**Homework #7 -- Further Adventures in MapReduce & Java 8**

Last Modified: *11/20/2017 19:30:45*

Due: Monday December 4th Grading: 15%

You will be working with a set of files containing [JSON](http://www.json.org/)-formatted records of the form (URL, timestamp, set of tags). One record represents the fact that someone visited the URL at the given time, and labelled the visit with a number of tags. Here is some sample data:

{"url": "http://freegasmoney.org/", "timestamp": 1257034785, "tags": ["free", "gas", "coupon"]}

{"url": "http://www.autodriver.com/car-loans/pre-approved-auto-loans/index.html", "timestamp": 1257119583, "tags": ["approved", "car", "loans", "auto", "loan", "approval", "pre"]}

{"url": "http://www.simplykidstoys.com/precious-princess-perfect-dollhouse.html", "timestamp": 1257107683, "tags": ["toys", "birthday", "christmas", "dollhouse"]}

{"url": "http://thenuweigh.com/", "timestamp": 1257042871, "tags": ["diet", "diets", "fatloss", "weight", "loss", "program"]}

The timestamps represent the number of seconds since January 1, 1970. This is a very common way of representing time in software.

This data is from a [research project](http://cnets.indiana.edu/groups/nan/givealink). It is organized into 15 files, for a total of 11MB.

**Download the data**

Download the data from [here](https://courses.dce.harvard.edu/~cscie55/links.tgz). You can unpack the data using the command **tar xvzf links.tgz**. The data will be unpacked in a subdirectory **files.1**.

The data originates [here](http://carl.cs.indiana.edu/data/websci2014/givealink-nov-2009.tgz), but it needed some cleaned-up, and reorganization into two directories. Do NOT use this original data.

**Problems to solve**

**Problem 1.** Using the **Hadoop API's** write a **Map/Reduce** solution that lists the url's with the (comma-separated) union of the tags listed in all the records for each url. For example, suppose the following two records were seen in the stream:

{"url": "http://freegasmoney.org/", "timestamp": 1257034785, "tags": ["free", "gas"]}

{"url": "http://freegasmoney.org/", "timestamp": 1257034877, "tags": ["free", "gas", "vacation"]}

The final output from your solution would have one line:

**http://freegasmoney.org/ free,gas,vacation**

Note: the tags are comma-separated, no white space, and the key/value pair is separated by a single white space. Run your solution on the local file system, i.e. not against the Hadoop cluster. Capture the first ten lines of the final output and submit it with your solution. For correctness you can compare your results with this reference result: [prob1-first-ten](https://courses.dce.harvard.edu/~cscie55/prob1-first-ten.txt)

*Hint:* One way to accumulate the tags is for the **Reducer** to defer writing any results until it executes a **cleanup** after the last **reduce** method runs, meanwhile storing tags in a **java.util.Set**, which **cleanup** finally uses to write the result.

To parse the lines of data use this Java code [Link.zip](https://courses.dce.harvard.edu/~cscie55/Link.zip)    **Link.java** has a static method **parse** that takes a **String**. In **Mapper.map** you can run **Link.parse** on each incoming line. This will produce a **Link** object. Note that **Link.parser** returns **null** when it encounters bogus data. [Suggestion: extract a small set of records, say 5 lines, into a single file to run your code on while developing your solution. Two gotchas: (1) blank lines in a data file, (2) editor backup files that your editor may write out in the same directory, like emacs' ~ recovery files. Either of these will cause problems.]

**Problem 2.** Solve the same problem using Java 8 features, i.e. Lambda Expressions, Function References and **Stream**s. A single Java class should be all that is necessary. Call it **LinkStreamer**, in the same package as above, **cscie55.hw7**     Usage:

**java cscie55.hw7.LinkStreamer input-directory output-file [dd-MM-yyyy dd-MM-yyyy]**

When the program is called with two arguments it will look in the named input directory and stream the content of all the files it finds there, writing out the result to the named output file. Output will have the form:

**http://haberdashers.com/ hats, coats, gloves**

Add a new feature that counts occurrences of each URL between a range of dates when given on the command line. Thus, when the program is called with two additional arguments to specify a date range the program will write out a total count of the occurences, output looking like this:

**http://haberdahser.com/ 21**

The date format to support is **dd-MM-yyyy**

In order to generate a **long** value for comparison to the **timestamp** of each **Link** in the stream you will need some Java date API's. These will do the trick:

**SimpleDateFormat simpleDateFormat = new SimpleDateFormat("dd-MM-yyyy");**

**simpleDateFormat.setTimeZone(TimeZone.getTimeZone("EST"));**

**Date startDate = simpleDateFormat.parse(args[2]);**

**Long startseconds = startDate.getTime()/1000;**

The division by 1000 converts milliseconds to seconds.

Considerations:

* First command line argument will be the directory name. Solution will have to iterate over the files in that directory.
* In order to integrate all the data in those files the processing of the files as individual streams will need to maintain some state from stream to stream, all the while parsing the lines of input and turning each one into a **Link**.
* Once your solution has collected the result, have it print to standard out the pairs **<URL, Tags>**, as above in the problem statement for Hadoop.
* Collect the output in a file and include it with your solution.
* Skip generating javadoc. Instead, comment the code to narrate the algorithm.
* Run your solution, giving start date as November 8, 2009 and end date as November 9, 2009.

**Submission requirements**

A zip file containing

* All the Java files for each problem. [All classes should belong to package **cscie55.hw7**]
* The output files created by running the MR code (locally) for each driver class containing **main**.
  + first 10 lines [**part-r-00000**] for problem 1,
  + the entire output file for problem 2. [Ref: [prob2-part-r-00000](https://courses.dce.harvard.edu/~cscie55/time-link-part-2-00000)]
* A jar file containing all class files in a form suitable to run solutions for all two problems.
* A solution overview, explaining each of the solutions.

http://09hannah09.blogspot.com/ pets,cars,jobs,massage,events,life

http://09hannah09.blogspot.com/2009/04/my-journey-begins.html pets,cars,jobs,massage,life,events

http://09hannah09.blogspot.com/2009/06/free-classified-job-on-not-so-ordinary.html hannah,ad,service,online,classifieds

http://09hannah09.blogspot.com/2009/06/is-classified-posting-service-can-be.html service,classifiedsposting,free

http://09hannah09.blogspot.com/2009/07/photography.html service,free

http://09hannah09.blogspot.com/2009/07/summertime-and-garden-needs-sorting.html service,free

http://1000paydayloan.blogsarena.com/ loan,1000,cash

http://10poundslighter.com/ loss,weight

http://10starmovies.com/ movies,watch,documentaries,online,free,tv-shows

http://192.168.1.254/ compter&amp,language

http://affordable-medical-equipment-supply.com/150-practice-ecgs-interpretation-and-review/ 1

http://affordable-medical-equipment-supply.com/lippincott-manual-of-nursing-practice-series-ecg-interpretation/ 1

http://alturl.com/5tzz 2

http://alturl.com/7q6g 1

http://alturl.com/cyzd 1

http://alturl.com/czrr 2

http://alturl.com/tizh 1

http://antispyware.wetpaint.com/ 1

http://antispywarefix.livejournal.com/ 1

http://autoauctionexpress.com/ 1

http://autoauctionexpress.com/auto-auction-camarillo.html 1

http://autoauctionexpress.com/auto-auction-del-mar.html 1

http://autoauctionexpress.com/category/los-angeles 1

http://berkline.livejournal.com/895.html 1

http://berklinesofa.vox.com/library/post/palliser-leather-recliner-a-comfort-brand.html 1

http://bit.ly/2SUBCm 1

http://bit.ly/2uuWz4 1

http://bit.ly/3AhkKn 2

http://bit.ly/41Y23Y 2

http://bluetoothheadset01.blogspot.com/2009/09/your-guide-to-bluetooth-headsets.html 1

http://cloudjobs.net/?page=register\_employer 1

http://condohotelsplayadelcarmenmexico.blogspot.com/2009/11/playa-del-carmen-condo-rental-for-great.html 1

http://dogjointpainrelief.com/ 2

http://emailaffirmations.com/ 1

http://ezinearticles.com/?Building-Your-Own-Wind-Turbine-to-Save-You-Some-Money-Using-Diy-Wind-Power&id=3213466 1

http://ezinearticles.com/?Discover-How-to-Tell-a-Girl-You-Love-Her-the-Right-Way-and-How-to-Keep-the-Fire-Burning&id=3193849 1

http://ezinearticles.com/?Discover-the-Incredible-Weapon-Against-a-Slow-Computer---The-XP-Registry-Repair-Utility&id=3181538 1

http://ezinearticles.com/?Finding-a-Really-Effective-Win-XP-Registry-Cleaner&id=3194875 1

http://ezinearticles.com/?Four-Different-and-Surprising-Ways-to-Make-a-Girl-Like-You&id=3193859 1

http://ezinearticles.com/?Get-to-Know-the-Four-Codes-of-the-XBox-360-Red-Rings-of-Death&id=3194759 1

http://ezinearticles.com/?Great-and-Proven-Bacterial-Vaginosis-Home-Remedy&id=3213793 1

http://ezinearticles.com/?How-to-Prevent-the-Xbox-360-Two-Red-Lights&id=3207518 1

http://ezinearticles.com/?Increase-Your-Computers-Speed-Significantly-With-an-XP-Registry-Repair-Utility&id=3181547 1

http://ezinearticles.com/?Kick-Out-Those-Vista-Registry-Errors-in-Your-System&id=3194863 1

http://ezinearticles.com/?Learn-How-You-Can-Make-Girl-Want-You-With-a-Few-Smart-Tricks&id=3194044 1

http://ezinearticles.com/?Prevent-the-Vicious-Red-Lights-on-Xbox-360-Cycle&id=3207494 1

http://ezinearticles.com/?The-Surefire-Method-For-a-Fast-Microsoft-XP-Registry-Repair-That-Will-Last-Forever&id=3167448 1

http://ezinearticles.com/?Top-Three-Xbox-360-Red-Light-Fix-Options&id=3207507 1

http://ezinearticles.com/?What-Are-Registry-Repair-Programs-Designed-to-Do?&id=3194857 1

http://ezinearticles.com/?What-Xbox-360-3-Light-Fix-is-Available-on-the-Internet?&id=3207498 1

http://family.webshots.com/video/3040544300105558921EHuDBQ 1

http://freecallindia.org/call-india-for-free 1

http://freecallindia.org/call-india-mobile-free 1

http://funtennis.blogspot.com/2009/11/roger-federer-vs-novak-djokovic-swiss.html 1

http://healthcaretips.t35.com/ 1

http://home-decorating-made-easy.blogspot.com/ 1

http://internship-tips.com/ 1

http://ipright.vn/Home/ 2

http://koransaya.blogspot.com/2009/11/rain-asia-tour-konser-musik-jadwal.html 1

http://landprofitgeneratorbonusreview.com/47/do-you-want-to-learn-how-to-buy-land-for-pennies-then-flip-that-land-for-thousands/ 2

http://lasvegas-show.net/las-vegas-season-four/ 1

http://orderpropecia01.blogspot.com/2009/02/super-quality-propecia-or-proscar.html 2

http://powerpoint-dvd-presentation.blogspot.com/2009/11/smiths-new-idea-for-thanksgiving-family.html 1

http://probleminventory.com/ 1

http://sandiego.prohomeinspector.org/energy-inspector%e2%80%99s-guide-%e2%80%93-commercial-and-residential-provisions-of-the-2006-international-energy-conservation-code/ 2

http://sandiego.prohomeinspector.org/home-inspection-report/ 1

http://sandiego.prohomeinspector.org/inspect-before-you-buy-insider-secrets-you-need-to-know-about-home-inspection-with-companion-cd-rom/ 2

http://sandiego.prohomeinspector.org/the-professional-inspectors-guide-to-residential-roof-inspection/ 1

http://santabarbaraminister.com/ 1

http://televerus.net/ 1

http://tgblogsite.com/dyoseftg/?p=27 1

http://tiny.cc/2GbzS 1

http://tiny.cc/Dmghf 1

http://tiny.cc/IOYed 1

http://tiny.cc/InQGk 1

http://tiny.cc/KeiT4 1

http://tiny.cc/M84tL 1

http://tiny.cc/aCVg9 1

http://tiny.cc/lAlmK 1

http://tradecarcalifornia.com/category/trade-car-california 1

http://tradecarcalifornia.com/sell-your-car-for-cash.html 1

http://tradecarcalifornia.com/trade-car-for-cash-california.html 1

http://virtual-assistants.outsourcing-services-india.com/ 1

http://www.123employee.com/ 1

http://www.2020image.com/ 1

http://www.aboutcaillou.com/ 1

http://www.actt.edu.au/ 1

http://www.acupuncturefordepression.info/depression/ 1

http://www.adsbyvideo.com/ 2

http://www.afroromance.com/members/United\_States/New\_York/Babylon/1 1

http://www.ankaweb.com/ideas-on-wedding-invitations-for-a-stylish-wedding/ 1

http://www.arcurs.com/ 1

http://www.articlesbase.com/video-games-articles/a-mind-blowing-xbox-360-repair-guide-1431524.html 1

http://www.articlesbase.com/video-games-articles/a-simple-but-effective-xbox-360-red-ring-fix-1431516.html 1

http://www.articlesbase.com/video-games-articles/excellent-xbox-360-troubleshooting-steps-1431531.html 1

http://www.articlesbase.com/video-games-articles/fixing-the-xbox-360-two-red-lights-syndrome-1431538.html 1

http://www.articlesbase.com/video-games-articles/great-methods-for-xbox-rrod-fix-1431545.html 1

http://www.articlesbase.com/video-games-articles/the-best-and-ultimate-xbox-rrod-fix-1431544.html 1

http://www.articlesbase.com/video-games-articles/the-solution-to-xbox-360-red-lights-flashing-1431514.html 1

http://www.articlesbase.com/video-games-articles/tips-and-tricks-in-fixing-the-xbox-360-three-flashing-lights-problem-1431529.html 1

http://www.articlesbase.com/video-games-articles/top-three-360-rrod-repair-approaches-1431198.html 1

http://www.articlesbase.com/video-games-articles/xbox-360-3-light-fix-easy-way-to-fix-the-error-yourself-1431486.html 2

http://www.articlesbase.com/video-games-articles/xbox-360-3-red-lights-2-top-tips-to-end-xbox-360-red-lights-and-enjoy-uninterrupted-gaming-fun-1431160.html 1

http://www.articlesbase.com/video-games-articles/xbox-360-red-light-blinking-amazingly-simple-way-to-get-rid-of-the-xbox-360-blinking-red-lights-for-good-1431157.html 1

http://www.articlesbase.com/video-games-articles/xbox-360-red-rings-of-death-put-a-full-stop-to-the-error-with-these-3-simple-tips-1431153.html 1

http://www.articlesbase.com/video-games-articles/xbox-360-two-red-lights-problem-solutions-1431536.html 1

http://www.articlesbase.com/video-games-articles/you-can-fix-your-consoles-xbox-360-red-light-blinking-issue-1431550.html 1

http://www.articlesbase.com/video-games-articles/your-basic-xbox-360-troubleshooting-guide-1431535.html 1

http://www.asiftosay.com/health-of-colon-is-indeed-important/ 1

http://www.ayushitravel.com/ 1

http://www.bestnonslip.com.au/ 1

http://www.besttirepressuregauge.com/ 1

http://www.businesscreditcardprocessing.com/industry.php?ind=15 1

http://www.cafechorti.org/excellent-turbo-colon-cleanse-products/ 1

http://www.catspecialties.com/custom\_tshirts-W25C282.aspx 1

http://www.ceplainmaculada.com/hgh-play-vital-role-for-human-body/ 1

http://www.chinaorientaltravel.com/adoption-tour.php 1

http://www.churchilldentalpractice.com.au/ 1

http://www.condohotelsplayadelcarmen.com/around\_playa\_del\_carmen.php 1

http://www.conjoingroup.com/about\_conjoin/richard\_garnick.html 2

http://www.constructionwest.com/ 1

http://www.cosmeticmedical.com.au/proc/bleph.html 1

http://www.creditsolutions.com/ 1

http://www.dallascosmeticcenter.com/ 3

http://www.dallascosmeticcenter.com/injectables.html 1

http://www.dallascosmeticcenter.com/laser.html 2

http://www.dallascosmeticcenter.com/procedures\_restylane.html 1

http://www.dallascosmeticcenter.com/procedures\_sculptra.html 1

http://www.duckriver.biz/creative-wedding-stationary-ideas 1

http://www.e-tradecounter.co.uk/p-407-va100-axial-bathroom-extractor-fan.aspx 1

http://www.ebest24.co.uk/car-36-led-turn-signal-parking-tail-backup-light-lamp-bulb-panel-3427.html 1

http://www.ebest24.com/en/eyes-exercise-black-pinhole-glasses-vision-eyesight-improve-7469.html 1

http://www.ebottleo.com.au/products\_new.php 1

http://www.enhancecosmeticsurgery.com.au/ 1

http://www.erectarack.com.au/mezzanine\_floor.php 1

http://www.fairviewgolfcenter.com/ 1

http://www.futuresustainability.com.au/future-insulation.php 1

http://www.gemablog.com/how-to-cure-a-yeast-infection/ 1

http://www.gemablog.com/natural-cure-for-yeast-infection/ 1

http://www.getbackincontrol.co.uk/ 1

http://www.global-information-network-society.com/ 2

http://www.goalsforwomen.com.au/polly\_peres.html 1

http://www.goarticles.com/cgi-bin/showa.cgi?C=2099079 1

http://www.goarticles.com/cgi-bin/showa.cgi?C=2154254 1

http://www.granicreteaustralia.com.au/concrete\_solu.php 1

http://www.gregwriter.com/ceospace 1

http://www.hatchandhatch.com.au/ 2

http://www.herbalcureindia.com/female-health/m2-tone.html 1

http://www.herbalcureindia.com/products/shilajit.htm 1

http://www.hirewebdevelopersindia.com/hire-aspstorefront-developers.html 2

http://www.lindachristas.org/education/ex-president-nelson-mandela-offered-omar-dullah-chair-of-freedom-linda-christas-college/ 1

http://www.livesexcalls.co.uk/recruitment.html 1

http://www.livesexcalls.co.uk/sexgirls.html 1

http://www.maddenjinx.com/hgh-supplements-acting-as-energizer-for-the-body/ 1

http://www.meatgasm.com/how-to-cure-for-skin-yeast-infection/ 1

http://www.nswcashregisters.com.au/nexa-printer-pos-bundle.html 2

http://www.palmbeachvirtualoffice.com/services.php 2

http://www.palmphone.com/ 1

http://www.pinkelectricguitar.org/ 2

http://www.preferredlipodocsofedmonton.com/ 1

http://www.prlog.org/10401875-playa-del-carmen-condo-rentals-by-condohotelsplayadelcarmencom.html 2

http://www.registrycleanergeeks.com/error-doctor/review/ 1

http://www.saeshaflowers.com/send-flowers-to.aspx/Indonesia/0/0/0/0 1

http://www.sameshow.com/ppt2flash/article/tips-to-create-thanksgiving-ecard.html 1

http://www.santabarbaraminister.com/ 2

http://www.sblaw.vn/ 1

http://www.sell-sport.com/adam-vinatieri-blue-nfl-jersey4-p-317.html 1

http://www.sell-sport.com/anthony-gonzalez-blue-team-color-nfl-jerseyindianapolis-colts-11-p-1125.html 1

http://www.sell-sport.com/anthony-munoz-premier-throwback-nfl-jerseybengals-78-p-184.html 2

http://www.sell-sport.com/brandon-marshall-navy-nfl-jersey15-p-263.html 1

http://www.sell-sport.com/brandon-marshall-orange-nfl-jersey15-p-261.html 1

http://www.sell-sport.com/brandon-marshall-white-jersey15-p-262.html 1

http://www.sell-sport.com/carson-palmer-black-replia-jersey9-p-176.html 1

http://www.sell-sport.com/carson-palmer-white-jersey9-p-177.html 1

http://www.sell-sport.com/chad-johnson-orange-alternate-nfl-jerseyscincinnati-bengals-85-p-767.html 1

http://www.sell-sport.com/chad-johnson-white-jersey85-p-178.html 1

http://www.sell-sport.com/champ-bailey-navy-nfl-jersey24-p-279.html 1

http://www.sell-sport.com/champ-bailey-orange-alternate-jersey24-p-278.html 1

http://www.sell-sport.com/champ-bailey-white-nfl-jersey24-p-272.html 1

http://www.sell-sport.com/eddie-royal-navy-nfl-jersey19-p-280.html 1

http://www.sell-sport.com/john-elway-premier-throwback-alternate-nfl-jerseydenver-broncos-p-265.html 1

http://www.sell-sport.com/john-elway-throwback-navy-team-color-nfl-jerseydenver-broncos-7-p-775.html 1

http://www.sell-sport.com/mitchell-and-ness-carson-palmer-throwback-orange-nfl-jerseycincinnati-bengals-9-p-1228.html 1

http://www.sell-sport.com/mitchell-ness-tjhoushmandzadeh-throwback-black-nfl-jerseybengals-84-p-1230.html 1

http://www.sell-sport.com/rudi-johnson-white-jersey32-p-173.html 1

http://www.sell-sport.com/steve-atwater-throwback-navy-team-color-nfl-jerseydenver-broncos-27-p-1041.html 1

http://www.sell-sport.com/terrell-davis-throwback-navy-team-color-nfl-jerseydenver-broncos-30-p-949.html 1

http://www.sell-sport.com/tj-houshmandzadeh-orange-jersey84-p-170.html 1

http://www.semaphore-software.com/software-solution/mobile-website-development.htm 1

http://www.settlementagentsperth.com/ 1

http://www.southeastmnfoodnetwork.com/cure-for-yeast-infection-the-best-ways/ 1

http://www.spiritvoyage.com/yogablog/index.php/2009/11/05/meditation-music-for-kundalini-yoga/ 1

http://www.stacey-solomon-music.com/ 1

http://www.stacey-solomon-music.com/?p=135 1

http://www.superbass.com/what-makes-a-colon-effective/ 1

http://www.teepack.com/en/zenobia.htm 1

http://www.thefearlessfactor.com/ 1

http://www.towelmania.com/embato.html 2

http://www.usedengine.org/ 2

http://www.usedengines.me/ 1

http://www.vimeo.com/7470270 1

http://www.weaversonline.org/ 1

http://www.webmarketingofnorthflorida.com/Jacksonville\_InternetMarketing.php 1

http://www.weddingtransport.net/ 1

http://www.yogasuppliesonline.com/yoga\_information/yogaawareness.html 2