Jiann-Liang Tsay <jlt245@nyu.edu>

Project requirements

Raman Kannan <rk2153@gmail.com> Sat, Sep 24, 2016 at 5:03 AM
To: Jiann-Liang Tsay <jlt245@nyu.edu>

I approve this project.

JLT, don't be stressed about the schedule. If you need more time to do it right
you have it.

regards
Raman

On Sat, Sep 24, 2016 at 4:17 AM, Jiann-Liang Tsay <jlt245@nyu.edu> wrote:

> Hi Raman,
>
> Before I forward my project-1 proposal, I would like to give you a draft as below and present my concerns in terms of timing and available resources.
>
> Splunk product is an IR system specializing in time-series events of query and retrieval with rich visualization. However we use very primitive parsing skill to parse unstructured text or data into events. It also misses text analytics or algorithms that can be applied widely to many areas and industries such as e-commerce, new product information or trends etc.
>
> I consider to use the project-1 as a POC (Proof of Concept) for a possible product candidate as: clustering time-series text stream. My idea is:
>
> 1. In feature selection, I plan to format events from text as the events of the attributes as (time, location, person or people, topic) to present the text. Those are interested features for now. (In project-2, perhaps I can extend it to more general feature selection algorithm and strategy)
> 2. Topic extraction may be from the popular algo such as topic modeling, LDA (Latent Dirichlet Allocation) with keywords in the topic. Keywords are important for calculating distance and similarities.
> 3. I would like to try out an algo called OSKM (online spherical k-means) which is basically useful for clustering data streams. This algo partitions data set in terms of time interval and fades old clusters if not frequently used to response the most recent trends.
> 4. The dataset may be from twitter, weblogs or emails that I will decide later.
> 5. The system can work as event detection as well that is an outliers analytics. For example, within a time interval, an event comes up without going into any cluster, that is an outlier.
> 6. I will answer (visualize) a few questions or applications from this system. Questions are like the recent trend of clothing color or style for example...
>
> My concerns are:
>
> 1. How much R libraries can support those algorithms and calculations? I am not familiar with R. Since I only have three weeks for the project, if I need to use programming languages such as python or java to implement most of them, I doubt I can finish it on time.
> 2. I may not be able to do anything in performance evaluation as it will heavily rely on R, neither it is parallel nor comparable to benchmarks?
> 3. In project-2, I may consider to use java to implement the system (like solr is a good text indexing system that I can extend it for that.)
>
> references:
> http://www.sciencedirect.com/science/article/pii/S0893608005001413
> A survey of text clustering algorithms by Charu C. Aggarwal
>
> Let me know any comments or recommendation in anything, resources, goals etc...
>
> Thanks,

> Jiann jlt245.
>
> On Fri, Sep 23, 2016 at 7:35 PM, Raman Kannan <rk1750@nyu.edu> wrote:
>
>> Anna
>> I have never misled anyone about the amount of work required for the course.
>>
>> Malcolm Gladwell says it takes 10000 hours
>> to learn a new subject.
>>
>> But please take a note:
>> I have combined requirements for all three
>> Courses I teach. May be that makes it appear.
>>
>> Also note that every item is not required for 6923.
>>
>> Regards
>> Raman
>>
>> --
>> You received this message because you are subscribed to the Google Groups "NYU-CS-6923-fall-2016" group.
>> To unsubscribe from this group and stop receiving emails from it, send an email to nyu-cs-6923-fall-2016+unsubscribe@googlegroups.com.
>> To post to this group, send email to nyu-cs-6923-fall-2016@googlegroups.com.
>> To view this discussion on the web visit https://groups.google.com/d/msgid/nyu-cs-6923-fall-2016/CAJ78gn-mBS5DLUSrm%3DHPBEDdSTqssWHULQ19GLYtAb23P6AtLQ%40mail.gmail.com.
>> For more options, visit https://groups.google.com/d/optout.