Open Refine Instructions

Name Disambiguation and Normalization

OpenRefine - http://openrefine.org/

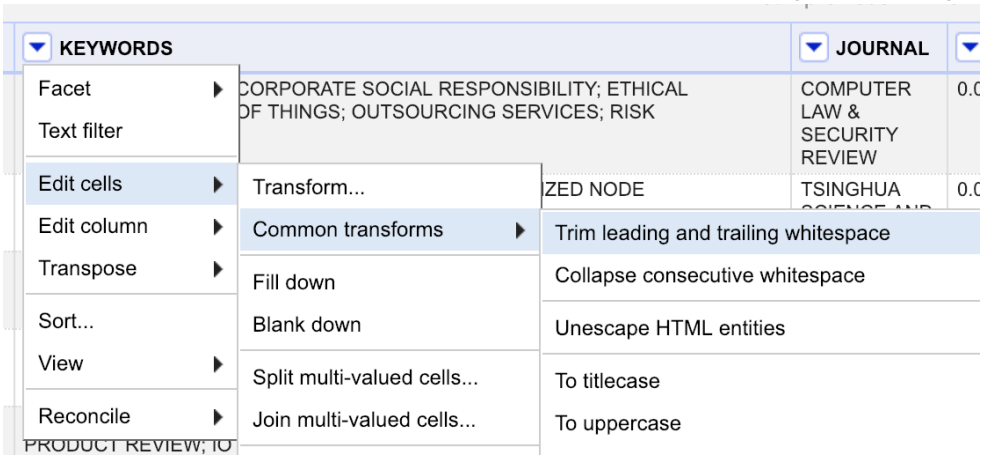
Demo how to use crane data (ai, iot, robotics) to normalize keywords
Data: https://iu.box.com/s/cd0n4wlqw5z35mvg3decgwvyhriei7cy
Goal: Pre-process data prior feeding it to burst

Step 1: Importing Data

- Choose a file (txt, csv)
- Parse data (comma or tab delimited)
- Create project
- Check the number of rows **11371 rows** [example using iot data]

Step 2: Cleaning

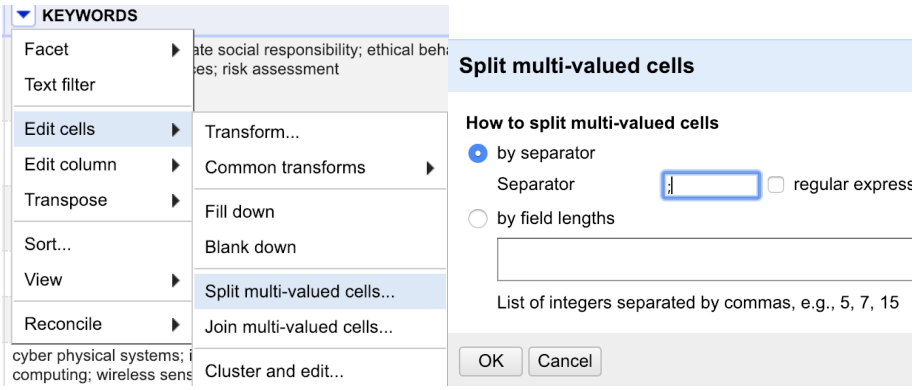
- Select a dropdown arrow in the column KEYWORDS
- Do the following as needed:
    - Edit cells > Common transforms > Trim leading and trailing spaces



    - Edit cells > Common transforms > Collapse whitespaces
    - Edit cells > Common transforms > To lower cases

Step 3: Splitting Cells

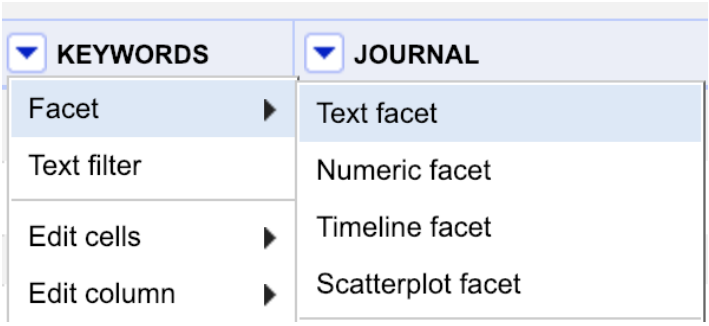
- Split Keywords items by a separator



- The number of row should be increased

Step 4: Text Facet

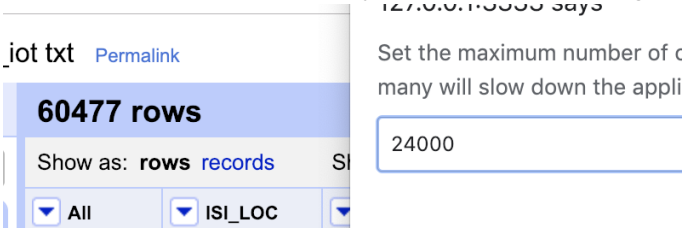
- Dropdown arrow on keywords: Select Facet > Text Facet



- Note: If you have the following message on the left top panel:

23813 choices total, too many to display
Set choice count limit

Select facet choice and set your number to be higher than the current number of rows



- Review how many unique choices you have [example - 23813 choices]

Step 5: Clustering

- Select cluster option [top right corner in your text facet]

york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel

Method key collision Keying Function fingerprint

- Keep default: Method - key collision and fingerpring [more details - https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth]

- Make a note of how many clusters are found [top right] - e.g. 1904
- Review clusters and naming suggestion

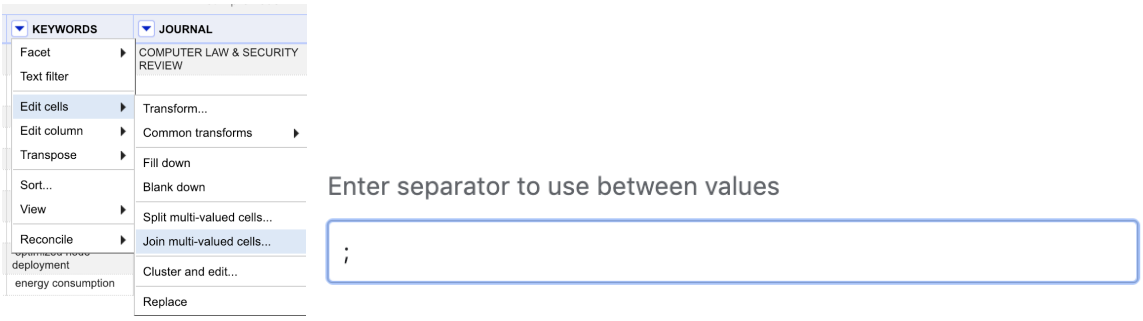
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
10	1294	• internet of things (iot) (745 rows) • internet of things (iot) (493 rows) • iot (internet of things) (30 rows) • iot (internet of things) (17 rows) • internet of things iot (3 rows) • iot - internet of things (2 rows) • internet of things (iot) (1 rows) • internet of things - iot (1 rows) • internet of things (iot)) (1 rows) • iot - internet of things (1 rows)	☐	internet of things (iot)

- If agree with the choices, select Merge
- After reviewing all clusters, select merge and re-cluster, as new clusters might be identified after the merger

- If no more clusters are found, close the cluster window. Wait until the facet gets refreshed, it might take a bit

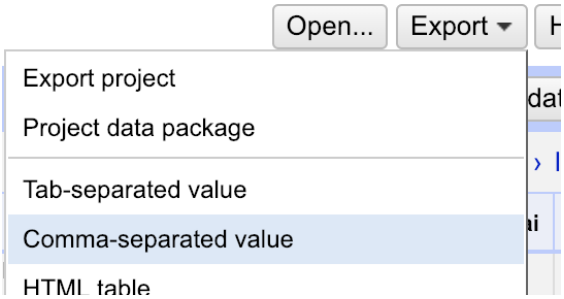
Step 6: Joining Cells

- Select Edit cells > Join multi-valued cells
- Indicate the separator as **;**



- Check the number of rows: **11371 rows** [example using iot data]
- If needed, you can revert all keywords to upper cases or initial cases only: Edit cells>common transform>To titlecases

Step 7: Exporting



Repeat for AI, ROBOTICS, and IOT