Introduction

You are thinking about starting research in data mining and related fields. Your friend who is already working on these fields has told you that conferences usually happen in nice and exotic locations and even though this should not affect your decision, knowing that you will get to travel to an exotic location to present your work is always some extra motivation. In this assignment you are going to (empirically) verify your friend’s statement by mining WikiCFP http://www.wikicfp.com/cfp/ a website that contains calls for papers for a wide variety of conferences for multiple fields. You will have to 1) crawl the data, 2) clean the data, 3) use Spark to compute various statistics of the data, and 4) visualize the results.

Phase 2 [35/100]

Spark [25/100]: In this part you will use Spark on the data you just crawled to compute various statistics. The size of the data you have crawled is actually pretty small and is not what Spark is made for, but the purpose of this assignment is to get you acquainted with Spark nevertheless. Since you have already successfully completed Phase 1.5, you should have a working installation of Spark in your machine.

Make sure you understand how the “WordCount” example, which we saw in class, works: https://spark.apache.org/examples.html

Finally, after you have successfully made “WordCount” work, you will have to do the following computations:
1. Compute and output the number of conferences per city. Which are the top 10 locations? [10/40]
2. Output the list of conferences per city [10/40]
3. For each conference regardless of the year (e.g., KDD), output the list of cities. [10/40]
4. For each city compute and output a time series of number of conferences per year [10/40]

You may have a separate Hadoop program per computation.

Note: If you can demonstrate to the instructor that you have extensive prior experience with Hadoop, only with instructor’s permission, you may use Spark instead of Hadoop for this part of the assignment.
**Result Visualization [10/100]:** As we discussed in class, visualizing the results in an intuitive way is extremely important, especially when external input from domain experts is needed. Here, you have to visualize the output of question #4 of the Hadoop section (i.e., the number of conferences per year, for each city); the visualization must be done on a map and the number of conferences can be represented as a heatmap (“red”/”hot” means a lot of conferences, “blue”/”cold” means few conferences) with an associated assignment of colors to numbers, or with variable size dots (large dot means a lot of conferences, small dots means a few). You are free to use any map interface you wish, however, you must clearly describe your approach (providing screenshots or code, whenever applicable) and provide image files with the visualizations.

**Deliverables**

Each phase consists of the following two deliverables:

1. A report where you document your solution for each part (including the screenshots of the data cleaning). The report should also include the results of the third part, as well as the plots required.
2. A copy of all the code you wrote for the crawler and the Hadoop programs. Please make sure you properly document your code with comments.

Each deliverable should be uploaded as an archive on iLearn.

**Grading scheme**

The respective points for each part of the assignment are shown in the description. For each part we are going to evaluate both how well you describe your approach (and this is based on the report you will submit) as well as whether you actually implemented everything successfully and generated the desired results. In particular, the breakdown (per each section of each phase) is:

1. 10% for describing your approach in detail.
2. 90% for implementation and results

**Academic Integrity**

If you use any external packages or help from the web (e.g., StackOverflow) please cite your sources! Same goes if one of your colleagues helped you with some part. You will not lose points using such help as long as you cite your sources.