# 任务

一个由10000条Instagram帖子组成的csv表格。通过数据分析和爬虫，进行数据清理，并载入其他参数。

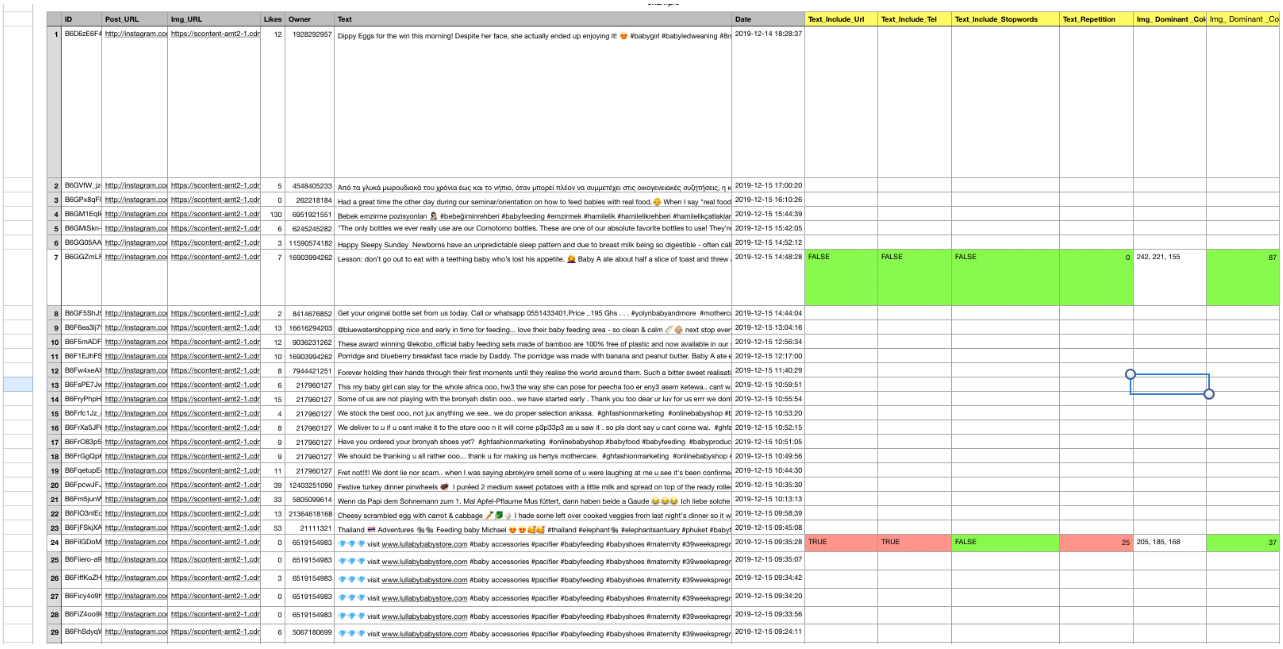
# 具体步骤

## 第一阶段：数据清理

根据以下参数，删除条目，并将剩下的帖子生成一个新的csv。

| 参数 | 定义 | If…，则删除此条目 | 代码参考 |
|---|---|---|---|
| Text_Include_Url | “Text”中是否含有网址 | TRUE | |
| Text_Include_Tel | “Text”中是否含有电话 | TRUE | |
| Text_Include_Stopwords | “Text”中是含有“stopwords.txt”中的词条 | TRUE | |
| Text_Repetition | 同一个“Text”是否重复>=2次 | TRUE | |
| Img_ Dominant _Color_RGB | “Img_URL”链接到的图片，其主导颜色色值(r,g,b) | | https://adamspannbauer.github.io/2018/03/02/app-icon-dominant-colors/ |
| Img_ Dominant _Color_Difference | r g b 中最大值与最小值的差值=a | a>150 | |

见下图：出现红色代表需要删除该条目



| | ID | Post_URL | Img_URL | Likes | Owner | Text | Date | Text_Include_Url | Text_Include_Tel | Text_Include_Stopwords | Text_Repetition | Img_ Dominant _Col | Img_ Dominant _Co |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | B6D6zE6F4 | http://instagram.coı | https://scontent-amt2-1.cdr | 12 | 1928292957 | Dippy Eggs for the win this morning! Despite her face, she actually ended up enjoying it! 😍 #babygirl #babyledweaning #8n | 2019-12-14 18:28:37 | | | | | | |
| 2 | B6GVfW_jz | http://instagram.coı | https://scontent-amt2-1.cdr | 5 | 4548405233 | Από τα γλυκά μωρουδιακά του χρόνια έως και το νήπιο, όταν μπορεί πλέον να συμμετέχει στις οικογενειακές συζητήσεις, η κ | 2019-12-15 17:00:20 | | | | | | |
| 3 | B6GPx8qFl | http://instagram.coı | https://scontent-amt2-1.cdr | 0 | 262218184 | Had a great time the other day during our seminar/orientation on how to feed babies with real food.👶 When I say "real food | 2019-12-15 16:10:26 | | | | | | |
| 4 | B6GM1Eqlk | http://instagram.coı | https://scontent-amt2-1.cdr | 130 | 6951921551 | Bebek emzirme pozisyonları 🤱 #bebeğiminrehberi #babyfeeding #emzirmek #hamilelik #hamilelikrehberi #hamilelikçatlaklar | 2019-12-15 15:44:39 | | | | | | |
| 5 | B6GMiSkn- | http://instagram.coı | https://scontent-amt2-1.cdr | 6 | 6245245282 | "The only bottles we ever really use are our Comotomo bottles. These are one of our absolute favorite bottles to use! They're | 2019-12-15 15:42:05 | | | | | | |
| 6 | B6GG05AA | http://instagram.coı | https://scontent-amt2-1.cdr | 3 | 11590574182 | Happy Sleepy Sunday  Newborns have an unpredictable sleep pattern and due to breast milk being so digestible - often call | 2019-12-15 14:52:12 | | | | | | |
| 7 | B6GGZmLF | http://instagram.coı | https://scontent-amt2-1.cdr | 7 | 16903994262 | Lesson: don't go out to eat with a teething baby who's lost his appetite. 🤦‍♀️ Baby A ate about half a slice of toast and threw a | 2019-12-15 14:48:28 | FALSE | FALSE | FALSE | 0 | 242, 221, 155 | 87 |
| 8 | B6GF5ShJ | http://instagram.coı | https://scontent-amt2-1.cdr | 2 | 8414676852 | Get your original bottle set from us today. Call or whatsapp 0551433401.Price ..195 Ghs . . . #yolynbabyandmore  #motherc | 2019-12-15 14:44:04 | | | | | | |
| 9 | B6F6ea3lj7 | http://instagram.coı | https://scontent-amt2-1.cdr | 13 | 16616294203 | @bluewatershopping nice and early in time for feeding... love their baby feeding area - so clean & calm 🍼👶🏽 next stop ever | 2019-12-15 13:04:16 | | | | | | |
| 10 | B6F5mADF | http://instagram.coı | https://scontent-amt2-1.cdr | 12 | 9036231262 | These award winning @ekobo_official baby feeding sets made of bamboo are 100% free of plastic and now available in our | 2019-12-15 12:56:34 | | | | | | |
| 11 | B6F1EJhFS | http://instagram.coı | https://scontent-amt2-1.cdr | 10 | 16903994262 | Porridge and blueberry breakfast face made by Daddy. The porridge was made with banana and peanut butter. Baby A ate e | 2019-12-15 12:17:00 | | | | | | |
| 12 | B6Fw4xeA) | http://instagram.coı | https://scontent-amt2-1.cdr | 8 | 7944421251 | Forever holding their hands through their first moments until they realise the world around them. Such a bitter sweet realisati | 2019-12-15 11:40:29 | | | | | | |
| 13 | B6FsPE7Jw | http://instagram.coı | https://scontent-amt2-1.cdr | 6 | 217960127 | This my baby girl can slay for the whole africa ooo, hw3 the way she can pose for peecha too er eny3 asem ketewa.. cant w | 2019-12-15 10:59:51 | | | | | | |
| 14 | B6FryPhpH | http://instagram.coı | https://scontent-amt2-1.cdr | 15 | 217960127 | Some of us are not playing with the bronyah distin ooo.. we have started early . Thank you too dear ur luv for us errr we dont | 2019-12-15 10:55:54 | | | | | | |
| 15 | B6Frfc1Jz_ | http://instagram.coı | https://scontent-amt2-1.cdr | 4 | 217960127 | We stock the best ooo, not jux anything we see.. we do proper selection ankasa.  #ghfashionmarketing  #onlinebabyshop #b | 2019-12-15 10:53:20 | | | | | | |
| 16 | B6FrXa5JF | http://instagram.coı | https://scontent-amt2-1.cdr | 8 | 217960127 | We deliver to u if u cant make it to the store ooo n it will come p3p33p3 as u saw it . so pls dont say u cant come wai.  #ghfa | 2019-12-15 10:52:15 | | | | | | |
| 17 | B6FrO83p5 | http://instagram.coı | https://scontent-amt2-1.cdr | 9 | 217960127 | Have you ordered your bronyah shoes yet?  #ghfashionmarketing  #onlinebabyshop #babyfood #babyfeeding  #babyproduc | 2019-12-15 10:51:05 | | | | | | |
| 18 | B6FrGgQpk | http://instagram.coı | https://scontent-amt2-1.cdr | 9 | 217960127 | We should be thanking u all rather ooo... thank u for making us hertys mothercare.  #ghfashionmarketing  #onlinebabyshop # | 2019-12-15 10:49:56 | | | | | | |
| 19 | B6FqetupE | http://instagram.coı | https://scontent-amt2-1.cdr | 11 | 217960127 | Fret not!!!! We dont lie nor scam.. when I was saying abrokyire smell some of u were laughing at me u see it's been confirme | 2019-12-15 10:44:30 | | | | | | |
| 20 | B6FpcwJFJ | http://instagram.coı | https://scontent-amt2-1.cdr | 39 | 12403251090 | Festive turkey dinner pinwheels 🦃  I puréed 2 medium sweet potatoes with a little milk and spread on top of the ready rolle | 2019-12-15 10:35:30 | | | | | | |
| 21 | B6Fm5junW | http://instagram.coı | https://scontent-amt2-1.cdr | 33 | 5805099614 | Wenn da Papi dem Sohnemann zum 1. Mal Apfel-Pflaume Mus füttert, dann haben beide a Gaude 😂😂😂 Ich liebe solche | 2019-12-15 10:13:13 | | | | | | |
| 22 | B6FlO3nlEc | http://instagram.coı | https://scontent-amt2-1.cdr | 13 | 21364618168 | Cheesy scrambled egg with carrot & cabbage 🥕🥬🥚 I hade some left over cooked veggies from last night's dinner so it w | 2019-12-15 09:58:39 | | | | | | |
| 23 | B6FjFSkjXA | http://instagram.coı | https://scontent-amt2-1.cdr | 53 | 21111321 | Thailand 🇹🇭 Adventures 🐘🐘 Feeding baby Michael 😍😍🥰🥰 #thailand #elephant🐘 #elephantsantuary #phuket #babyf | 2019-12-15 09:45:08 | | | | | | |
| 24 | B6FiIGDoM | http://instagram.coı | https://scontent-amt2-1.cdr | 0 | 6519154983 | 💎💎💎 visit www.lullabybabystore.com #baby accessories #pacifier #babyfeeding #babyshoes #maternity #39weekspregr | 2019-12-15 09:35:28 | TRUE | TRUE | FALSE | 25 | 205, 185, 168 | 37 |
| 25 | B6Fiiero-a9 | http://instagram.coı | https://scontent-amt2-1.cdr | 0 | 6519154983 | 💎💎💎 visit www.lullabybabystore.com #baby accessories #pacifier #babyfeeding #babyshoes #maternity #39weekspregr | 2019-12-15 09:35:07 | | | | | | |
| 26 | B6FiffKoZH | http://instagram.coı | https://scontent-amt2-1.cdr | 3 | 6519154983 | 💎💎💎 visit www.lullabybabystore.com #baby accessories #pacifier #babyfeeding #babyshoes #maternity #39weekspregr | 2019-12-15 09:34:42 | | | | | | |
| 27 | B6Ficy4o9h | http://instagram.coı | https://scontent-amt2-1.cdr | 0 | 6519154983 | 💎💎💎 visit www.lullabybabystore.com #baby accessories #pacifier #babyfeeding #babyshoes #maternity #39weekspregr | 2019-12-15 09:34:20 | | | | | | |
| 28 | B6FiZ4oo9 | http://instagram.coı | https://scontent-amt2-1.cdr | 0 | 6519154983 | 💎💎💎 visit www.lullabybabystore.com #baby accessories #pacifier #babyfeeding #babyshoes #maternity #39weekspregr | 2019-12-15 09:33:56 | | | | | | |
| 29 | B6FhSdyqV | http://instagram.coı | https://scontent-amt2-1.cdr | 6 | 5067180699 | 💎💎💎 visit www.lullabybabystore.com #baby accessories #pacifier #babyfeeding #babyshoes #maternity #39weekspregr | 2019-12-15 09:24:11 | | | | | | |

## 第二阶段：帖子分析

对于新scv，分析其每条条目，补全以下参数：

| | | 定义 | 示例 | 代码参考 |
|---|---|---|---|---|
| | Img_ Dominant _Color_RGB | 保留已经填好的数据 | 242, 221, 155 | |
| 文本分析 | Text_Emoji | “Text”中含有的Emoji | 🤦‍♀️ | |
| | Text_Emoji_Number | 统计“Text_Emoji”的数量 | 1 | |
| | Text_Hashtags | “Text”中含有的hashtag，用”;”隔开 | #weaning; #blw; #blw10months; #baby; #babyled; #babyledweaning; #babyledweaningideas; #blwideas; #babyledweaning10months; #babyledweaninguk; #weaningjourney; #weaningbaby; #babyfood; #babyfoodideas; #babyfeeding; #babymeals; #whatifeedmybaby; #breastfedbaby; #food; #lunch; #blwlunch; #eatingout | |

| | | 定义 | 示例 | 代码参考 |
|---|---|---|---|---|
| | Text_NoHashtag | “Text”中去掉hashtag余下的文本 | Lesson: don’t go out to eat with a teething baby who’s lost his appetite. 🤦‍♀️ Baby A ate about half a slice of toast and threw about half an egg to the floor before we gave up. Sigh. At least the staff at @crystalrockscafe were lovely as ever (and Mummy and Daddy enjoyed their meals!). | |
| | Text_NoHashtag_Length | “Text_NoHashtag”字数 | 55 | |
| | Text_NoHashtag_Sentiment | 链接到api，分析每句话的sentiment | Entire Document<br>Score: -0.2<br>Magnitude: 2.7<br>“Lesson: don’t go out to eat with a teething baby who’s lost his appetite.”<br>Score: -0.4<br>Magnitude: 0.4<br>“🤦‍♀️ Baby A ate about half a slice of toast and threw about half an egg to the floor before we gave up.”<br>Score: -0.6<br>Magnitude: 0.6<br>“Sigh.”<br>Score: -0.7<br>Magnitude: 0.7<br>“At least the staff at @crystalrockscafe were lovely as ever (and Mummy and Daddy enjoyed their meals!).”<br>Score:  0.9<br>Magnitude: 0.9 | https://cloud.google.com/natural-language/#how-automl-natural-language-works;<br>https://pypi.org/project/vaderSentiment/ |
| | Text_NounPhrases | 数据”Text”中含有的名词 | | https://pypi.org/project/spacy/ |
| | Text_Verbs | 数据”Text”中含有的动词 | go; eat; ate; sigh; threw, gave up; enjoyed | |
| | Text_emotions | 数据”Text”中含有的情感 | Lost; lovely; enjoyed | |

|  |  | 定义 | 示例 | 代码参考 |
|---|---|---|---|---|
| 图片识别 | Img_labels | 链接到api，分析图片的label | {<br>"status": {<br>"code": 200,<br>"msg": "OK"<br>},<br>"head": {<br>"method": "/predict",<br>"service": "classification_21k",<br>"time": 41<br>},<br>"body": {<br>"predictions": [<br>{<br>"classes": [<br>{<br>"prob": 0.073883056640625,<br>"cat": "Atrium"<br>},<br>{<br>"prob": 0.06305904686450958,<br>"cat": "Library"<br>},<br>{<br>"prob": 0.06247774139046669,<br>"cat": "Roof"<br>},<br>{<br>"prob": 0.03711871802806854,<br>"cat": "Proton accelerator"<br>}<br>],<br>"uri": "/data/example.jpg"<br>}<br>]<br>}<br>} | https://www.deepdetect.com/models/classification_21k/ |

|  |  | 定义 | 示例 | 代码参考 |
|---|---|---|---|---|
| 爬虫 | Owner_Account | “Owner”链接到的用户id | another_blw_account | https://github.com/realsirjoe/instagram-scraper |
|  | Owner_Username | “Owner”链接到的用户名 | Weaning Baby A |  |
|  | Owner_Followers | “Owner”链接到的用户粉丝数 | 614 |  |
|  | Owner_Posts_Number | “Owner”链接到的帖子数 | 487 |  |
|  | Owner_Intro | “Owner”链接到的用户简介 | No fancy photography or bamboo tableware, just a couple of foodie parents trying to figure out BLW with an 11 month old (24.01.19, EBF, 4 teeth). |  |