STAT 607 - Assignment 2

Name: Zhen Qin, Uniqname: qinzhen

1.1

The top 10 websites by pagerank are:

> ['yahoo.com' 'rea-group.com' 'wikimedia.org' 'tumblr.com' 'google.com' 'youtube.com' 'canalblog.com' 'wikipedia.org' 'creativecommons.org' 'blogspot.com']

Search the top 10 nodes according to in-degree:

```
tmp = np.argsort(in_link)[-10:]
page_indices[tmp]
```

> array([['creativecommons.org', '32'], ['tumblr.com', '88'], ['amazon.cn', '6'], ['amazon.co.uk', '8'], ['shopbop.com', '82'], ['amazon.fr', '12'], ['flickr.com', '44'], ['amazon.ca', '5'], ['wikipedia.org', '95'], ['blogspot.com', '25']], dtype='|S30')

There are four nodes in common. According to the algorithm, more in-degrees tend to result in higher pagerank. This is partially verified by the outcome.

1.2

Power iteration computation result `pageranks` agree with result of `eig`.

> Pagerank computations via power method and numpy.linalg.eig agree

2.1

Linear kernel: 0.894, polynomial kernel: 0.901, rbf/gaussian kernel: 0.927, sigmoid/arctan kernel: 0.907.

rbf/gaussian kernel gave the best accuracy.

2.2

False Positive Rate is extremely higher than False Negative Rate. That is to say, Positive is far more difficult to classify.

3.1

Without using loops, the accuracy was calculated as:

> Accuracy: 0.921

3.2

After test, the test accuracy seemed to converge to a number that is less than 1.

First, I do not think the accuracy will increase to 1 because you cannot generate a perfect classifier from a part of the sample space. Once there is one single outlier that is way different from training samples, the accuracy will never be 1.

Second, I do not think the accuracy will increase up to a point and then decrease. According to Adaboost theory, the false rate is less than $\prod_m Z_m$. When accuracy is close to 1 after enough iterations, for more indexes, $e^{-\alpha y_i G(x_i)}$ is less than 1. That is to say Z_m is not likely to increase a lot. Thus I believe the accuracy will not decrease.

When it comes to oscillation, maybe in some special cases the accuracy will oscillate but I believe the situation is not common. Actually, in the model I cannot find any periodical factors leading to this kind of results.

In fact, I think in most cases the accuracy will not increase or increase slowly but converge to a value less than 1. Since theoretically the accuracy will not decrease a lot and be approximately monotone and have an upper bound 1, it will converge to a point due to monotone bounded theorem.