tech_review

November 15, 2020

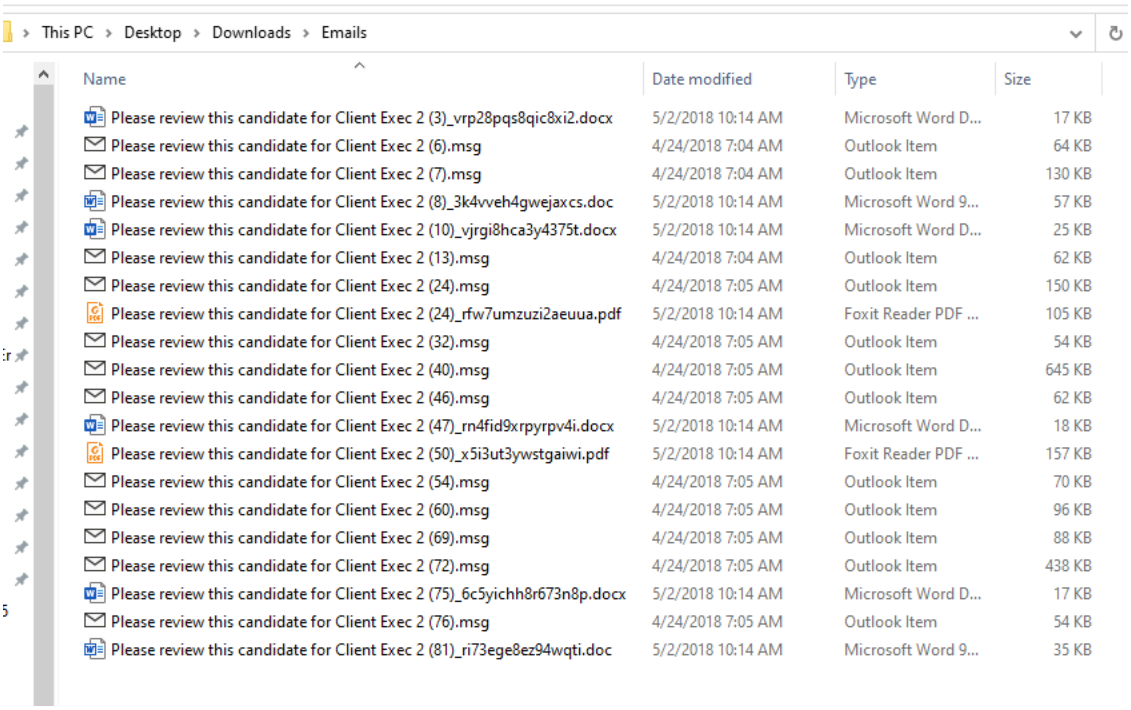
1 TF/IDF with KNIME

In this tech review, I am going to show how to compute TF/IDF with KNIME. KNIME is a drag-n-drop advanced analytics platform. More at KNIME.com

KNIME Desktop is a free software. It can do all of the analytics that one might need to do.

In this demo, I am going to show how to parse documents (resumes) and compute TF/IDF.

I have a folder with resumes in it that looks like this:

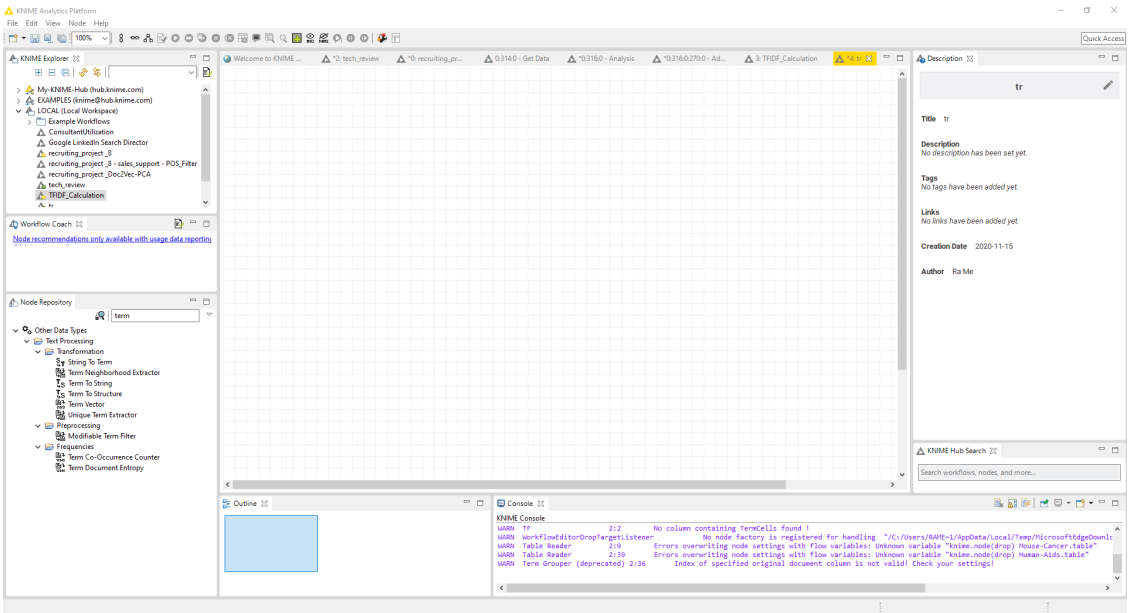


This PC > Desktop > Downloads > Emails

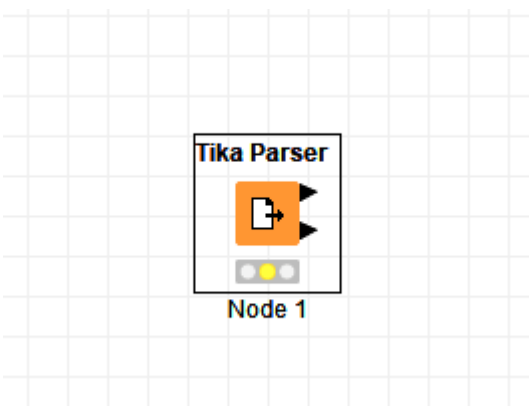
Name	Date modified	Type	Size
Please review this candidate for Client Exec 2 (3)_vrp28pqs8qic8xi2.docx	5/2/2018 10:14 AM	Microsoft Word D...	17 KB
Please review this candidate for Client Exec 2 (6).msg	4/24/2018 7:04 AM	Outlook Item	64 KB
Please review this candidate for Client Exec 2 (7).msg	4/24/2018 7:04 AM	Outlook Item	130 KB
Please review this candidate for Client Exec 2 (8)_3k4vveh4gwejaxcs.doc	5/2/2018 10:14 AM	Microsoft Word 9...	57 KB
Please review this candidate for Client Exec 2 (10)_vjrgi8hca3y4375t.docx	5/2/2018 10:14 AM	Microsoft Word D...	25 KB
Please review this candidate for Client Exec 2 (13).msg	4/24/2018 7:04 AM	Outlook Item	62 KB
Please review this candidate for Client Exec 2 (24).msg	4/24/2018 7:05 AM	Outlook Item	150 KB
Please review this candidate for Client Exec 2 (24)_rfw7umzuzi2aeuua.pdf	5/2/2018 10:14 AM	Foxit Reader PDF ...	105 KB
Please review this candidate for Client Exec 2 (32).msg	4/24/2018 7:05 AM	Outlook Item	54 KB
Please review this candidate for Client Exec 2 (40).msg	4/24/2018 7:05 AM	Outlook Item	645 KB
Please review this candidate for Client Exec 2 (46).msg	4/24/2018 7:05 AM	Outlook Item	62 KB
Please review this candidate for Client Exec 2 (47)_rn4fid9xrpyrpv4i.docx	5/2/2018 10:14 AM	Microsoft Word D...	18 KB
Please review this candidate for Client Exec 2 (50)_x5i3ut3ywstgaiwi.pdf	5/2/2018 10:14 AM	Foxit Reader PDF ...	157 KB
Please review this candidate for Client Exec 2 (54).msg	4/24/2018 7:05 AM	Outlook Item	70 KB
Please review this candidate for Client Exec 2 (60).msg	4/24/2018 7:05 AM	Outlook Item	96 KB
Please review this candidate for Client Exec 2 (69).msg	4/24/2018 7:05 AM	Outlook Item	88 KB
Please review this candidate for Client Exec 2 (72).msg	4/24/2018 7:05 AM	Outlook Item	438 KB
Please review this candidate for Client Exec 2 (75)_6c5yichh8r673n8p.docx	5/2/2018 10:14 AM	Microsoft Word D...	17 KB
Please review this candidate for Client Exec 2 (76).msg	4/24/2018 7:05 AM	Outlook Item	54 KB
Please review this candidate for Client Exec 2 (81)_ri73ege8ez94wqti.doc	5/2/2018 10:14 AM	Microsoft Word 9...	35 KB

You can see that we have 20 documents: `.MSG`, `.DOC`, `.DOCX`, and `.PDF`.

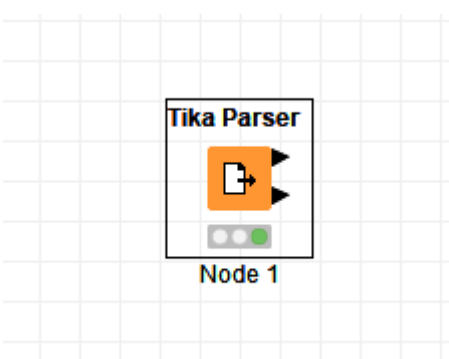
First, we start with a blank workbench.



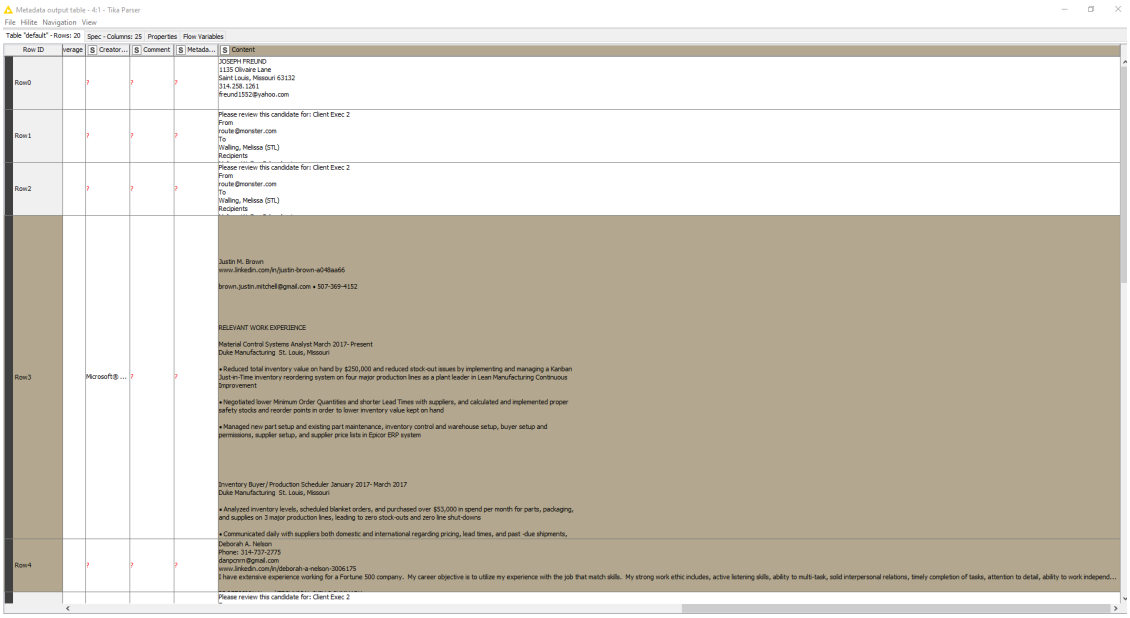
We want to parse all of our documents at once. In KNIME, there is a processing node called `Tika Parser`. We can drop it into our workbench.



The next step is to specify where the documents are. For that, we right-click, click `Configure`, then click on `Browse`, then navigate to the folder where the documents are and click `OK`. After that, right-click on the node and click execute. You will see that the yellow collor under the node will turn to green.

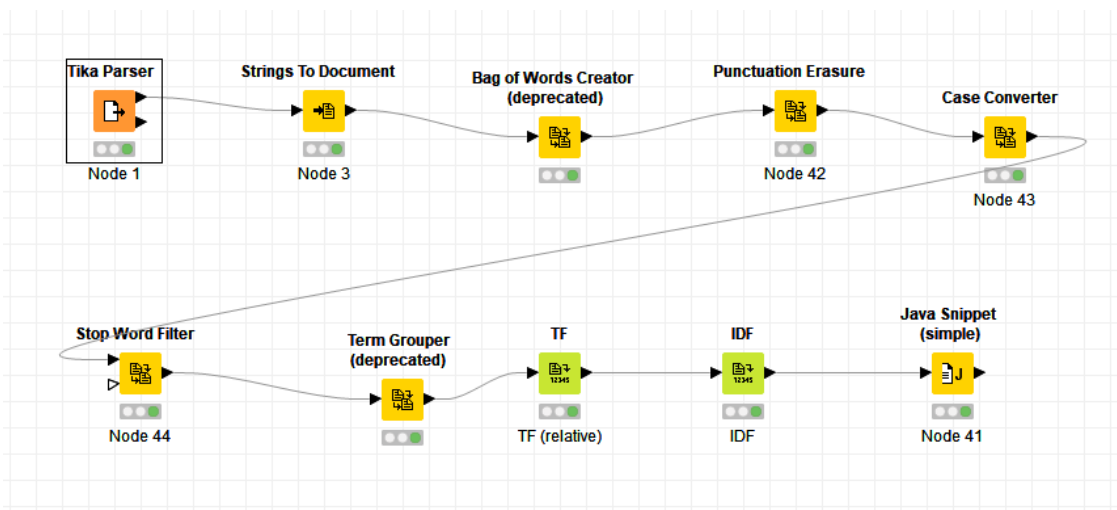


If we'll right-click on the node again, and if we'll click on `Metadata output table`, we'll see a table with parsed data. The last column, `Content`, contains the actual document data.



You can see that each cell in the `Content` column is a blob of text that contains resume data.

Here is the final workflow so that we can see what we are trying to accomplish:



From the `Tika Parser` node, we are going to the `Strings To Document` node. This is needed because we have to feed a document to the `BoW` node. This is the way KNIME works. When we right-click on the `BoW` node and click on `Documents Output Table`, we can see that our row count is now in thousands because for each term in a document we have a row.

· Rows: 8920 Spec - Columns: 2 Prop

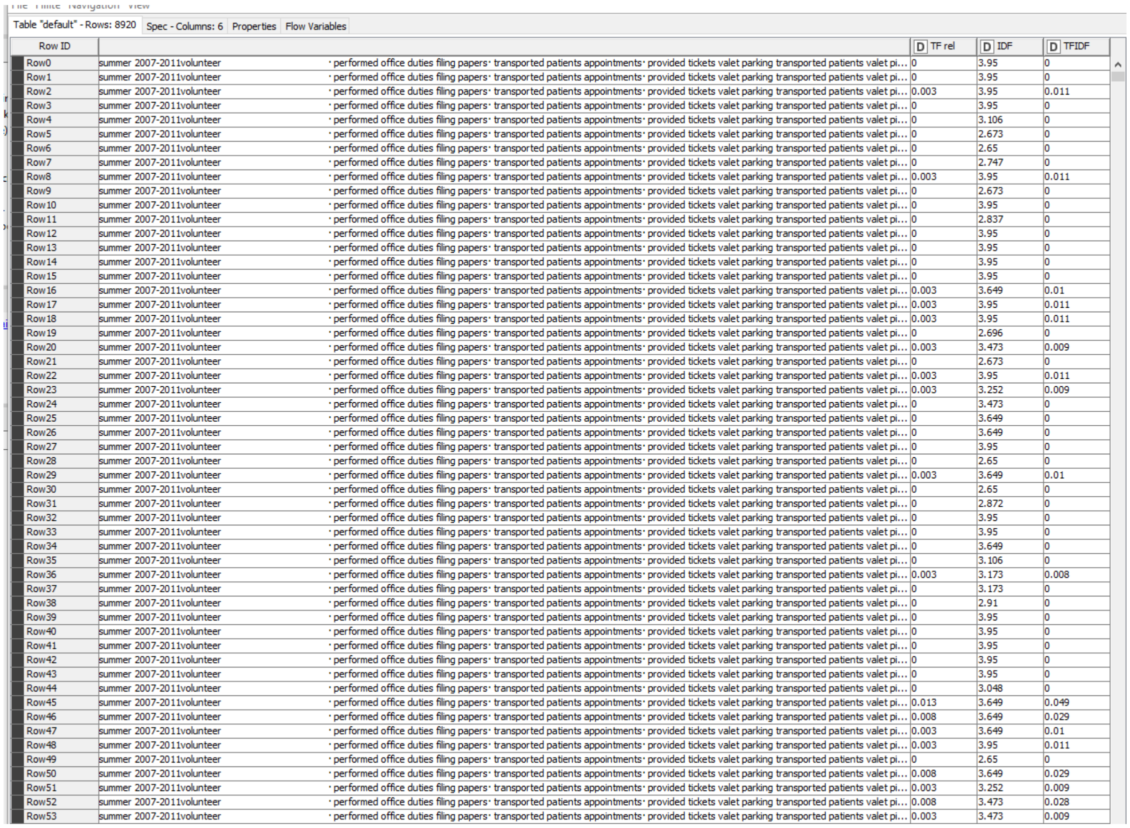
Term
finalist[]
TRULASKE[]
COLLEGE[]
Columbia[]
1999-2004[]
Bachelor[]
Science[]
Major[]
:[]
Banking[]
Financial[]
Management[]
Association[]
Vice-President[]
Treasurer[]
Student[]
Council[]
Phi[]
Kappa[]
Fraternity[]
Rush[]
Chairman[]
EXPERIENCE[]
Solid[]
Gold[]
Pet[]
2017-2017[]
Senior[]
Analyst[]
Developed[]
sales[]
marketing[]
analysis[]
for[]
our[]
domestic[]
international[]

We then use the nodes `Punctuation Erasure`, `Case Converter`, and `Stop Word Filter` to do what we have learn in our lectures when we were working with tokenizers.

The next node is a `Term Grouper` node. This node groups all terms of a document by their text and deletes all tags.

The final 3 nodes are `TF`, `IDF`, and `Java Snippet`. All 3 nodes compute a column for our table. The last node computes a product of `TF` * `IDF`.

When we right-click on the `Java Snippet` node and click on `Appended Table`, we see the following table:



Row ID		TF rel	IDF	TFIDF
Row0	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.95	0
Row1	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.95	0
Row2	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0.003	3.95	0.011
Row3	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.95	0
Row4	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.106	0
Row5	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	2.673	0
Row6	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	2.65	0
Row7	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	2.747	0
Row8	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0.003	3.95	0.011
Row9	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	2.673	0
Row10	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.95	0
Row11	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	2.837	0
Row12	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.95	0
Row13	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.95	0
Row14	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.95	0
Row15	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.95	0
Row16	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0.003	3.649	0.01
Row17	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0.003	3.95	0.011
Row18	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0.003	3.95	0.011
Row19	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	2.696	0
Row20	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0.003	3.473	0.009
Row21	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	2.673	0
Row22	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0.003	3.95	0.011
Row23	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0.003	3.252	0.009
Row24	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.473	0
Row25	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.649	0
Row26	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.649	0
Row27	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.95	0
Row28	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	2.65	0
Row29	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0.003	3.649	0.01
Row30	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	2.65	0
Row31	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	2.872	0
Row32	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.95	0
Row33	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.95	0
Row34	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.649	0
Row35	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.106	0
Row36	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0.003	3.173	0.008
Row37	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.173	0
Row38	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	2.91	0
Row39	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.95	0
Row40	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.95	0
Row41	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.95	0
Row42	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.95	0
Row43	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.95	0
Row44	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	3.048	0
Row45	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0.013	3.649	0.049
Row46	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0.008	3.649	0.029
Row47	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0.003	3.649	0.01
Row48	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0.003	3.95	0.011
Row49	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0	2.65	0
Row50	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0.008	3.649	0.029
Row51	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0.003	3.252	0.009
Row52	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0.008	3.473	0.028
Row53	summer 2007-2011volunteer · performed office duties filing papers· transported patients appointments· provided tickets valet parking transported patients valet pi…	0.003	3.473	0.009

We can see that KNIME has computed what we needed.

This concludes our demo.