Text Analysis and Entity Resolution

Entity resolution is a common, yet difficult problem in data cleaning and integration. In this lab exercise, we will use powerful and scalable text analysis techniques to perform entity resolution across two data sets of commercial products.

Entity Resolution, or "[Record linkage](https://en.wikipedia.org/wiki/Record_linkage)" is the term used by statisticians, epidemiologists, and historians, among others, to describe the process of joining records from one data source with another that describe the same entity. Our terms with the same meaning include,  "entity disambiguation/linking", "duplicate detection", "deduplication", "record matching", "(reference) reconciliation", "object identification", "data/information integration", and "conflation".

Entity Resolution (ER) refers to the task of finding records in a data set that refer to the same entity across different data sources (e.g., data files, books, websites, databases). ER is necessary when joining data sets based on entities that may or may not share a common identifier (e.g., database key, URI, National identification number), as may be the case due to differences in record shape, storage location, and/or curator style or preference. A data set that has undergone ER may be referred to as being cross-linked.

The lab is due Jun 26, 2015 at 00:00 UTC. There is a three day grace period for late submissions until Jun 29, 2015 at 00:00 UTC. Submissions after that time will lose 20 points.

HOW TO COMPLETE THIS ASSIGNMENT

This assignment is broken up into sections with bite-sized examples for demonstrating Spark functionality for Entity Resolution For each problem, you should start by thinking about the algorithm that you will use to *efficiently* process the log in a parallel, distributed manner. This means using the various [RDD](http://spark.apache.org/docs/latest/api/python/pyspark.html#pyspark.RDD) operations along with [lambda functions](https://docs.python.org/2/tutorial/controlflow.html#lambda-expressions) that are applied at each worker.

INSTRUCTIONS

Here are the instructions.

1. Start the VM - To start the VM, from a DOS prompt (Windows) or Terminal (Mac/Linux), issue the command "vagrant up", while in the custom directory [created](https://courses.edx.org/courses/BerkeleyX/CS100.1x/1T2015/courseware/d1f293d0cb53466dbb5c0cd81f55b45b/920d3370060540c8b21d56f05c64bdda/) for this course.
2. Once the Virtual Machine is running, access the Jupyter web UI for running IPython notebooks by navigating your web browser to "<http://localhost:8001/>" (or "<http://127.0.0.1:8001/>").
3. If you have any running notebooks they **SHOULD BE** shutdown.  Only **ONE** notebook should be run at a time.  Running notebooks have a green icon to the left of the notebook name and green text to the right of the screen that says "Running".  Shutdown running notebooks by clicking the checkbox next to the notebook and then clicking the orange "Shutdown" button.
4. Download the Lab 3 IPython notebook.  Make sure that the file extension is .ipynb.  If the download adds an extension (e.g. ".txt"), rename the file so that the extension is just .ipynb.
   * Lab 3 Text Analysis and Entity Resolution: <https://raw.githubusercontent.com/spark-mooc/mooc-setup/master/lab3_text_analysis_and_entity_resolution_student.ipynb>. You can view this lab exercise online [here](http://nbviewer.ipython.org/github/spark-mooc/mooc-setup/blob/master/lab3_text_analysis_and_entity_resolution_student.ipynb).
5. Upload the IPython notebook.  This process was explained during "[Setting up the Course Software Environment](https://courses.edx.org/courses/BerkeleyX/CS100.1x/1T2015/courseware/d1f293d0cb53466dbb5c0cd81f55b45b/920d3370060540c8b21d56f05c64bdda/)" in week one.
6. In the **Lab 3 Text Analysis and Entity Resolution** notebook, please follow the instructions in the notebook and replace <FILL IN> sections with your solution.  After you confirm that your code passes all of the tests while running in the VM, please export it as a **python file (.py)**and submit it to the autograder server.  The submission process is the same as in the previous week for Lab 2 - *make sure you submit to the correct autograder.*  In the next module, the instructions are provided again for your convenience.
7. An outline of what will be covered in the notebook is included below.
8. When you have submitted successfully, you can shutdown the VM by issuing the command "vagrant halt".

TEXT ANALYSIS AND ENTITY RESOLUTION EXERCISE

This exercise consists of 5 parts and quiz questions:

* *Part 1*: ER as Text Similarity - Bags of Words
* *Part 2*: ER as Text Similarity - Weighted Bag-of-Words using TF-IDF
* *Part 3*: ER as Text Similarity - Cosine Similarity
* *Part 4*: Scalable ER
* *Part 5*: Analysis (this is part where you will click through and view plots of your work from part 4)

After you complete the lab, make sure you answer the quiz questions in the next section.

**Please do not post your programming exercises in publicly visible repositories, such as GitHub.**