Unstructured Data Analytics **Homework 2**
Fall 2017 **October 27, 2017**

Deadlines Homework 2 is due on Nov.10th 11:55pm. Therefore, you have in total 14 days to finish this homework. The late policy for all our homework has been carefully discussed in the first lecture.

How to submit: Please submit a zipped file to the following email address: kaizhaofrank AT gmail.com. The title of the email should be 'Unstructured Data Analytics-HW2' and the file name should be 'Yourname-Pantherid.zip'. In the zipped folder it should contain the homework data set and two separate ipython notebook files '1-topic-model.ipynb' and '2-topic-change.ipynb', for the first and second problem respectively.

Data Set: The instructor has prepared a small collection of Amazon reviews (626 review comments from 19 products) in the file Homework_1.csv. The data set is in the iCollege under the folder Homeworks. The columns are named and organized in the following manner: ProductID, ReviewID, ReviewTitle, ReviewTime, Verified, ReviewContent, ReviewRating.

---

1. (10 points) Topic Modeling

   We have in total 19 products in the data set, and the comments from each product can be viewed as a separate document. Please build a topic model and show the top 2 related topics for each product and its corresponding probability.

   Hint: you should perform stop-word removal and stemming before generating the topic models. You can choose the number of topics yourself and there should be no less than three topics.

   What to submit:

   An ipython notebook file displaying the product id and the corresponding topic and its probability. For example:

   Product ID, Probability*Topic 1, Probability*Topic 2

   1 (e.g., Apple IphoneX), 0.34*[screen, battery, mobile], 0.22*[good, great, excellent]

   2 (e.g., Benz car), 0.33*[car, expensive, fast], 0.21*[quality, german, make]

   ...

   19 (e.g., Coca Cola), 0.21*[cool, diet, young], 0.15*[good, great, excellent]

2. (5 points) Change of topics over time

   Product ID 8 is a coffee grinder and we want to examine the customer opinions on it over time. We can achieve it by examining the change of topics of the comments. The comments for each day from product ID 8 can be viewed as a separate document. Please build a topic model and show the top 2 related topics for each day and its corresponding probability. The data cleaning and preprocessing process should follow the guidance from the first problem.

What to submit: An ipython notebook file displaying the time (in an ascending order), and the corresponding topic and its probability. For example:

Time, Probability*Topic 1, Probability*Topic 2

2015-11-10, 0.21*[great, coffee], 0.15*[bean, coffee]

...

2016-01-17, 0.34*[easy, use], 0.22*[bean, coffee]

2016-01-19, 0.33*[not, buy], 0.21*[huge, mess]

...