Analysis of Posts and Comments on Tech In Asia

Introduction

In this assignment, we crawled the posts and related comments by using the API provided. We clean, explore, and analyze the data, in order to reveal the correlation between features such as length and sentiment of posts and the number of comments.

Exploration and Analysis

In this analysis, we use Python 3.5, together with textblob as the NLP tool.

Firstly we crawl the post data from API and load it as JSON object. It has the following fields showing in the picture.

```
Every post contains following fields:
dict_keys(['seo', 'external_scripts', 'comment_count', 'companies', 'excerpt', 'id', 'editor', 'link', 'sponsor', 'slug'
, 'author', 'permissions', 'mobile', 'title', 'comment_status', 'is_sponsored', 'date_gmt', 'featured_image', 'modified_
gmt', 'comments', 'categories', 'tags', 'type', 'status', 'content', 'is_partnership', 'read_time'])
```

We are mainly interested in 'comment_count' ,'title' and 'content'.

Before we analyze the content, we need to clean the content data, as many special notations such as html tags appear in this part. Here is an example.

```
This is a piece of content before cleaning:
b'<div id="attachment_440717" class="wp-caption alignnone">\n<img class="size-full wp-image-440717" src="https://cdn.tec
hinasia.com/wp-content/uploads/2017/11/59cb10acN6f8aad2d.png" alt="" width="1365" height="767"><p class="wp-caption-text
">JDs automated sorting center in Kunshan. Photo credit: JD.</'
```

This is the result after cleaning.

```
This is the same content after cleaning:
b'JDs automated sorting center in Kunshan. Photo credit: JD.Heres a rundown of news from today and over the weekend.Ecom
merceGoogle to plunk down US$550 million in JD.com (China). The US tech giant is pouring US$550 million into the Chinese
 ecommerce powerhouse as part of a strategic partnership bet'
```

We explore the data by doing some basic stats.

For example, the average reading time for posts is 3.97 minutes, while the maximum is 22.

```
The average reading time for posts is     3.97 minutes.
The longest reading time for posts is    22.00 minutes.
```

The average number of comments is 0.69, while the maximum is 74!

```
The average number of comments for posts is     0.69.
The largest number of comments for posts is    74.00.
```

Now let us find out the most frequent nouns in the titles by using textblob. The following picture shows the most frequent nouns together with their frequency. Here, we are interested in nouns other than verbs or other words, as frequent nouns will tell us the frequent country name, company name and so on.

```
These are the most frequent nouns in title found by textblob, together with their frequency.

['asia:184', ''  s:139', 'news roundup:108', 'china:58', 'singapore:52', 'video:41', 'tech n
ews roundup  -:39', 'grab:38', 'alibaba:32', 'opinion:31', 'tia singapore:28', 'japan:24', '
uber:22', 'india:21', 'tech:20', 'go-jek:20', 'dec:20', 'ai:17', 'discuss:16', 'ipo:16', 'in
donesia:15', 'google:12', 'didi:11', 'nov:11', 'malaysia:11', 'asia singapore:10', 'ceo:9',
'facebook:9', 'carousell:9', 'wechat:8', ''  m:8', 'tencent:8', 'ico:8', 'startups:8', 'jan:
8', 'infographic:7', 'baidu:7', 'blockchain:7', 'flipkart:7', 'vc:7', 'wework:7', 'propertyg
uru:7', 'sea:7', 'vietnam:7', 'vr:6', 'ant:6', 'lazada:6', 'ola:6', 'thailand:6', 'xiaomi:6'
, 'mobike:6', 'grab-uber:5', 'startup:5', 'iflix:5', '$ 1b:5', ''  t:5', 'softbank:5', 'tech
 companies:5', 'vcs:5', 'amazon:5', 'reddoorz:4', 'ninja van:4', 'malaysian:4', 'meituan:4',
 'jd:4', 'samsung:3', 'myrepublic:3', 'zte:3', 'airbnb:3', 'philippines:3', 'icos:3', 'rotim
atic:3', 'rocket:3', 'japan asia:3', 'singaporean:3', 'eduardo saverin:3', 'razer:3', 'learn
:3', 'australia:3', 'don '  t:3', 'ama:3', ''  re:3', 'ecommerce:3', 'binance:3', '$ 500m:3'
, '$ 25m:3', ''  s tech giants:3', 'shopback:3', 'jack ma:3', 'toyota:3', 'ant financial:3',
 'singapore-based:3', 'taiwan:3', 'entrepreneurs:3', 'paytm:3', 'sequoia:3', 'will:3', 'kore
a:3', 'toutiao:3', '# tiasg2018:3']
```

(textblob is not totally accurate on finding the nouns.)

We can see that the most mentioned countries are: **China, Singapore, Japan, India, Indonesia, and Malaysia.** The most mentioned companies are: **Grab, Alibaba, Uber, go-jek, Google, Didi, Facebook, Carousell, Wechat and Tencent.**

If we have a look at the sentiment of post content, we can see that on average they are positive to neutral. (Polarity score great than 0 means positive while less than 0 means negative sentiment.) On average, the posts are of objective type. (subjectivity score less than 0.5 means objective.)

```
The average polarity score is   0.1076.
   91.00 percent of posts have a positive polarity score.
The average subjectivity score is   0.4150.
   86.00 percent of posts are of objective type.
```

What is related to the number of comments?

We compute the correlation coefficient between length, polarity, and subjectivity of posts and the number of comments. Surprisingly, none of the first three is strongly related with the number of comments!

```
correlation between length of content and no. comments:
[[ 1.         -0.04017234]
 [-0.04017234  1.        ]]
correlation between content positivity and no. comments:
[[1.         0.00516453]
 [0.00516453 1.        ]]
correlation between content subjectivityList and no. comments:
[[1.         0.06387298]
 [0.06387298 1.        ]]
```

However, if we consider the posts whose topic is one of the high frequency companies we mentioned before, we can see that the number of comments on these posts is above average. In other words, there is a correlation between the famous companies and number of comments. The following picture shows some examples.

```
china {'average no. comments': 0.7076923076923077, 'average polarity': 0.03807648650294531}
singapore {'average no. comments': 0.5566037735849056, 'average polarity': 0.033126172869302
487}
japan {'average no. comments': 0.2222222222222222, 'average polarity': 0.01782524607705767}
jack ma {'average no. comments': 2.3333333333333335, 'average polarity': 0.03201515151515152
}
grab {'average no. comments': 0.32, 'average polarity': 0.033549783549783545}
facebook {'average no. comments': 1.2222222222222223, 'average polarity': 0.0420497134038800
7}
go-jek {'average no. comments': 4.05, 'average polarity': 0.044953785448119246}
didi {'average no. comments': 0.18181818181818182, 'average polarity': 0.10886363636363637}
google {'average no. comments': 1.25, 'average polarity': 0.07433913975279108}
```

We can see that, Facebook, go-jeck, and Google have number of comments above average. (The average number is 0.69 as we computed before.) However Didi has very few comments although it is of high frequency. I think the reason is that its business is mainly in China. We can try more companies name in future work to make this correlation clear.

It seems that a frequently mentioned country name does mean a large number of comments.

Which post has the most comments? Here is the answer. It is a very short and in a colloquial style. It has 74 comments!

```
I'm Crystal Widjaja, SVP of business intelligence and growth at Go-Jek. AMA!
```