590 with label **is_entity**
800 with label **is_entity** & **category**
**10 categories:**

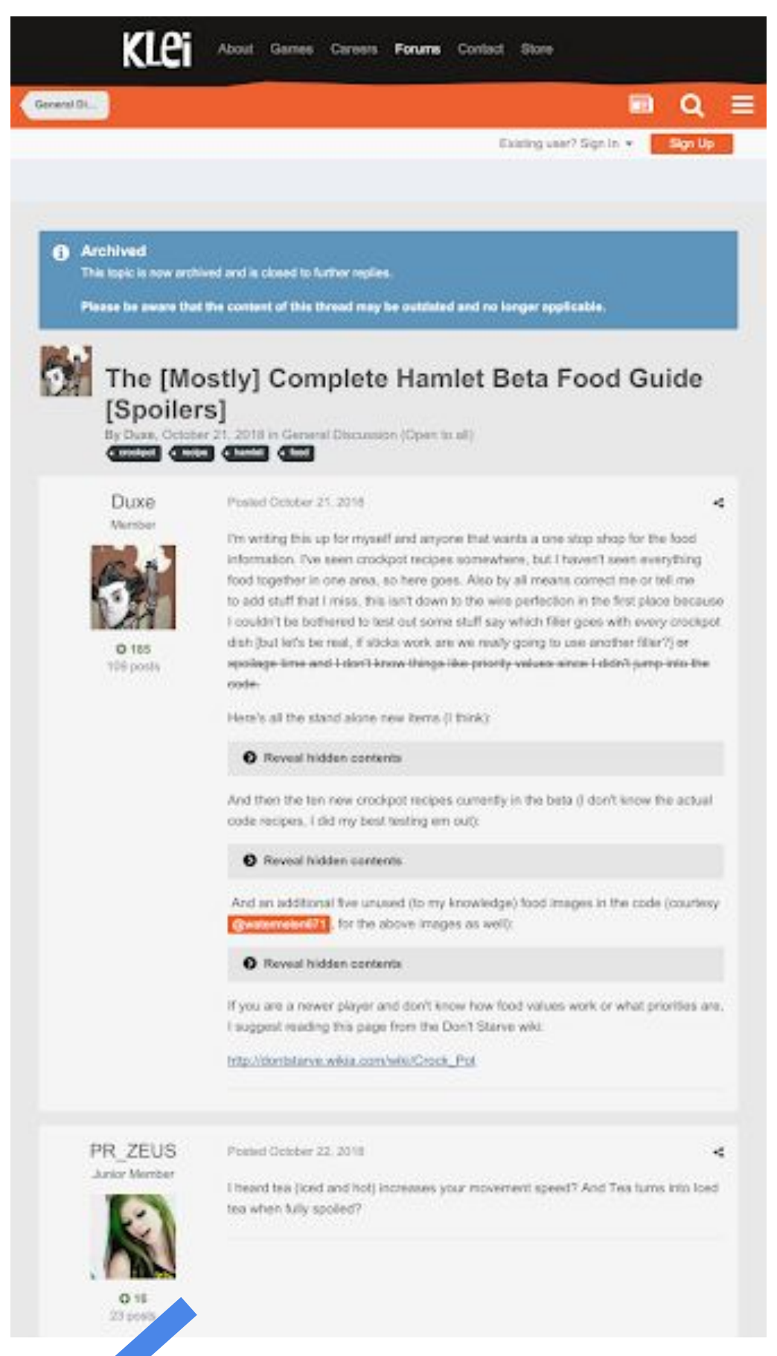
- media_introduction
- others
- location
- social_media_profile
- encyclopedia
- qa_forum
- shopping_item
- list
- media_player
- article

example:

https://forums.kleientertainment.com/forums/topic/97192-the-mostly-complete-hamlet-beta-food-guide-spoilers/



```
<!DOCTYPE html>

<html lang="en-US" dir="ltr"><head>

</head><body class="ipsApp ipsApp_front ipsClearfix ipsJS_has" data-controller="core.fron
t.core.app" data-message="" data-pageapp="forums" data-pagelocation="front" data-pagemodul
e="forums" data-pagecontroller="topic" itemscope="" itemtype="http://schema.org/WebSite"><
header id="fixed-header">

  <div id="header">

    <div class="container">

          <a href="/" title="Klei Entertainment" rel="home" id="logo">

        <img src="https://www.kleientertainment.com/sites/all/themes/klei_theme/logo.pn
g" alt="Klei Entertainment"/>

        </a>




        <div id="header-region">

          <div class="block block-system block-menu block-odd first" data-bid="system-ma
in-menu">



    <ul class="menu"><li class="first leaf about mid-504"><a href="http://klei.com/klei-in
dependent-game-studio-located-vancouver-canada">About</a></li>

<li class="leaf active-trail games mid-500"><a href="http://klei.com/games">Games</a></li>

<li class="leaf careers mid-501"><a href="http://klei.com/careers">Careers</a></li>

<li class="leaf forums mid-503"><a href="http://forums.kleientertainment.com/" title="" cl
ass="active-trail active">Forums</a></li>

<li class="leaf contact mid-499"><a href="http://klei.com/contact">Contact</a></li>

<li class="last leaf store mid-502"><a href="http://store.kleientertainment.com" title="">
Store</a></li>

 </ul><div>

</div> <!-- /block -->        </div>
```



qa_forum

2 Other Promising Applications

- Assist building an extension for Chrome which can beautify the UI with different strategies according to the category of the webpage. Similar to switch omega.
- Help the browser collecting information about which category the user visits most frequently to decide what ADs to present.

投票区域