Xuanyue Zhou
Bomoda Data Exercise
January 7, 2016

1a. Count the number of posts mentioning each of the included brand names as well as the users who mentioned them. Hint: Pay attention to the variation of names.

Michael Kors: 4512 posts, 3564 unique users.
Kate Spade: 5426 posts, 3564 unique users.

1b. List the top 10 users and locations (as province level in China and nation level worldwide) for total posts.

Userid	Count
3956831443	156
5671523257	145
5689266274	144
3737544502	140
2417793200	70
3219312864	66
2263431535	65
3481576475	63
2204576563	60
2179810147	60

Location	Count
海外 美国	2958
海外	736
其他	638
广□ 广州	346
香港 九□城区	296
广□ 深圳	236
四川 成都	235
上海 浦□新区	230
海外 其他	183
海外 加拿大	181

2a. Find the date that has the highest number of posts mentioning each of the brands

Date: Sep 04, number of posts mentioning each of the brands: 599.

2b. Find the peak hour with the most posts.

Hour: 10am, number of posts: 847.

3. Tokenize the comments and retrieve the top 10 mentioned Chinese terms associated with each brand from all texts.

Top 10 terms associated with Michael Kors:

```
u'\u54e6'
u'\u55e8'
u'\u897f\u5b89'
u'\u60a8'
u'\u60b2\u4f24’
u'\u6362'
u'\u75af'
u'\u6012'
u'\u662f'
u'\u649e'
```

Top 10 terms associated with Kate Spade:

```
u'\u7231\u5fc3'
u'\u4eb2\u4eb2'
u'\u535a\u7269'
u'\u6316'
u'\u5965\u83b1'
u'\u4e48'
u'\u9f3b'
u'\u5475\u5475'
u'\u98df\u8c31'
u'\u751f\u65e5'
```

4. Possible sampling bias:

- Large part of Chinese population that does not use Weibo; typically people younger than 30 are more likely to use Weibo, thus they are over-represented in the sample population; the elderly who may not be capable of technology do not use Weibo, middle-aged men and women who are busy with work may not have time to use Weibo, they are all under-represented in the sampling.
- There are also rural population or population in poverty in China that do not have access to internet, not to mention Weibo; people living in cities and urban area who are wealthy are more likely to have options to choose whether they want to use Weibo, so they are over-represented in the sampling.
- There are many segments of Chinese population who are more willing to use other social media platforms instead of Weibo, but they are equally likely to purchase MK or KS from the U.S.

- Gender bias also exists in sampling on Weibo because women are general more social, more likely to be active users on Weibo, while men are less likely to be on Weibo .
- Sampling technics such as topic sampling or geographic sampling are biased based on contents and location, not achieving the goal of random sampling.

5. Some possible algorithms to identify users who showed interest in Michael Kors over Kate Spade:

- Look at user's posted photos or tagged photos on Weibo which may contain information related to Michael Kors
- Look at user's favorites which many consist of posts related MK they saved because they show more interest in MK
- Look at user's likes and comments to posts that may be related to MK
- Look at activities of user's followers and followings and activities of user interacting with his or her followers or followings to extract any information relating to MK

Some data points that could be used to illustrate the algorithm's utility:

u'reposts_count', u'category', u'comments_count', u'favorited', u'statuses_count', u'friends_count', u'has_ability_tag', u'follow_me', u'followers_count', u'verified_trade', u'verified_source_url', u'profile_url', u'description', u'screen_name', u'remark', u'favourites_count', u'following', u'pagefriends_count', u'urank'

6. Algorithms to implement sentiment analysis:

- Named entity recognition – tag associated people and objects positive or negative
- Subjectivity classification – classify sentences as subjective or objective; subjective sentences will have sentiments
- Feature selection – use unigrams, bigrams, or n-grams with or without punctuations as features
- Supervised learning models to use: Naïve Bayes, support vector classifiers, logistic regression, tree classifiers: decision tree, random forest, gradient boosting and etc., maximum entropy models

Pros and cons:
tree classifiers may not be space optimal, lead to overfitting; naïve bayes model is very fast, would be a good baseline algorithm; support vector machines are not optimal for skewed or unbalanced datasets and they requires more computation power; maximum entropy classifier is a great parameterized method, uses search-based optimization, usually performs a little better than naïve bayes model.

7. How would you be able to identify brand names in text?

Identify variations of Chinese translations of brand names, examine word frequencies and co-occurrences, look for terms that often occurs with the brand names, check the context and make sure it fits with the mentioned brand names.