Directions for RapidMiner and Enron Analysis

RapidMiner is a free textual analysis software that combines data preparation, machine learning, and predictive model deployment. This guide was developed to assist students in navigating its basic functions and included the following steps:

Step 1   Download RapidMiner
Step 2   Open RapidMiner and Acquire Extensions
Step 3   Add AYLIEN Text Analysis Extension
Step 4   Download Data (in this case 10-Ks)
Step 5   Organize Data into Folders
Step 6   Process 10-Ks
Step 7   Find List of Words
Step 8   Modify List of Words Results
Step 9   Find Passive Verbs
Step 10   Find Uncertainty Words using Loughran-McDonald Sentiment Word Lists
Step 11   Analyze Sentiment

Appendix   Solutions – Mattel vs Hasbro using 2016 10-Ks

Step 1: Download RapidMiner.

1. Go to www.rapidminer.com and select download in the right hand corner. (Download the appropriate version for your computer)
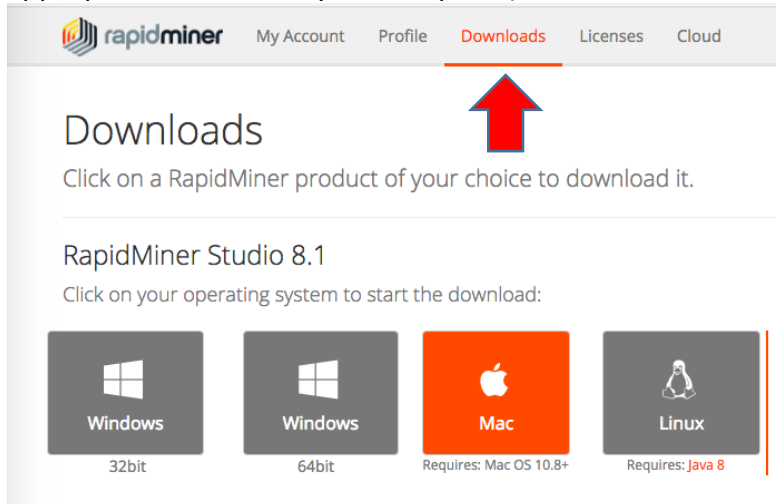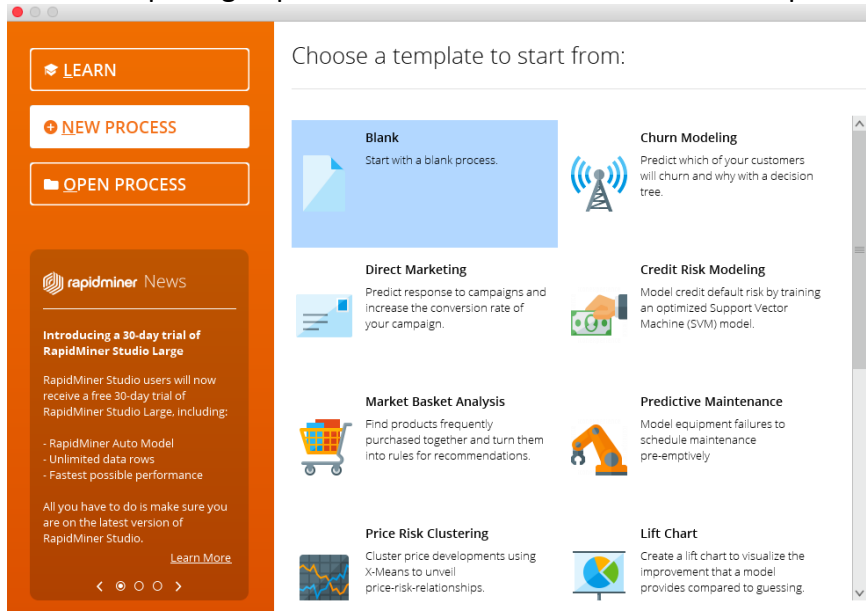

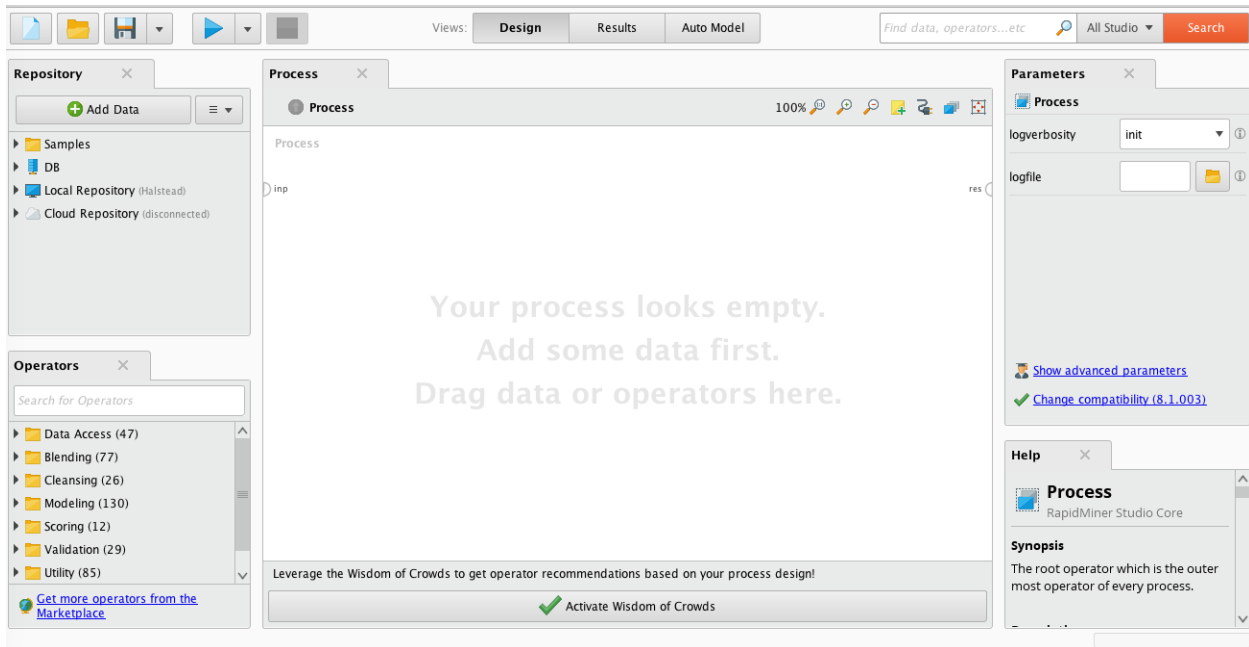2. Once downloaded run Install to add RapidMiner to your computer.

3. RapidMiner ask you to confirm your account via email. (Note, it may take RapidMiner a few minutes to generate this email.)

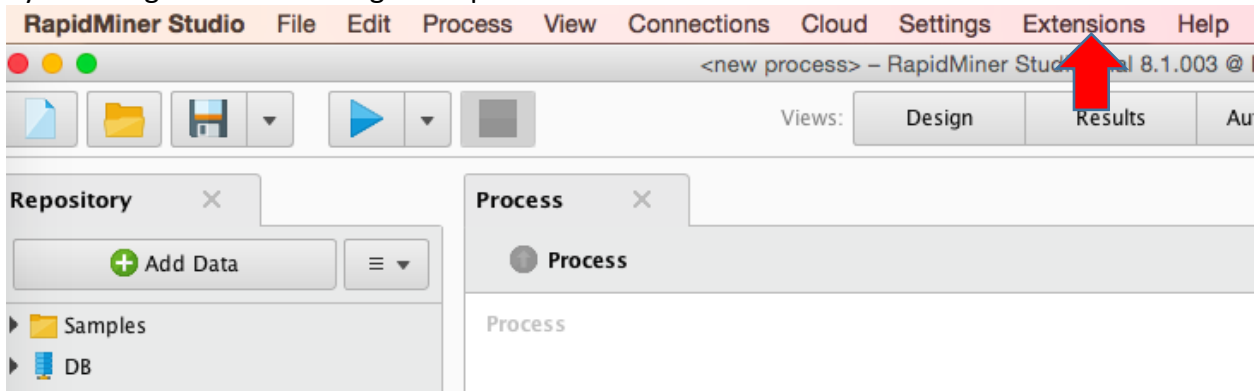**Step 2: Open RapidMiner and Acquire Appropriate Extensions.**

1. When opening RapidMiner be sure to select a new blank process.
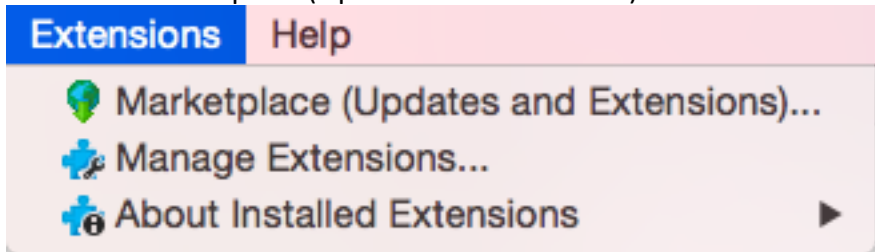

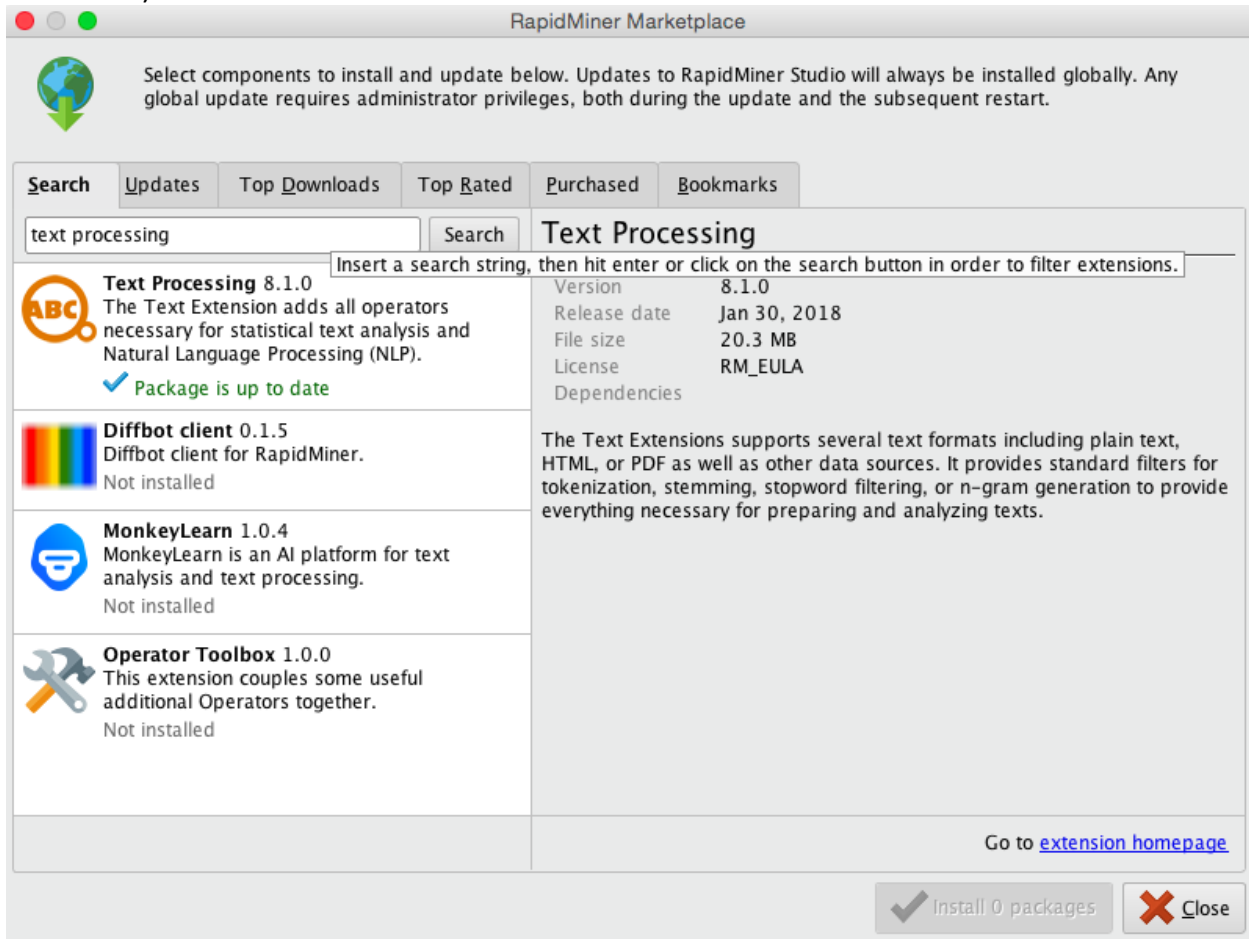2. Your home page of RapidMiner should look like this:

3. To add the appropriate extensions to analyze text, download the "text processor" extension by selecting extensions along the top border.

4. Select "Marketplace (Updates and Extensions).

5. Search "text processing" in the search tab and click the install button. (Accept the terms and conditions)

6. After installing the extension from the market place, RapidMiner will restart.

Repeat the steps above to download the extension "Text Analysis by AYLEIN" to RapidMiner.

Step 3: Add AYLIEN Text Analysis.

AYLIEN Text Analysis needs to be added separately since is it not developed by RapidMiner.

1. Find "AYLIEN Text Analysis" under the extensions in the Operators section.
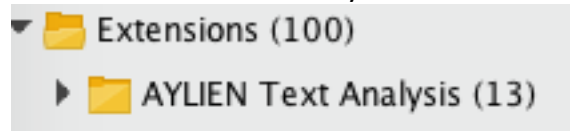
2. Double Click on AYLIEN Text Analysis until you get to the Analyze Sentiment Operator and drag it into the process area.

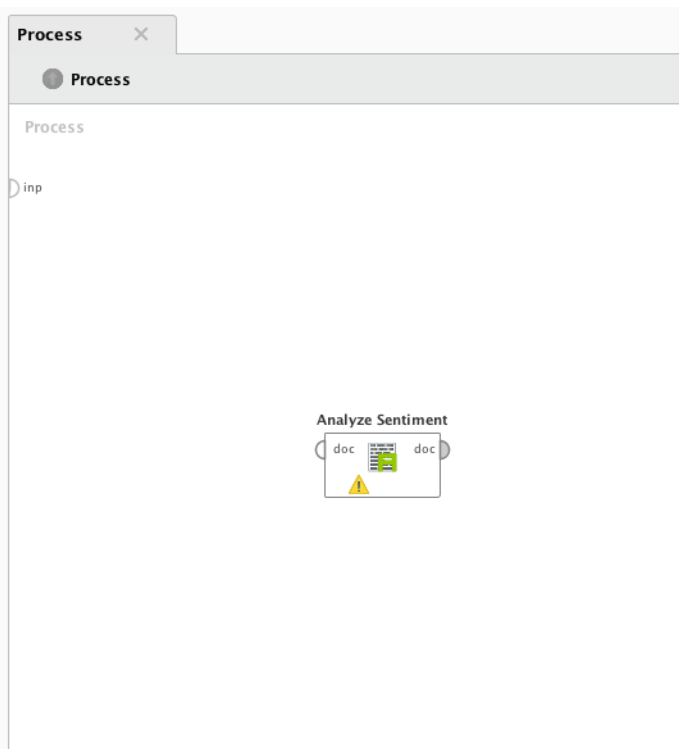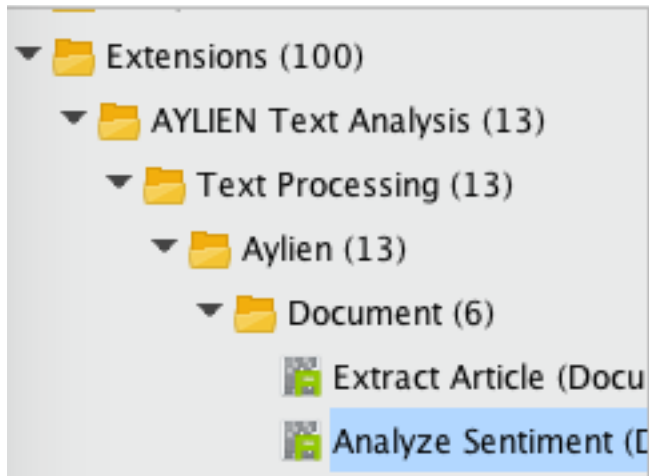3. Click on the "Analyze Sentiment". The parameters screen will appear on the right side of the screen.
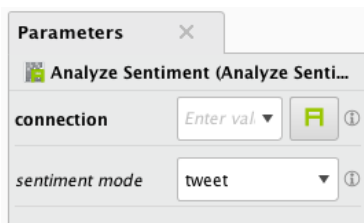
4. Since this is outside of RapidMiner, we need to connect to this extension to use it. Click on the "info" icon to the right of the A in the connection tab. Read the description and then click

on the blue "here" to get access your connection key. If this does not work go to
https://developer.aylien.com/ and click sign up.
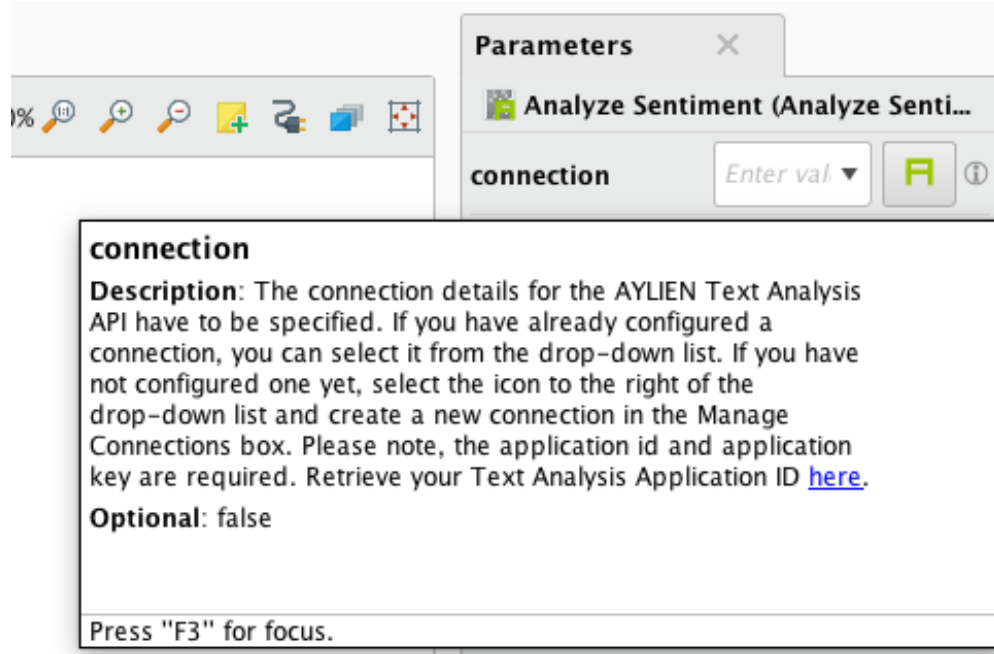
5. Fill this form out to sign up. (for organization type in "Iowa State University")

6. After you select sign up, the email you used will be sent an activation link to the account. Click the link:

Thanks for signing up to use our Text Analysis API. You're one step away from getting full access to our API on our Basic the link below.

Username: nphalstead1

Activate your account: https://developer.aylien.com/activate/69a2437e39a0357a67abbb71dd92a823a71c2e6f
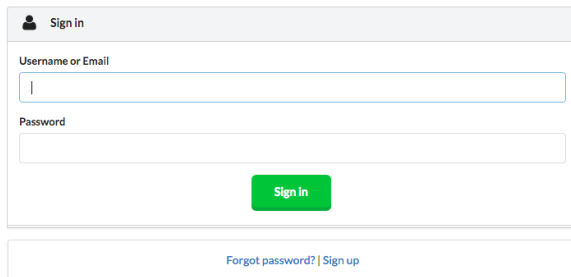
If you have problems activating your account please contact support support@aylien.com.

Onward,

AYLIEN Team
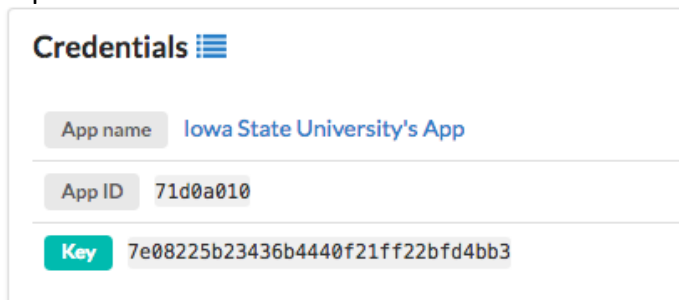
7. Enter your username and password that you just created.

_____

Sign in

Username or Email

Password

Sign in

Forgot password? | Sign up

8. Your APP ID and Key will show up. These are the connections we are looking for to put into RapidMiner.

Credentials ☰

App name   Iowa State University's App

App ID   71d0a010

Key   7e08225b23436b4440f21ff22bfd4bb3

9. Go back into RapidMiner and hit the A next to connection in the parameters for the operator we had selected earlier.
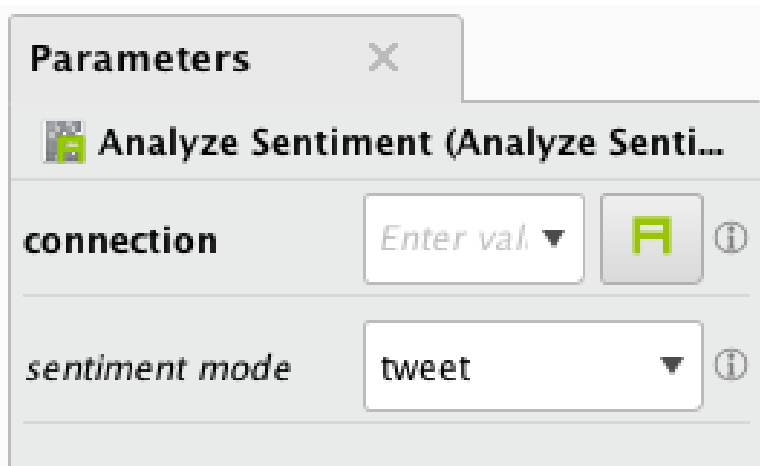
Type in the codes that you just received from the earlier steps. (You can click the pencil to rename this connection, I will rename it Text Analytics right now)
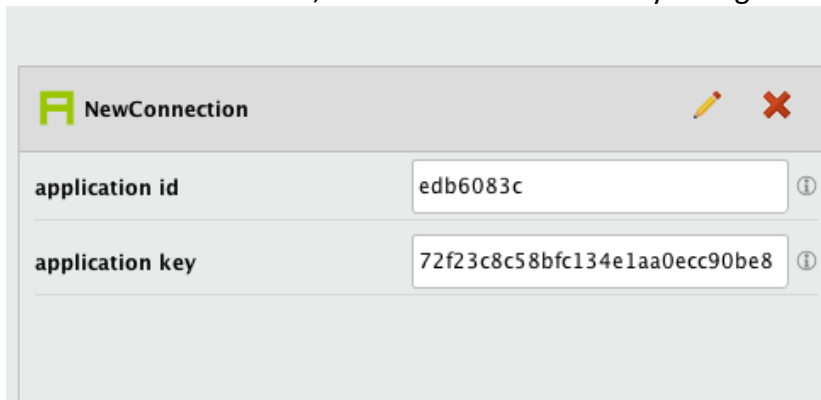
10. Hit "Apply". We will use this connection later in the assignment.

Step 4: Download Data (10-Ks).

1. For this assignment you were given three different companies, Enron Corp, CMS Energy, and FirstEnergy Corp. We will analyze these 10-K's that were released just prior to Enron's collapse to see if textual analysis could have predicted Enron's problems.

2. To start, locate each company's 10-K and save as a text file to your computer.

3. To start this process go to http://rankandfiled.com/ The home page should look like this:
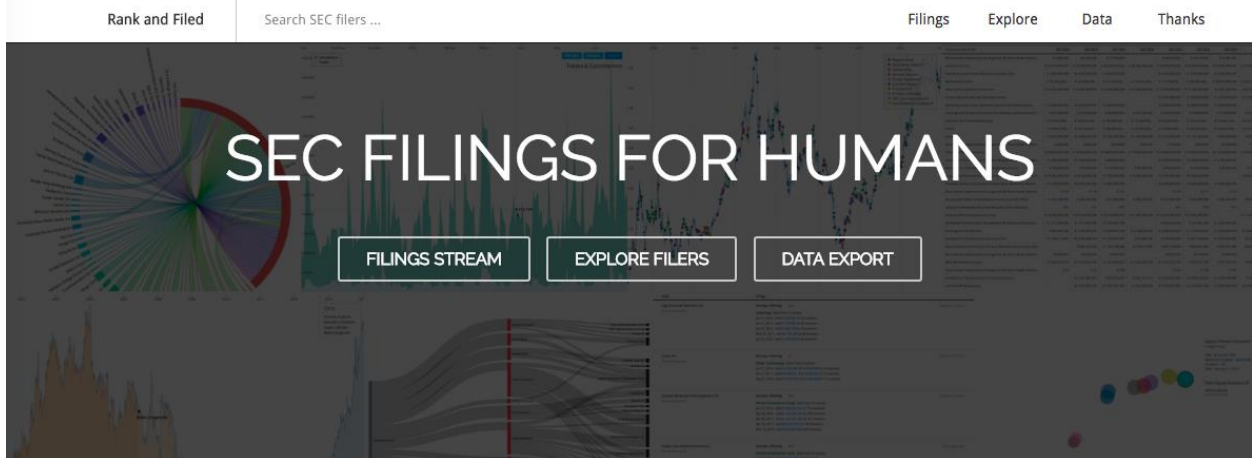
4. In the "Search SEC Filers.." type in Enron Corp and select the top result that you see below.
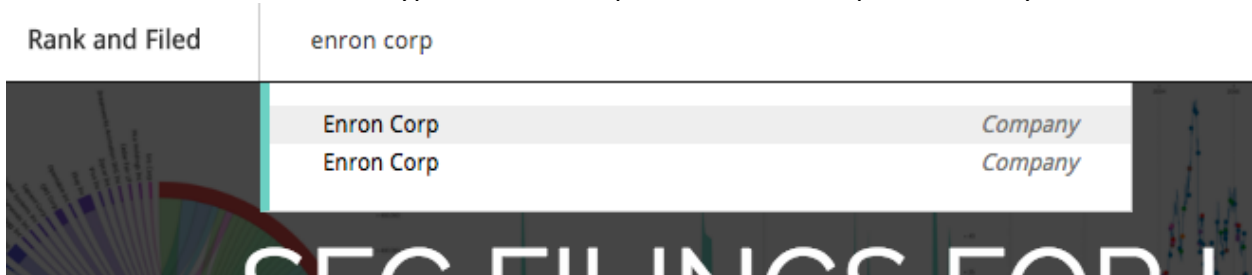
5. When you select Enron Corp this page should appear. Then select the table option and find the "Annual Report 10-K" for the year 2001.
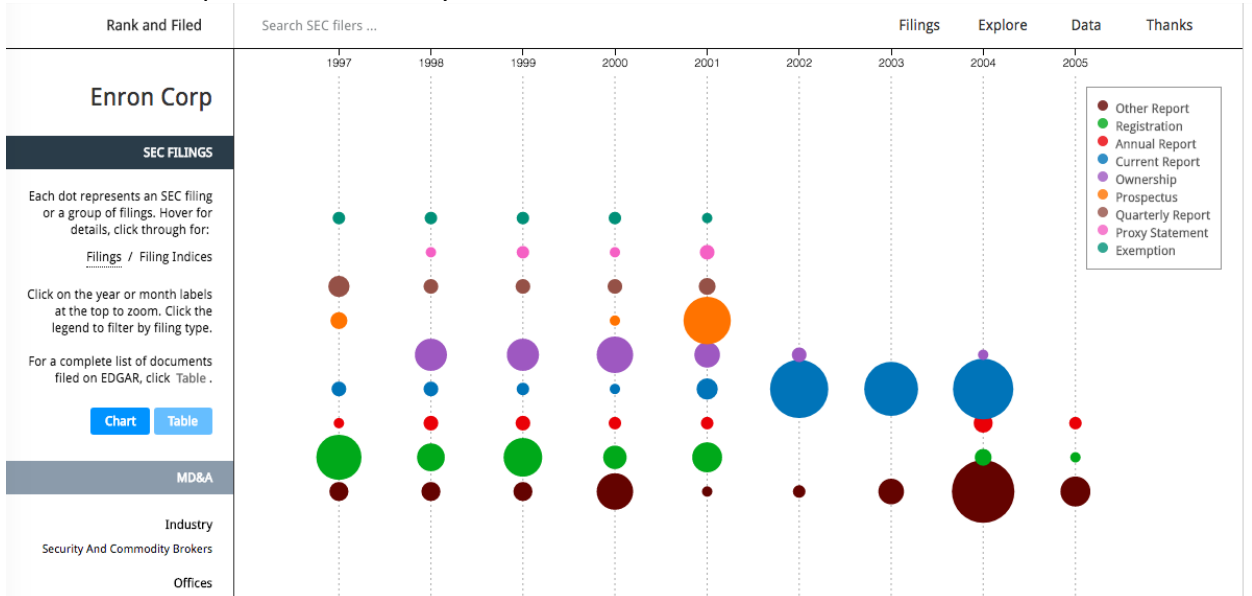
6. Below is what you will see in the table format. Once you find the report option that we see below, click on 10-K Annual Report (also where the page is highlighted).

			Statement Regarding Computation Of Ratios
Apr 2, 2001	10-K	Annual Report	Annual Report
			Subsidiaries Of The Registrant
			Statement Regarding Computation Of Ratios
			Power Of Attorney
			Material Contracts
			Material Contracts
			Financial Statements Of Atlantic Water Trust
			Consents Of Experts And Counsel
			Consents Of Experts And Counsel
Mar 27, 2001	DEF 14A	Proxy Statement	Definitive Proxy Statement
Mar 2, 2001	PRE 14A	Proxy Statement	Preliminary Proxy Statement
Feb 28, 2001	8-K	Current Report	Current Report

7. A document that looks like the one below should appear.

```
-----BEGIN PRIVACY-ENHANCED MESSAGE-----
Proc-Type: 2001,MIC-CLEAR
Originator-Name: webmaster@www.sec.gov
Originator-Key-Asymmetric:
 MFgwCgYEVQgBAQICAf8DSgAwRwJAW2sNKK9AVtBzYZmr6aGjlWyK3XmZv3dTINen
 TWSM7vrzLADbmYQaionwg5sDW3P6oaM5D3tdezXMm7z1T+B+twIDAQAB
MIC-Info: RSA-MD5,RSA,
 WVMqBDIh67AGAZHqEq9b+2C7mE5a9DiI4NiuXU+j8/U7MY4nL/lY9LZpo4elsZ/5
 uYSmX4MYDqjrU1O7b8UMYw==

<SEC-DOCUMENT>0001024401-01-500010.txt : 20010409
<SEC-HEADER>0001024401-01-500010.hdr.sgml : 20010409
ACCESSION NUMBER:             0001024401-01-500010
CONFORMED SUBMISSION TYPE:    10-K
PUBLIC DOCUMENT COUNT:        9
CONFORMED PERIOD OF REPORT:   20001231
FILED AS OF DATE:             20010402

FILER:

        COMPANY DATA:
                COMPANY CONFORMED NAME:            ENRON CORP/OR/
                CENTRAL INDEX KEY:                 0001024401
                STANDARD INDUSTRIAL CLASSIFICATION: SECURITY BROKERS, DEALERS & FLOTATION COMPANIES [6211]
                IRS NUMBER:                        470255140
                STATE OF INCORPORATION:            OR
                FISCAL YEAR END:                   1231

        FILING VALUES:
                FORM TYPE:         10-K
                SEC ACT:
                SEC FILE NUMBER:   001-13159
                FILM NUMBER:       1588971

        BUSINESS ADDRESS:
                STREET 1:          1400 SMITH ST
                CITY:              HOUSTON
                STATE:             TX
```

8. In the top part of your browser, select "File", then "Save-as." Save file as a text file on your desktop for now. Go ahead and rename it Enron 2001 10-K so that it is appropriately labeled.
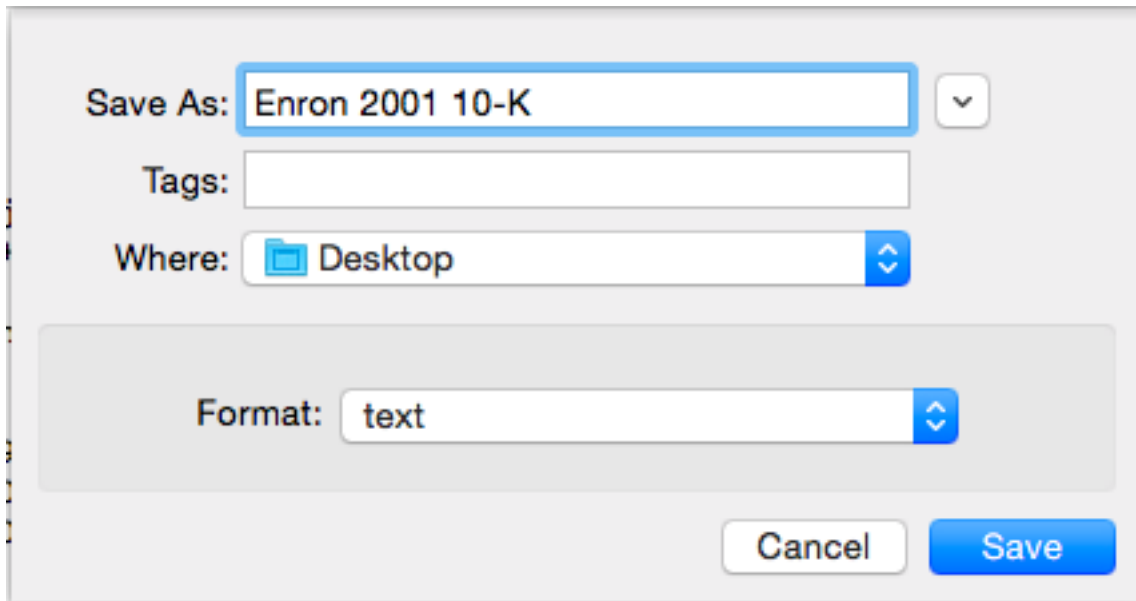
9. Repeat the above steps with the other two companies. Be sure to find the 2001 10-Ks to compare consistently with all your companies.

(Note: Any company that you were interested in comparing could be downloaded as text in the same steps that we see above.)

Step 5: Organize Folders.

1. To process data in Rapid Miner, create a folder on your desktop for this assignment. Name this folder Text Analytics.
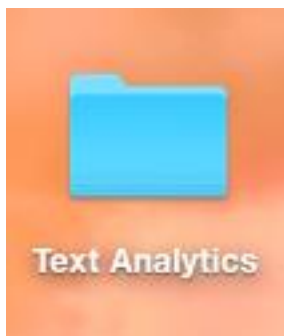
2. Inside the folder, create a folder for each company we are analyzing as shown below:
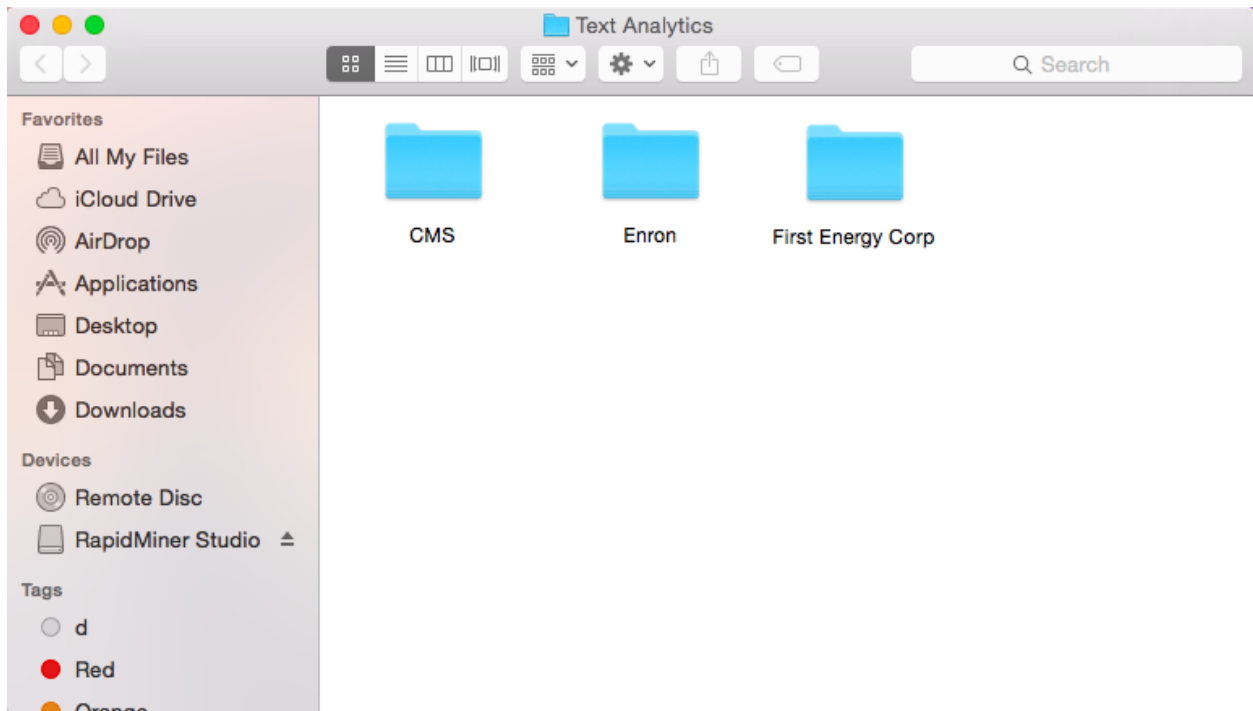
3. Inside the apprpriate folder, place the text document 10-K that we have saved from the earlier steps. For example in the CMS Energy Folder it would look like:
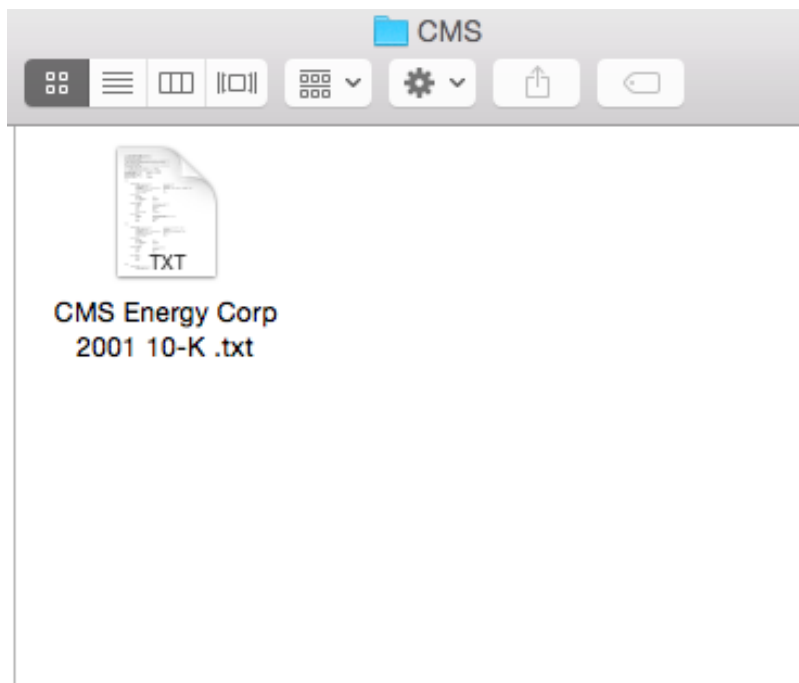
Step 6: Process Data in RapidMiner.

1. In RapidMiner in the lower left hand corner where we see the operators tab, type in "process documents from files."
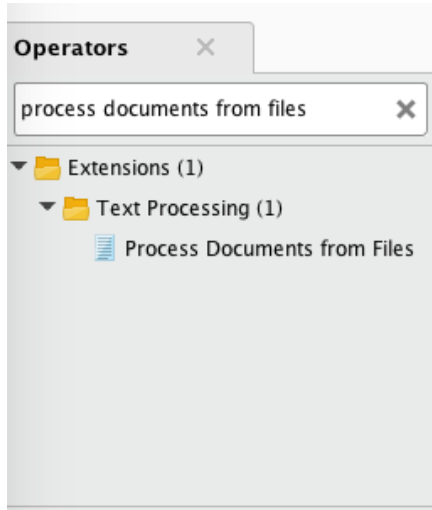
2. Drag the "process documents from files" into the process tab. Your screen should look like the one below:
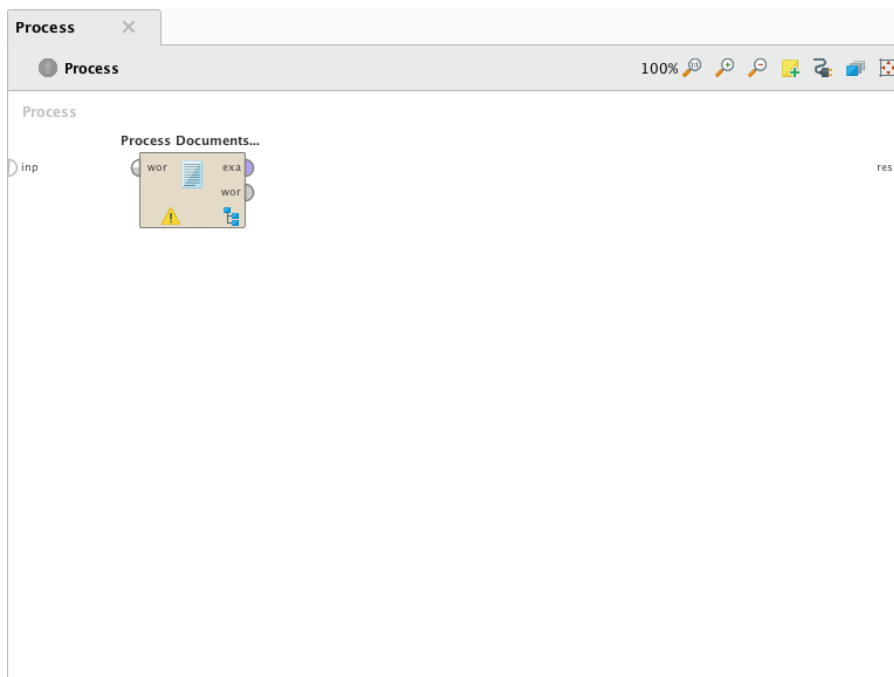
3. Right click on the process documents from file in the process tab and the following parameters will appear on the right side of your screen.
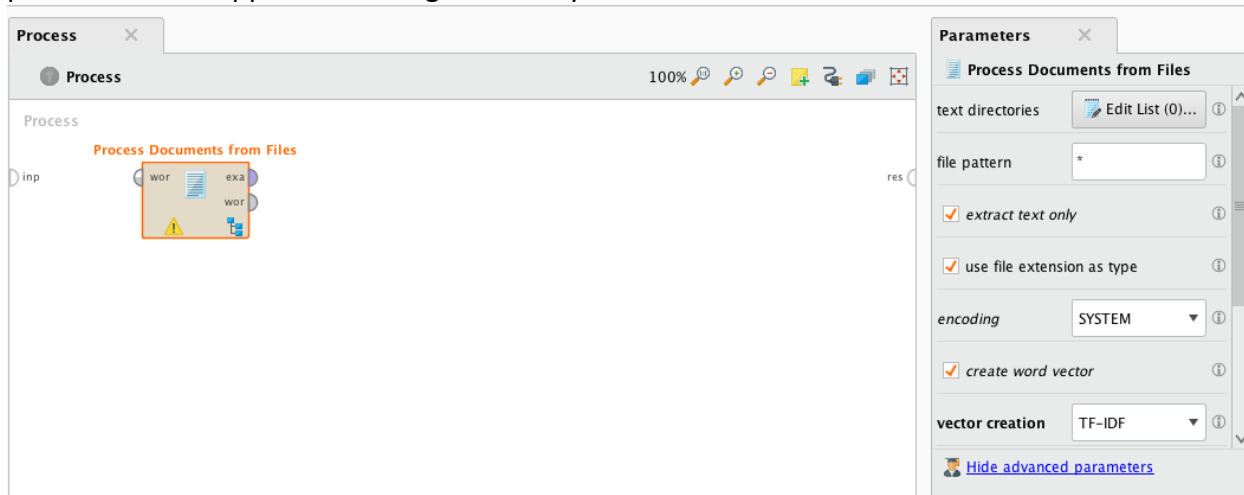
4. Select the "edit list" to the right of text directories in the parameters section. The following screen will appear:
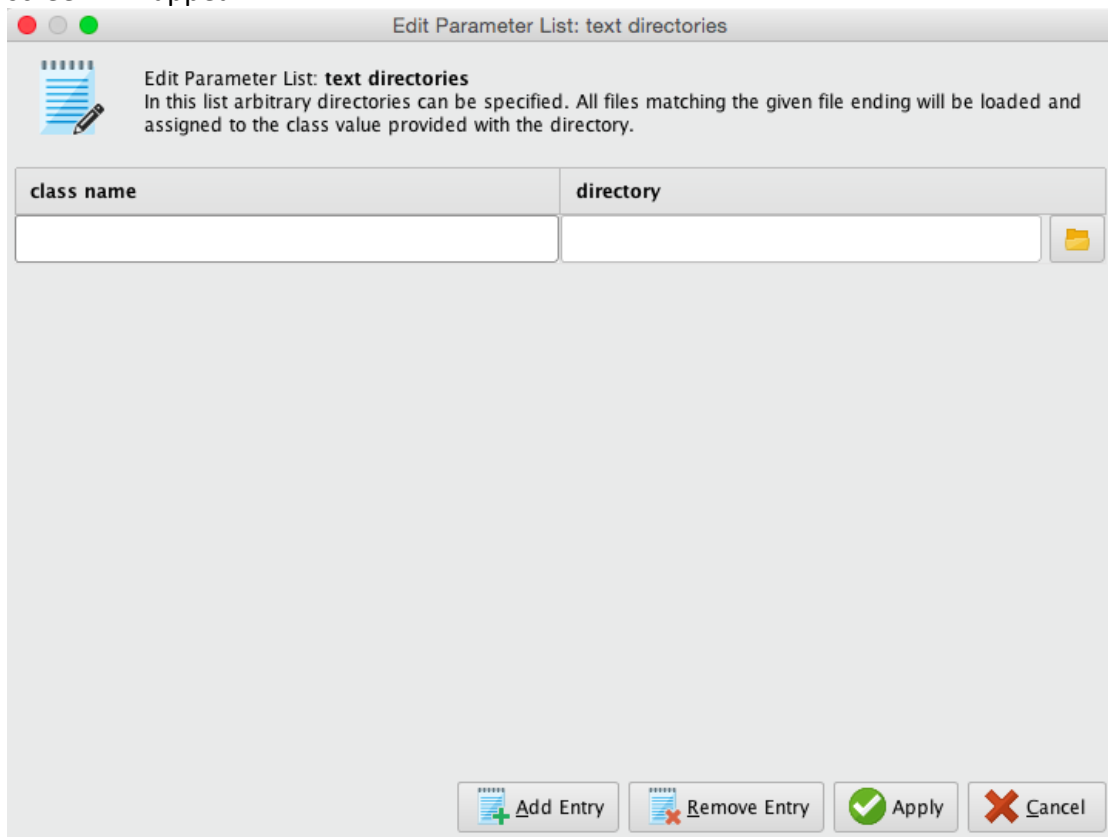
5. Click the folder underneath directory and the folders on your computer will appear. My text analytics folder is saved under my desktop and I right click that.

File Name	Size	Type	Last Modif...
📁 Applications		File Folder	Jul 27, 2015
📁 Desktop		File Folder	Apr 26, 2018
📁 Documents		File Folder	Apr 26, 2018
📁 Downloads		File Folder	Apr 26, 2018
📁 Library		File Folder	Mar 29, 2016
📁 Movies		File Folder	Nov 6, 2015
📁 Music		File Folder	Dec 9, 2015
📁 Pictures		File Folder	Dec 20, 2017
📁 Public		File Folder	Dec 31, 2013

6. Then right click on the text analytics folder. The three folders that we created earlier will appear.

📁 Text Analytics		File Folder	Apr 26, 2018

File Name	Size	Type	Last Modif...
📁 CMS		File Folder	Apr 24, 2018
📁 Enron		File Folder	Apr 24, 2018
📁 First Energy Corp		File Folder	Apr 26, 2018

7. Next, select the CMS folder and select "open" and we will be taken back to the beginning view. We will have to fill out the class name to give all of these documents their apporpiate labels. Type in "CMS" in class name like you see below.
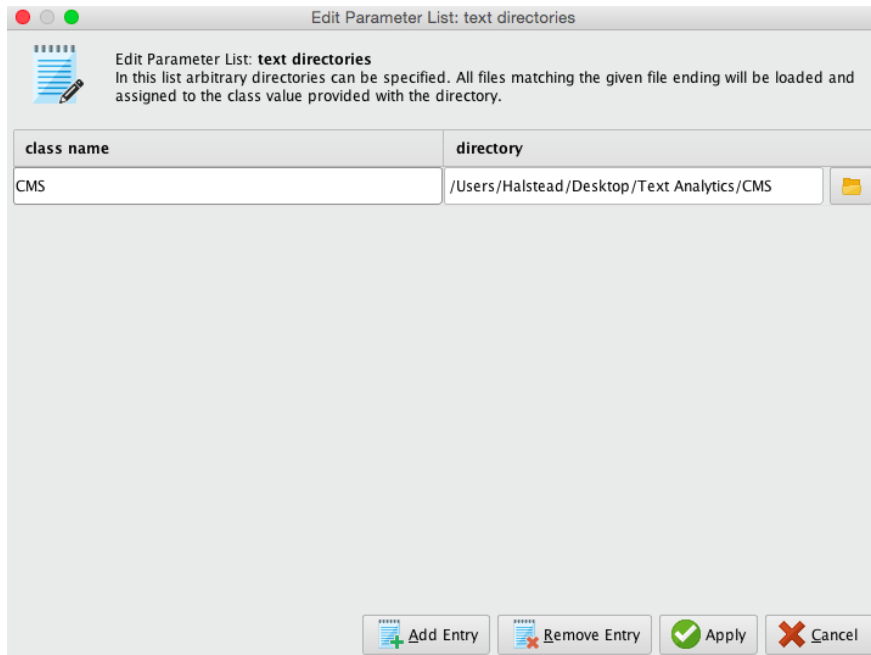
8. Repeat these steps with the other two companies. Your screen should look like this after recompleting these steps. Click "apply".
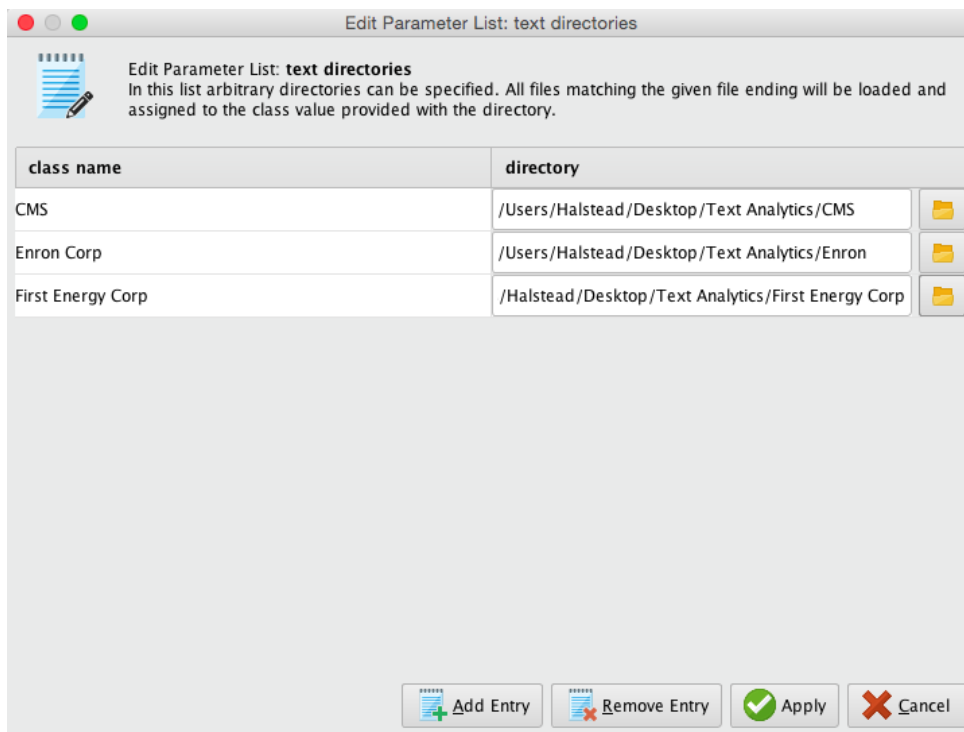
Step 7: Create List of Words.

1. Double click on the "process date from file" square in the process area of RapidMiner. The following screen should show up:
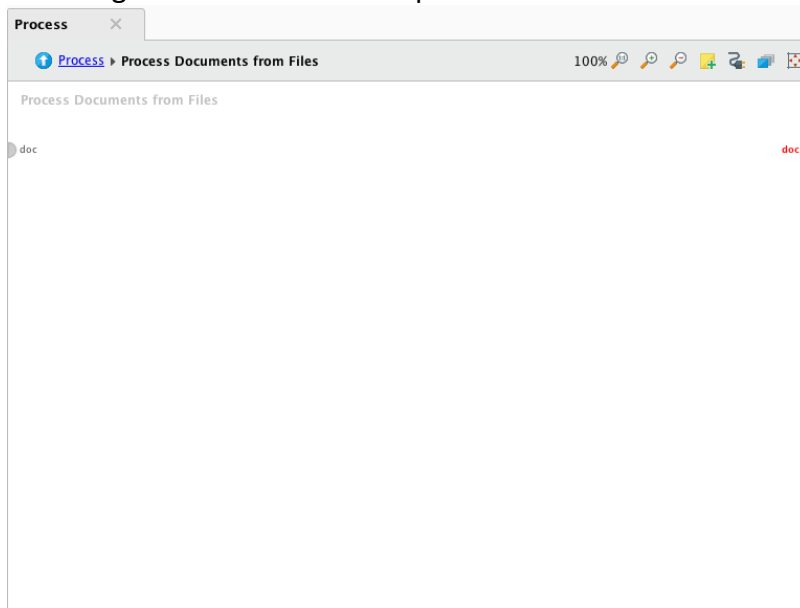
2. We are going to first process this text to show us all the lists of words in these 10-K documents. To do this, search "Tokenize" in the operatiors section of RapidMiner and drag it into the process.
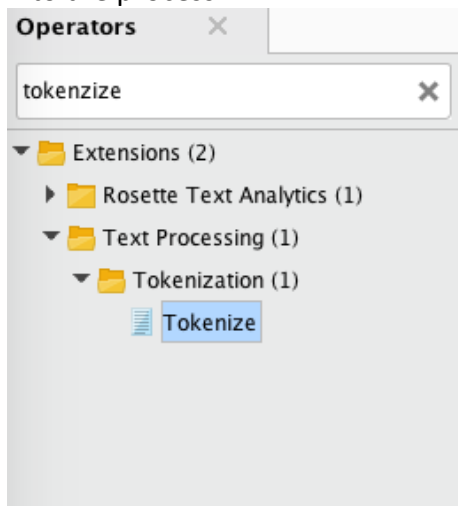
The tokenize process will give us a numerical number for the amount of times a word comes up in all of these different 10-Ks.

3. After dragging "tokenize" into the process section we need to connect the "doc" port on the left side of the screen to the "doc" port on the the left side of "tokenize". This step is shown below:
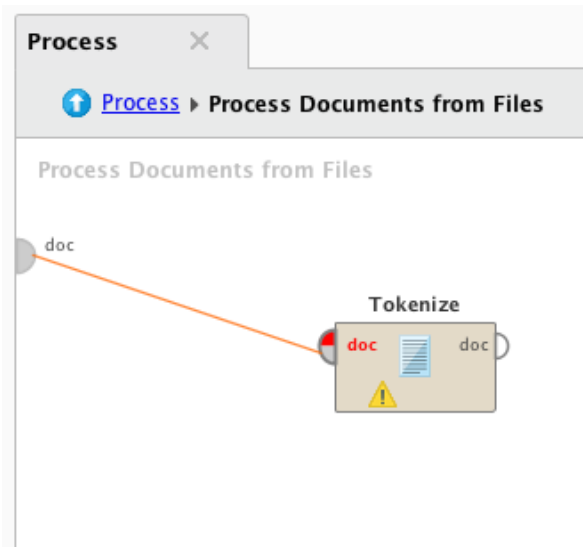
4. We then need to connect the right "doc" port of tokenize, to the "doc" port on the right side of the screen. After doing this, your screen should match the one below:
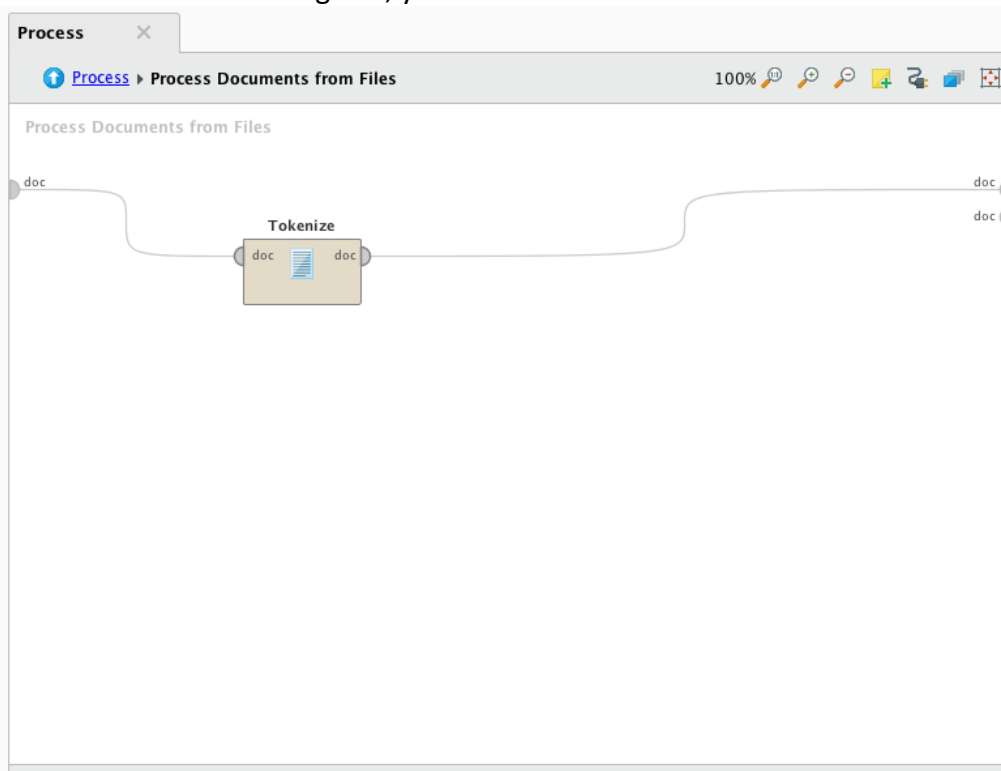
5. Select the blue "process" button to go back to the … and connect the word port of the "process documents from files" right side to the right side of the process screen, as you can see below:
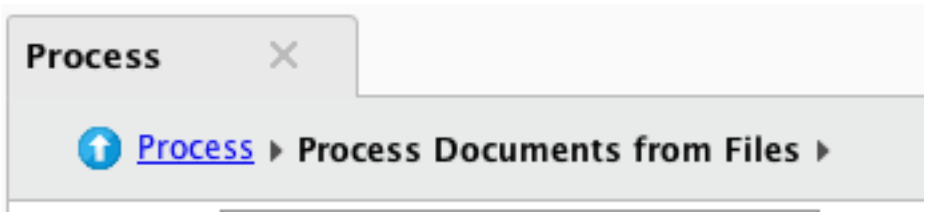
6. Next, select the blue "play" button to run the process.
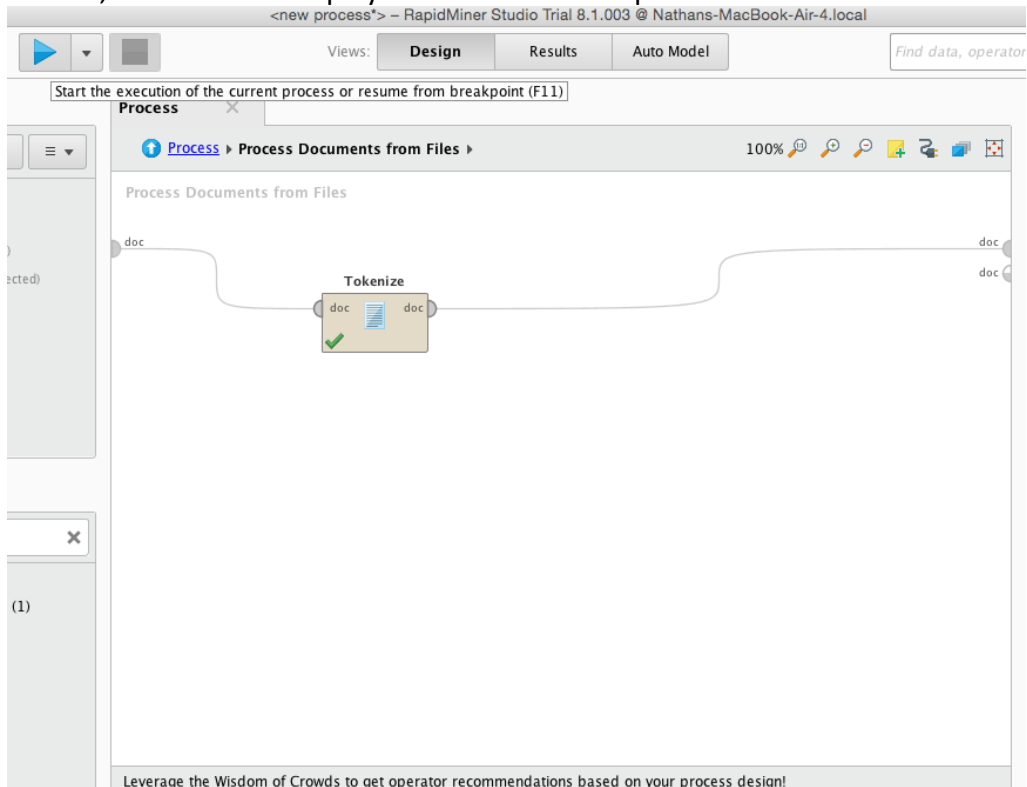
After a few seconds, we see the following results. This shows us all the words used in the document and how many times they are used. This displays total occurance, as well as the abilty to sort by the three different companies 10-Ks.

Word	Attribu...	Tot... ↓	Docum...	CMS	Enron ...	First E...
the	the	15391	3	4958	2510	7923
of	of	13674	3	4657	2545	6472
and	and	10075	3	3390	2694	3991
to	to	7193	3	2270	1751	3172
in	in	5501	3	1758	1281	2462
a	a	3221	3	1195	721	1305
as	as	3183	3	876	408	1899
for	for	2995	3	1028	623	1344
or	or	2860	3	1042	475	1343
Enron	Enron	2410	1	0	2410	0
The	The	2359	3	539	493	1327
Company	Company	2310	3	604	331	1375
by	by	2186	3	633	366	1187
on	on	1947	3	538	517	892
C	C	1863	5	646	685	532
be	be	1733	3	540	263	930
...	...	1607	3	634	416	647

Data

Step 8: Modify List of Words Results.

The amount of words that the initial text analysis displays can be a bit overwhelming. Using RapidMiner we will now add some operators to make the results more useful.

1. Double click again on the "process docments from file" square in the process. We will be adding more operators to this process. Your screen should look like this after you double click:
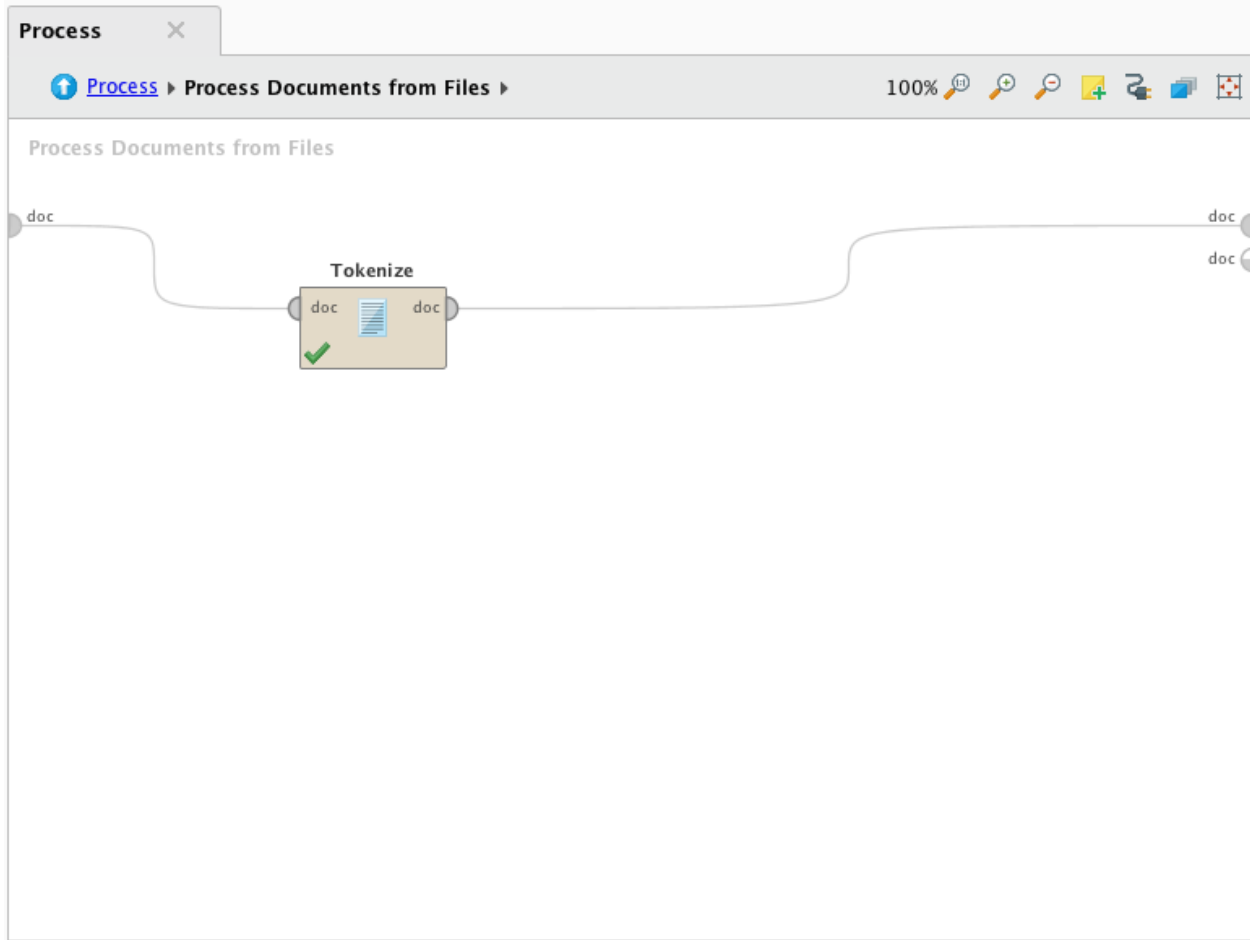
2. In the operators search tab, search "transform cases" and drag that operator to the process. Transform cases operator allows our document to read all uppercase and lowercase words as the same.

Operators ✕

transform cases ✕

▼ 📁 Extensions (1)
   ▼ 📁 Text Processing (1)
      ▼ 📁 Transformation (1)
         📄 Transform Cases

Process ✕

ⓘ Process ▸ **Process Documents from Files** ▸   100% 🔍 🔍 🔍 ▦ ⧉ ▦ ⊞

Process Documents from Files

doc   doc
                       Tokenize   doc
                     doc 📄 doc
                     ✓

                        Transform Cases
                     doc 📄 doc
                        ⚠

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
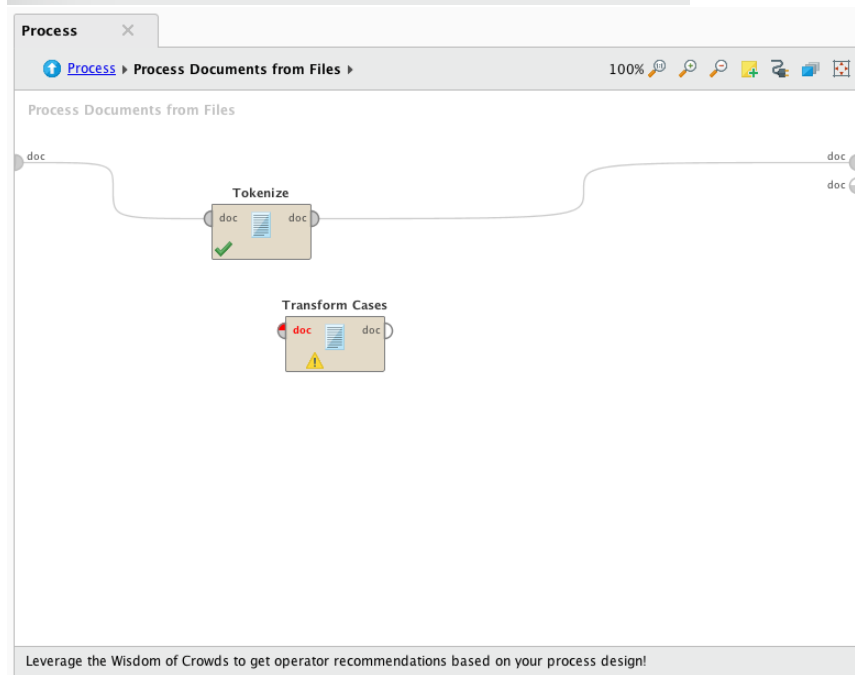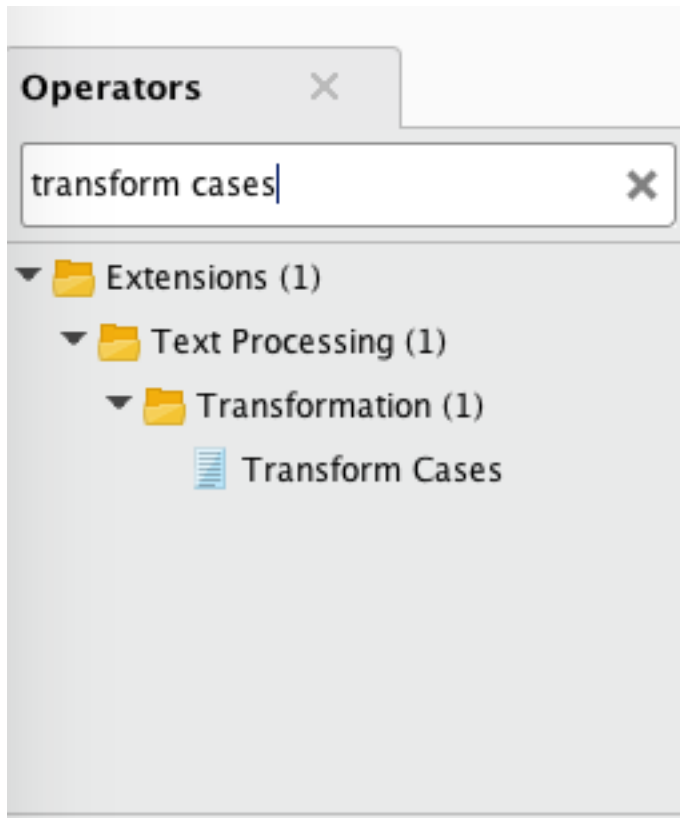
3. Follow the step in 2 to add in "Filter Tokens (by length)" and "Filter Stopworks (English)" to the process as well. Both of these operators will allow us to elimate common stopwords and eliminate certain words by length. Your screen will now look like this:
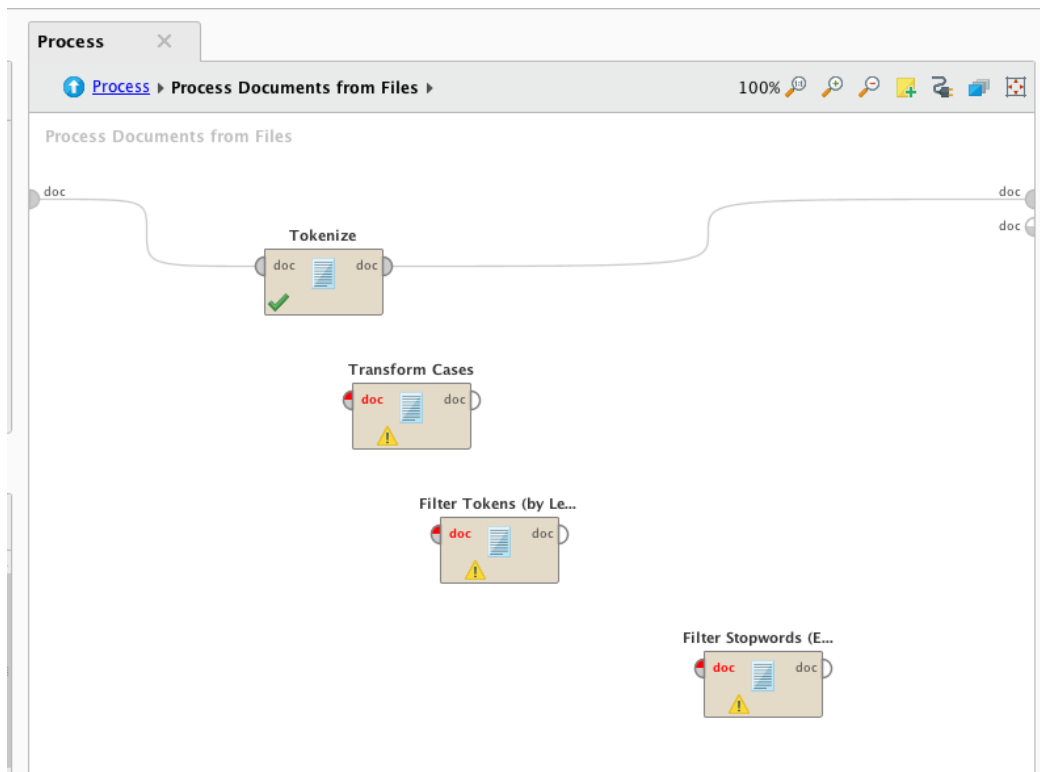
4. Next, we need to edit some of the operators. Click on the "Filter Tokens (by Length)" operator. We need to edit the parameters of this operator to decide how long of words we want to analyze. This will help clean up messy data by eliminating filler words and words that are less than 2 letter. Edit the "min chars" to 3 and the "max chars" to 999 as you see below.
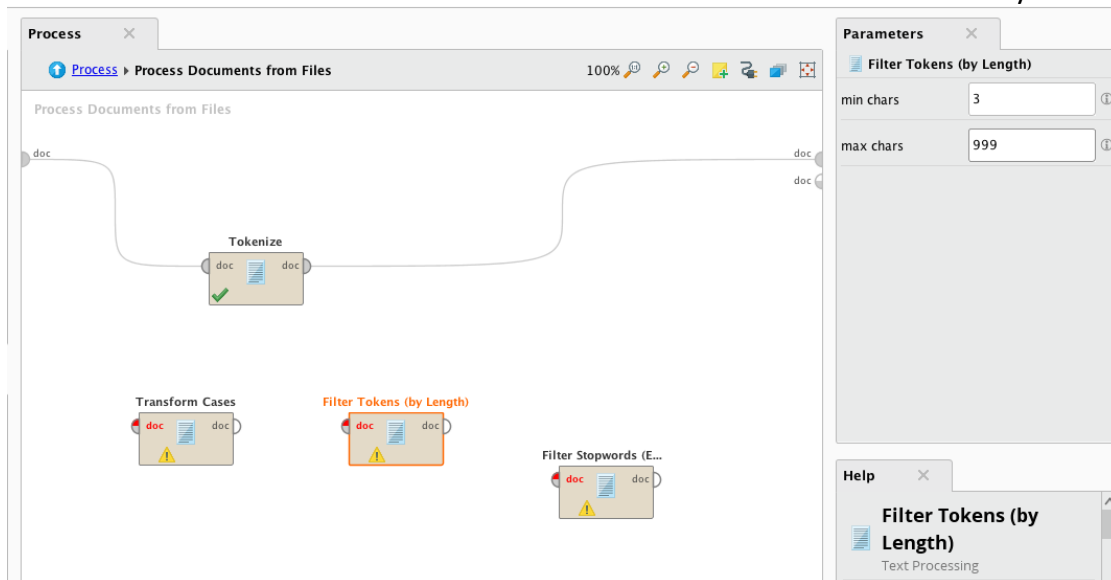
5. Like in the previous parts of this assignment, the new operators are currently not connected to the process. To update this, first eliminate the connect from the "tokenize" operator to the

"doc" port of the right side of the process screen. This can be done by hovering over the line and a red X will appear. Click on that red X.
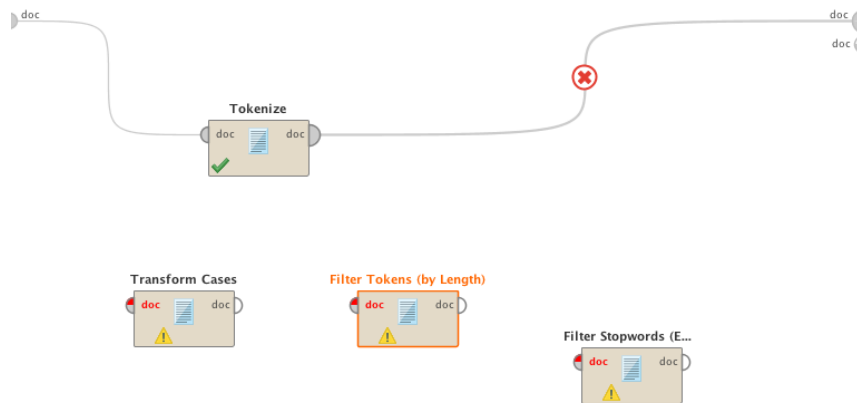
6. Now we need to connect all of the operators by their "doc" ports. Your screen should look somewhat similar to the screenshot below: (Hint, you can drag the operators to get them all in a straight line like the screenshot below, but this is not required)
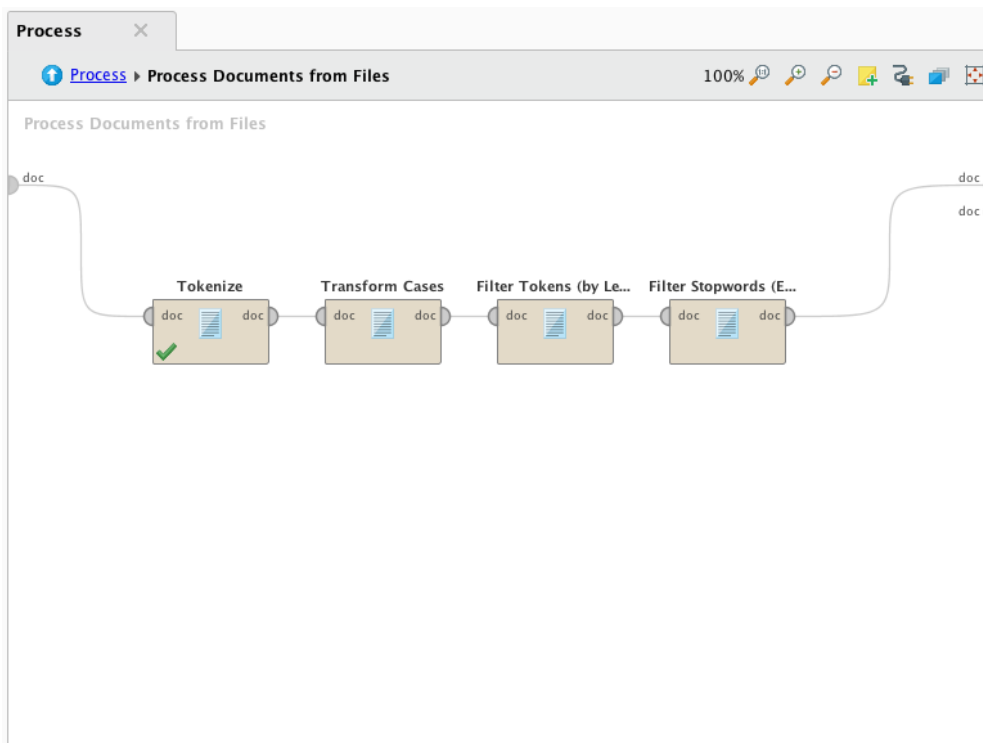
7. Now we can rerun our process by clicking the blue "Play" button and see how the results are now different. If we compare these results with the results from our previous step, the words that now are most common are completely different.

Word	Attribu...	Tot... ↓	Docum...	CMS	Enron ...	First E...
company	company	2664	3	739	399	1526
enron	enron	2535	1	0	2535	0
million	million	1720	3	640	418	662
stock	stock	1449	3	621	316	512
energy	energy	1387	3	995	282	110
gas	gas	1251	3	847	374	30
cms	cms	1161	1	1161	0	0
power	power	1136	3	232	623	281
interest	interest	1130	3	311	211	608
consum...	consum...	1114	3	1105	2	7
december	december	1055	3	382	248	425
delaware	delaware	1024	3	28	983	13
income	income	1000	3	379	196	425
bonds	bonds	937	3	206	7	724
series	series	913	3	180	38	695
financial	financial	856	3	396	218	242

Step 9: Find Passive Verbs.

Now that we have are data sorted into more useful information we want to be able to better analyze the differences between these companies. First, let's identify which 10-K had more passive tone verbs. We look for passive words as companies often use passive voice to explain something.

1. To start this process, double click the "process documents from file" operator in the process screen again.

2. We then need to search for a new operator titled "Filter Tokens (by POS Tag)". When we find it, drag it into the process. Your screen should look similar to the one below:
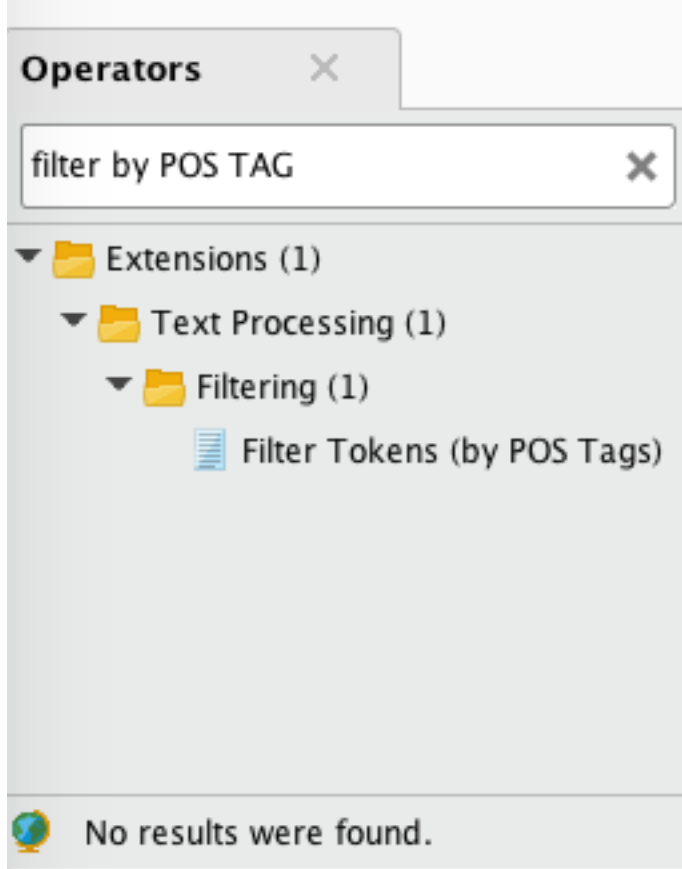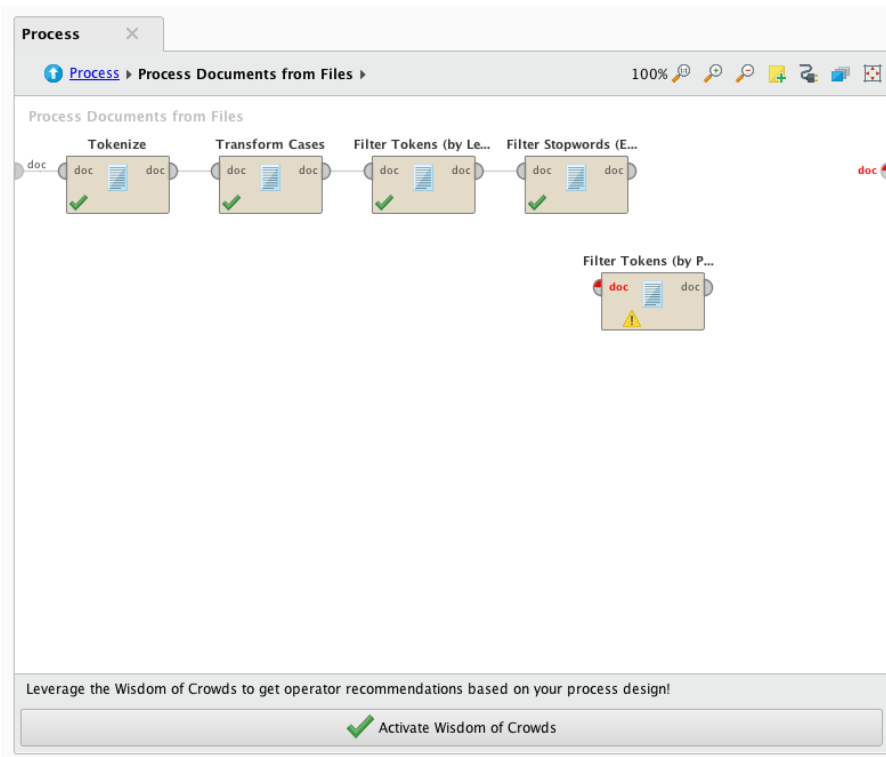
3. Next, attach the doc ports to the proper spaces. The process should look similar to this now:
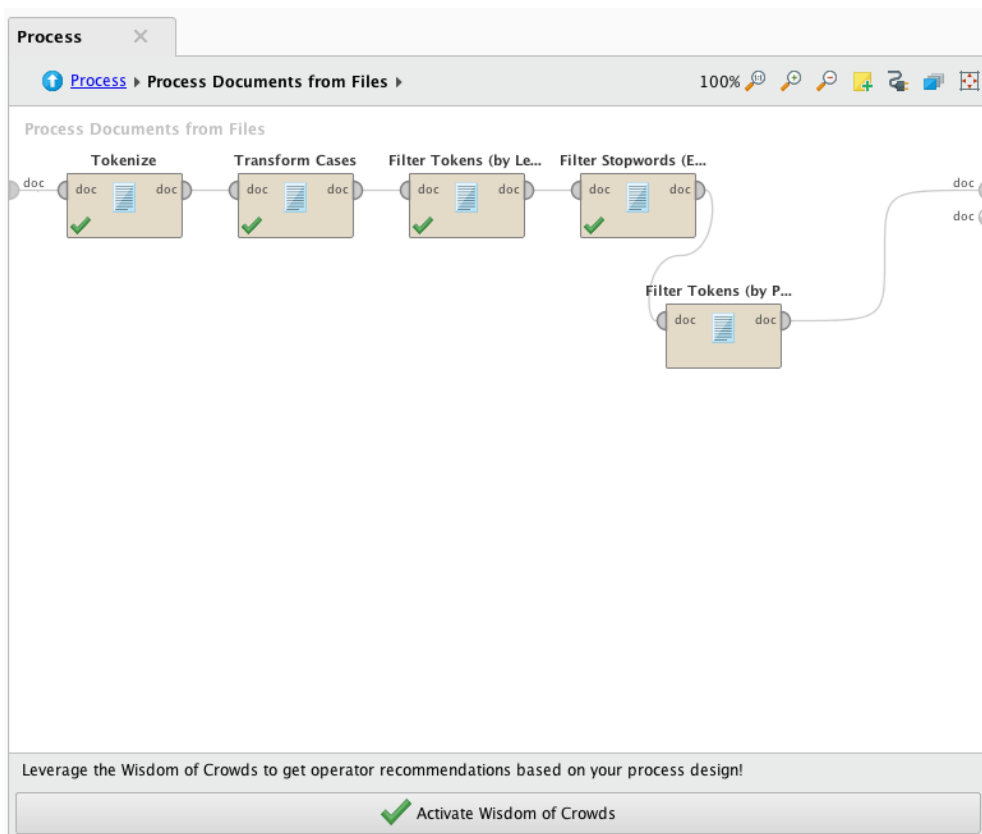
4. Click on the "Filter Tokens (by POS TAG)" so we can edit the parameters. In the parameters section we need to make sure that the languae is set to English and the expression is set to VBN. (Note: The VBN comes from the part-of-speech tagging model, this allows the data to searc for certain parts of speech. VBN in our example, is searching are data for past particle verbs.)
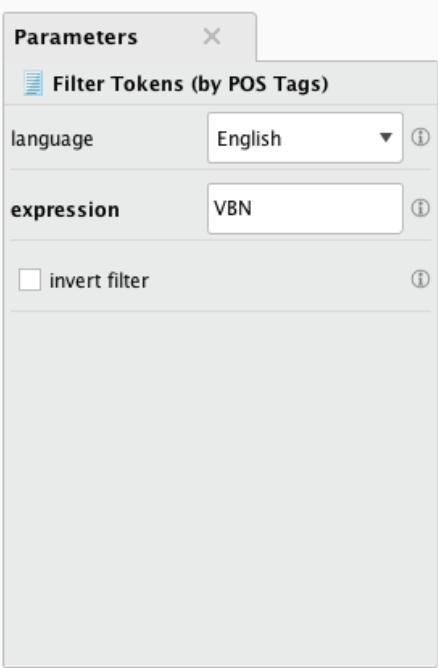
5. Click the play buttom on the process and view the results. Your results should look like this:

Word	Attribu...	Tot... ↓	Docum...	CMS	Enron ...	First E...
dated	dated	632	3	169	52	411
limited	limited	618	3	55	482	81
consolid...	consolid...	616	3	286	103	227
preferred	preferred	587	3	234	102	251
related	related	443	3	129	176	138
based	based	337	3	96	92	149
issued	issued	328	3	122	45	161
required	required	282	3	102	60	120
filed	filed	269	3	81	75	113
ended	ended	268	3	101	63	104
increased	increased	251	3	110	56	85
paid	paid	246	3	100	35	111
made	made	232	3	80	26	126
defined	defined	207	3	30	18	159
included	included	190	3	34	61	95
owned	owned	190	3	60	53	77

28

6. Next, analyze which company as more passive verbs in their 10-Ks. To do this, we will click on the names of the 10-K to sort by those documents. Let us first look at CMS. The screen shot below should be sorted by CMS Energy's 10K:

Word	Attribu...	Total O...	Docum...	CMS ↓	Enron ...	First E...
consolid...	consolid...	616	3	286	103	227
preferred	preferred	587	3	234	102	251
dated	dated	632	3	169	52	411
related	related	443	3	129	176	138
issued	issued	328	3	122	45	161
increased	increased	251	3	110	56	85
required	required	282	3	102	60	120
ended	ended	268	3	101	63	104
paid	paid	246	3	100	35	111
based	based	337	3	96	92	149
filed	filed	269	3	81	75	113
made	made	232	3	80	26	126
united	united	163	3	62	47	54
continued	continued	104	3	60	9	35
owned	owned	190	3	60	53	77
recorded	recorded	109	3	59	33	17

7. When we look at Enron's we see this:

Word	Attribu...	Total O...	Docum...	CMS	Enr... ↓	First E...
limited	limited	618	3	55	482	81
related	related	443	3	129	176	138
consolid...	consolid...	616	3	286	103	227
preferred	preferred	587	3	234	102	251
based	based	337	3	96	92	149
filed	filed	269	3	81	75	113
ended	ended	268	3	101	63	104
included	included	190	3	34	61	95
required	required	282	3	102	60	120
increased	increased	251	3	110	56	85
owned	owned	190	3	60	53	77
dated	dated	632	3	169	52	411
expected	expected	172	3	46	52	74
undersi...	undersi...	60	3	3	52	5
united	united	163	3	62	47	54
issued	issued	328	3	122	45	161

8. When we look at FirstEnergy's we see this:

Word	Attribu...	Total O...	Docum...	CMS	Enron ...	Firs... ↓
dated	dated	632	3	169	52	411
preferred	preferred	587	3	234	102	251
consolid...	consolid...	616	3	286	103	227
issued	issued	328	3	122	45	161
defined	defined	207	3	30	18	159
based	based	337	3	96	92	149
related	related	443	3	129	176	138
authoriz...	authoriz...	177	3	43	8	126
made	made	232	3	80	26	126
required	required	282	3	102	60	120
filed	filed	269	3	81	75	113
paid	paid	246	3	100	35	111
ended	ended	268	3	101	63	104
registered	registered	144	3	35	6	103
amended	amended	170	3	36	33	101
included	included	190	3	34	61	95

9. Students should form conclusions about these results in a short paragraph.

Step 10: Find Uncertainty Words using Loughran-McDonald Sentiment Word Lists.

Compare the number of words associated with uncertainty is another way to analyze 10-K's.

1. Click on the "Filter Tokens (By POS Tag)" and then click delete. We will be replacing this with another operator.

2. Search for the operator "Filter Tokens (by Content)" and drag it to the process. Your screen should look similar to this:
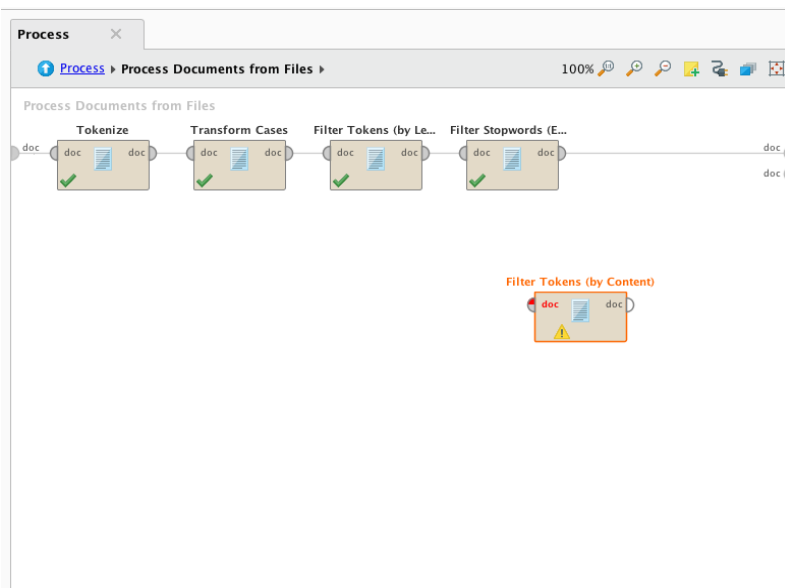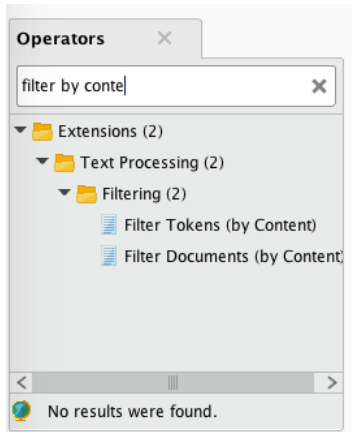
3. Connect the "Filter Tokens (by Content)" to the correct doc ports. Your screenshot should look similar to this:

31
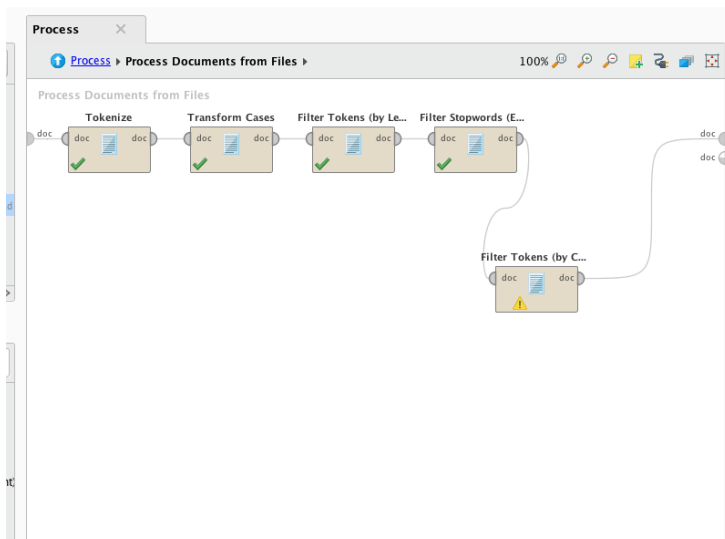
4. We need to search for uncertain words by using a list provided to us. This list is coming from Loughran-McDonald Sentiment Word Lists – Uncertainty. Open the document entitled uncertainty words.

5. The document should look like this:

```
.*ABEYANCE.*|.*ABEYANCES.*|.*ALMOST.*|.*ALTERATION.*|.*ALTERATIONS.*|.*AMBIGUITIE
S.*|.*AMBIGUITY.*|.*AMBIGUOUS.*|.*ANOMALIES.*|.*ANOMALOUS.*|.*ANOMALOUSLY.*|.
*ANOMALY.*|.*ANTICIPATE.*|.*ANTICIPATED.*|.*ANTICIPATES.*|.*ANTICIPATING.*|.*ANTICI
PATION.*|.*ANTICIPATIONS.*|.*APPARENT.*|.*APPARENTLY.*|.*APPEAR.*|.*APPEARED.*|.*
APPEARING.*|.*APPEARS.*|.*APPROXIMATE.*|.*APPROXIMATED.*|.*APPROXIMATELY.*|.*AP
PROXIMATES.*|.*APPROXIMATING.*|.*APPROXIMATION.*|.*APPROXIMATIONS.*|.*ARBITRAR
ILY.*|.*ARBITRARINESS.*|.*ARBITRARY.*|.*ASSUME.*|.*ASSUMED.*|.*ASSUMES.*|.*ASSUMI
NG.*|.*ASSUMPTION.*|.*ASSUMPTIONS.*|.*BELIEVE.*|.*BELIEVED.*|.*BELIEVES.*|.*BELIEVI
NG.*|.*CAUTIOUS.*|.*CAUTIOUSLY.*|.*CAUTIOUSNESS.*|.*CLARIFICATION.*|.*CLARIFICATIO
NS.*|.*CONCEIVABLE.*|.*CONCEIVABLY.*|.*CONDITIONAL.*|.*CONDITIONALLY.*|.*CONFUSE
S.*|.*CONFUSING.*|.*CONFUSINGLY.*|.*CONFUSION.*|.*CONTINGENCIES.*|.*CONTINGENCY.
*|.*CONTINGENT.*|.*CONTINGENTLY.*|.*CONTINGENTS.*|.*COULD.*|.*CROSSROAD.*|.*CRO
SSROADS.*|.*DEPEND.*|.*DEPENDED.*|.*DEPENDENCE.*|.*DEPENDENCIES.*|.*DEPENDENCY.
*|.*DEPENDENT.*|.*DEPENDING.*|.*DEPENDS.*|.*DESTABILIZING.*|.*DEVIATE.*|.*DEVIATED
.*|.*DEVIATES.*|.*DEVIATING.*|.*DEVIATION.*|.*DEVIATIONS.*|.*DIFFER.*|.*DIFFERED.*|.*
DIFFERING.*|.*DIFFERS.*|.*DOUBT.*|.*DOUBTED.*|.*DOUBTFUL.*|.*DOUBTS.*|.*EXPOSURE.
*|.*EXPOSURES.*|.*FLUCTUATE.*|.*FLUCTUATED.*|.*FLUCTUATES.*|.*FLUCTUATING.*|.*FLU
CTUATION.*|.*FLUCTUATIONS.*|.*HIDDEN.*|.*HINGES.*|.*IMPRECISE.*|.*IMPRECISION.*|.*I
MPRECISIONS.*|.*IMPROBABILITY.*|.*IMPROBABLE.*|.*INCOMPLETENESS.*|.*INDEFINITE.*|.
*INDEFINITELY.*|.*INDEFINITENESS.*|.*INDETERMINABLE.*|.*INDETERMINATE.*|.*INEXACT.*
|.*INEXACTNESS.*|.*INSTABILITIES.*|.*INSTABILITY.*|.*INTANGIBLE.*|.*INTANGIBLES.*|.*LIK
ELIHOOD.*|.*MAY.*|.*MAYBE.*|.*MIGHT.*|.*NEARLY.*|.*NONASSESSABLE.*|.*OCCASIONAL
LY.*|.*ORDINARILY.*|.*PENDING.*|.*PERHAPS.*|.*POSSIBILITIES.*|.*POSSIBILITY.*|.*POSSIBL
E.*|.*POSSIBLY.*|.*PRECAUTION.*|.*PRECAUTIONARY.*|.*PRECAUTIONS.*|.*PREDICT.*|.*PR
EDICTABILITY.*|.*PREDICTED.*|.*PREDICTING.*|.*PREDICTION.*|.*PREDICTIONS.*|.*PREDICTI
VE.*|.*PREDICTOR.*|.*PREDICTORS.*|.*PREDICTS.*|.*PRELIMINARILY.*|.*PRELIMINARY.*|.*P
RESUMABLY.*|.*PRESUME.*|.*PRESUMED.*|.*PRESUMES.*|.*PRESUMING.*|.*PRESUMPTION
.*|.*PRESUMPTIONS.*|.*PROBABILISTIC.*|.*PROBABILITIES.*|.*PROBABILITY.*|.*PROBABLE.*
|.*PROBABLY.*|.*RANDOM.*|.*RANDOMIZE.*|.*RANDOMIZED.*|.*RANDOMIZES.*|.*RANDO
MIZING.*|.*RANDOMLY.*|.*RANDOMNESS.*|.*REASSESS.*|.*REASSESSED.*|.*REASSESSES.*|.
*REASSESSING.*|.*REASSESSMENT.*|.*REASSESSMENTS.*|.*RECALCULATE.*|.*RECALCULATE
```

Copy the list of all words in this document. By selecting all the words and then right click copy.

6. Next we need to edit the parameters in the "Filter Tokens (by Content)" operator. Click on the operator and edit the parameters this way. First change the condition to "contains match".

Parameters ✕

▤ **Filter Tokens (by Content)**

condition	contains match ▼ ⓘ
regular expression	[ ] 🔍 ⓘ
☐ case sensitive	ⓘ
☐ invert condition	ⓘ

7. Next click on the document with the magnify glass to the right of "Regular Expression" in the parameter box. Paste the list of uncertainty words you had copied from the prior step into the regular expression column. (Make sure the .*|.* copied over from the uncertainty document since this allows RapidMiner to appropriately search for each word.)
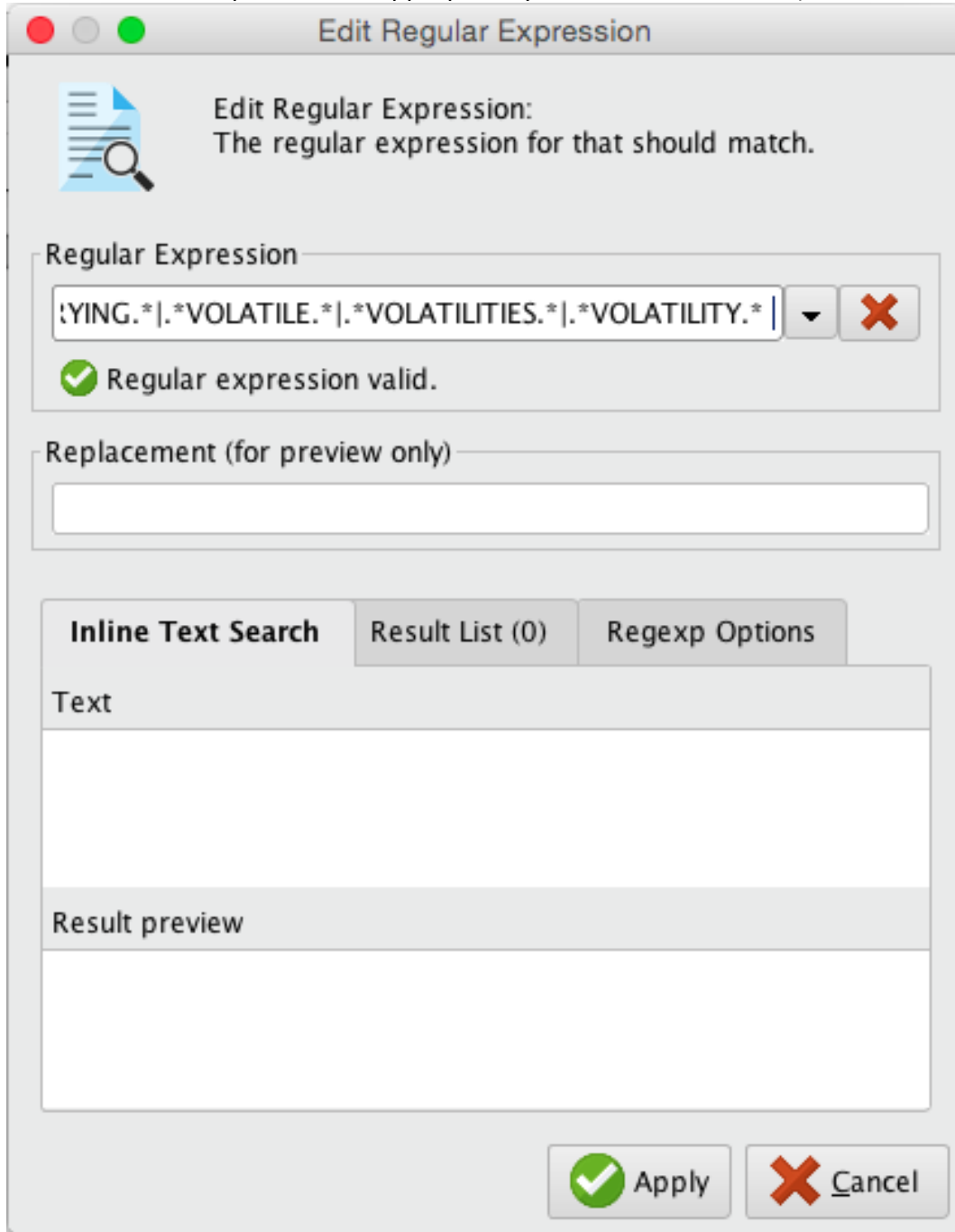
8. Click "apply" and then click the "blue" play button. THIS PROCESS WILL TAKE A FEW MINUTES BECAUSE OF THE AMOUNT OF WORDS RAPIDMINER IS SORTING THROUGH.

9. When analyzing the results, we see the following results:
When sorted by total occurrences:

Word	Attribu...	Tot... ↓	Docum...	CMS	Enron ...	First E...
approxi...	approxi...	374	3	64	110	200
risk	risk	260	3	93	141	26
indepen...	indepen...	123	3	61	23	39
uncertai...	uncertai...	86	3	78	3	5
believes	believes	78	3	41	33	4
assumed	assumed	60	3	13	22	25
risks	risks	57	3	20	21	16
assump...	assump...	41	3	19	10	12
exposure	exposure	41	3	26	9	6
predict	predict	41	3	21	5	15
believe	believe	39	3	9	10	20
variable	variable	37	3	19	9	9
differen...	differen...	36	3	7	8	21
anticipa...	anticipa...	35	3	12	7	16
revised	revised	35	2	13	0	22
conting...	conting...	33	3	18	5	10

When sorted by CMS Energy we see:

Word	Attribu...	Total O...	Document Occurences	CMS ↓	Enron ...	First E...
risk	risk	260	3	93	141	26
uncertai...	uncertai...	86	3	78	3	5
approxi...	approxi...	374	3	64	110	200
indepen...	indepen...	123	3	61	23	39
believes	believes	78	3	41	33	4
exposure	exposure	41	3	26	9	6
predict	predict	41	3	21	5	15
risks	risks	57	3	20	21	16
assump...	assump...	41	3	19	10	12
variable	variable	37	3	19	9	9
conting...	conting...	33	3	18	5	10
fluctuati...	fluctuati...	29	3	17	6	6
anticipa...	anticipa...	20	2	15	5	0
revise	revise	33	3	15	1	17

When sorted by Enron we see:

Word	Attribu...	Total O...	Document Occurences	CMS	Enr... ↓	First E...
risk	risk	260	3	93	141	26
approxi...	approxi...	374	3	64	110	200
believes	believes	78	3	41	33	4
indepen...	indepen...	123	3	61	23	39
assumed	assumed	60	3	13	22	25
risks	risks	57	3	20	21	16
assump...	assump...	41	3	19	10	12
believe	believe	39	3	9	10	20
exposure	exposure	41	3	26	9	6
variable	variable	37	3	19	9	9
differen...	differen...	36	3	7	8	21
exposur...	exposur...	19	3	10	8	1
anticipa...	anticipa...	35	3	12	7	16

When sorted by FirstEnergy Corp we see:

Word	Attribu...	Total O...	Document Occurences	CMS	Enron ...	Firs... ↓
approxi...	approxi...	374	3	64	110	200
indepen...	indepen...	123	3	61	23	39
risk	risk	260	3	93	141	26
assumed	assumed	60	3	13	22	25
revised	revised	35	2	13	0	22
differen...	differen...	36	3	7	8	21
assump...	assump...	32	3	10	2	20
believe	believe	39	3	9	10	20
revise	revise	33	3	15	1	17
anticipa...	anticipa...	35	3	12	7	16
risks	risks	57	3	20	21	16
approxi...	approxi...	19	3	2	2	15
predict	predict	41	3	21	5	15

10. Students should form conclusions about these results in a short paragraph.

Step 11: Analyze Sentiment.

Sentiment analysis indicates if the textual tone is positive or negative. It also identifies whether the document is subjective or objective. In summary, this will identify what 10-Ks are spun in a certain way.

1. First go back to your design screen of the process and delete out "Filter Tokens (by content)". Your screen will then again look like this:
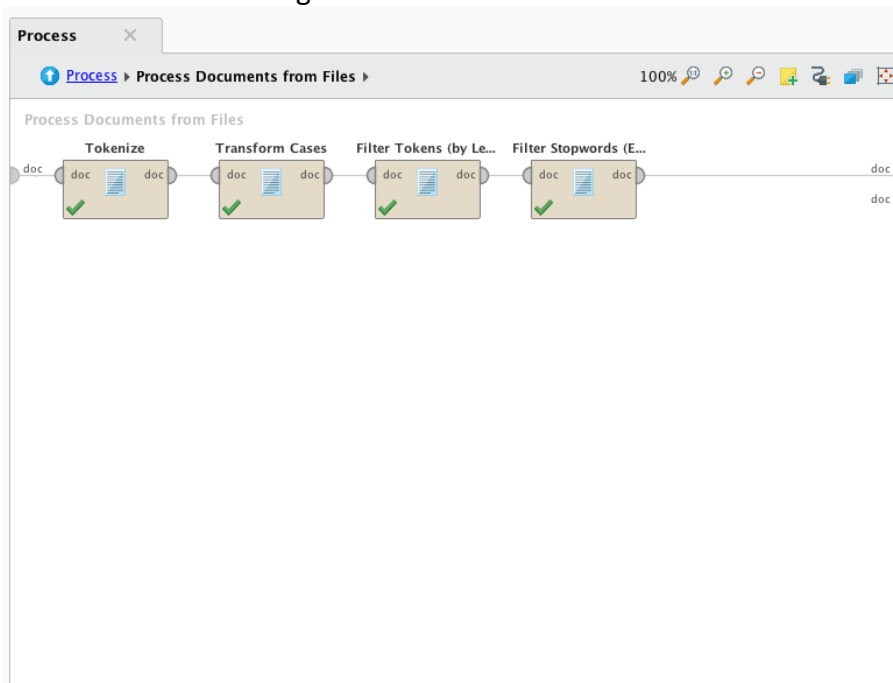
2. Click on the blue word "process" to get back into your "process Documents from file" design screen.
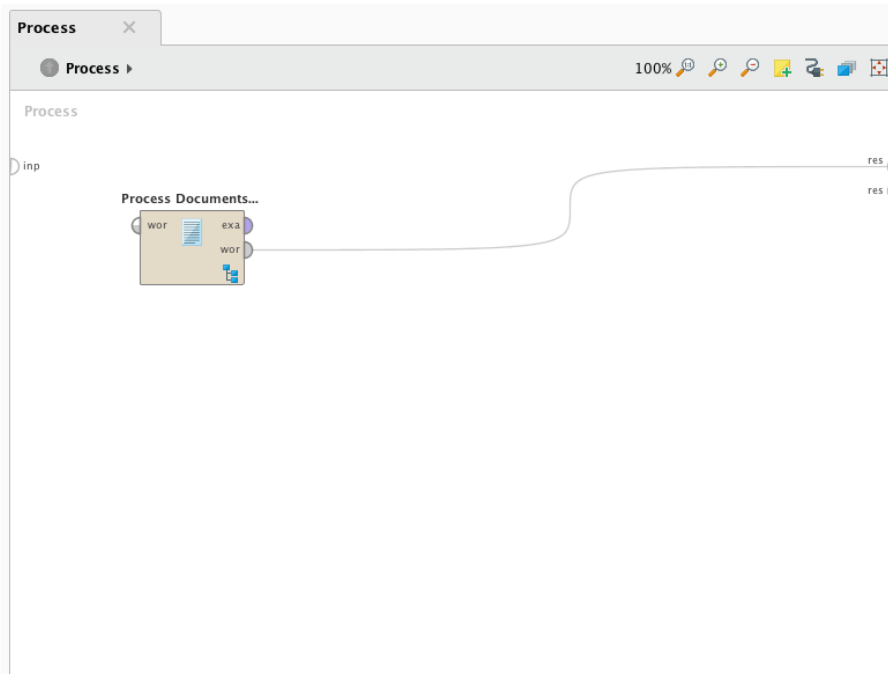
3. In the operators' tab this is where we will use the AYLIEN Text Analysis extension that we downloaded earlier in this assignment. Search AYLIEN and find the operator that is titled "Analyze Sentiment (document) and drag it into the process.
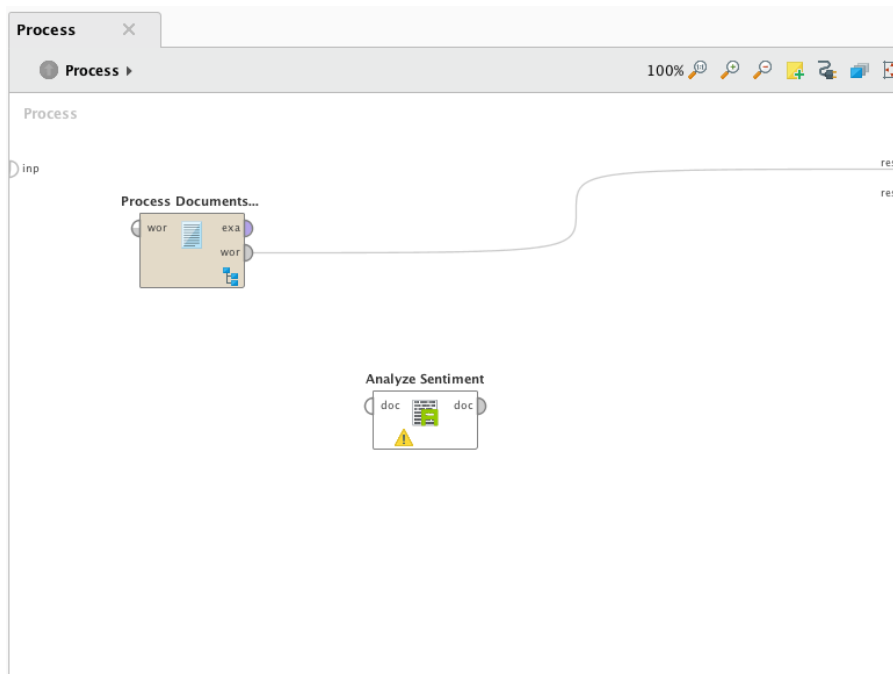
4. Search for the "Create Document" operator in the operator tab. Drag in the create document to the process. Repeat this step two more times. Your screen should look like this:
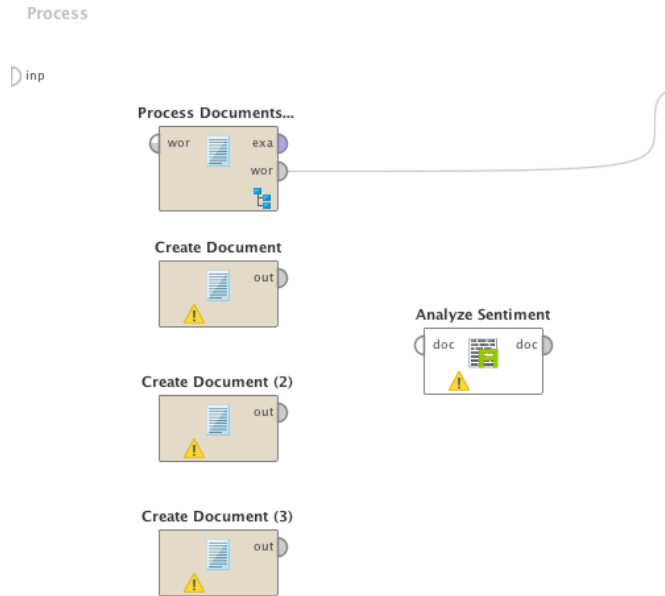
5. Right click on the first "create document" and select rename and name it CMS Energy.
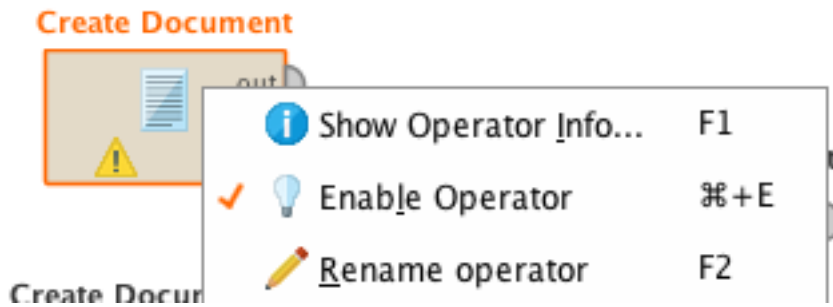
Follow this step with the other two create document operators as well. Your screen will now look similar to this one:

Process Documents...
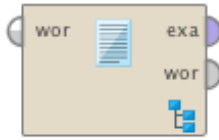
CMS Energy

Analyze Sentiment

Enron

FirstEnergy

6. Go back into the folders that we created with the text analytics 10-Ks and open CMS Energy and copy all of the text.

CMS Energy Corp 2001 10-K .txt

```
-----BEGIN PRIVACY-ENHANCED MESSAGE-----
Proc-Type: 2001,MIC-CLEAR
Originator-Name: webmaster@www.sec.gov
Originator-Key-Asymmetric:
 MFgwCgYEVQgBAQICAf8DSgAwRwJAW2sNKK9AVtBzYZmr6aGjlWyK3XmZv3dTINen
 TWSM7vrzLADbmYQaionwg5sDW3P6oaM5D3tdezXMm7z1T+B+twIDAQAB
MIC-Info: RSA-MD5,RSA,
 McES7xUZEL4p1JKza79R5pLWFj1gIrugfuTwucO4hC1KizEVezig18L5bN4iG+Ro
 gMTT5DIO/V8//Cb1EirTVw==

<SEC-DOCUMENT>0000950124-01-001548.txt : 20010326
<SEC-HEADER>0000950124-01-001548.hdr.sgml : 20010326
ACCESSION NUMBER:            0000950124-01-001548
CONFORMED SUBMISSION TYPE:   10-K
PUBLIC DOCUMENT COUNT:       11
CONFORMED PERIOD OF REPORT:  20001231
FILED AS OF DATE:            20010323

FILER:

     COMPANY DATA:
          COMPANY CONFORMED NAME:          CMS ENERGY CORP
          CENTRAL INDEX KEY:               0000811156
          STANDARD INDUSTRIAL CLASSIFICATION:  ELECTRIC & OTHER SERVICES COMBINED
[4931]
          IRS NUMBER:                      382726431
          STATE OF INCORPORATION:          MI
          FISCAL YEAR END:                 1231

     FILING VALUES:
```

7. Now go back to RapidMiner and select the operator titled "CMS Energy" and get the parameters to appear. Select Edit Text.
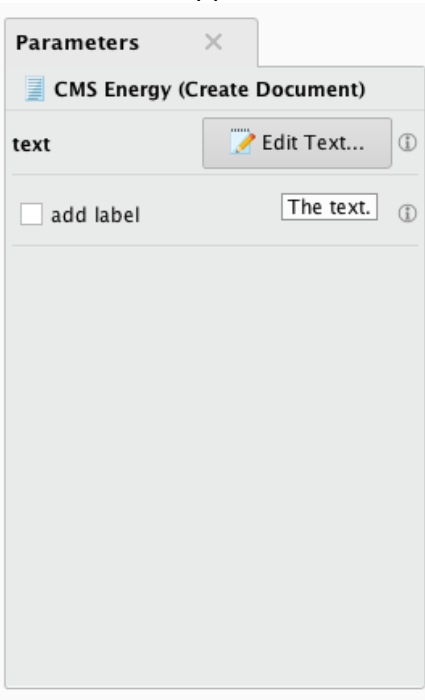
8. This edit text will appear and then paste in what you had just copied from the previous step. After this hit "apply".
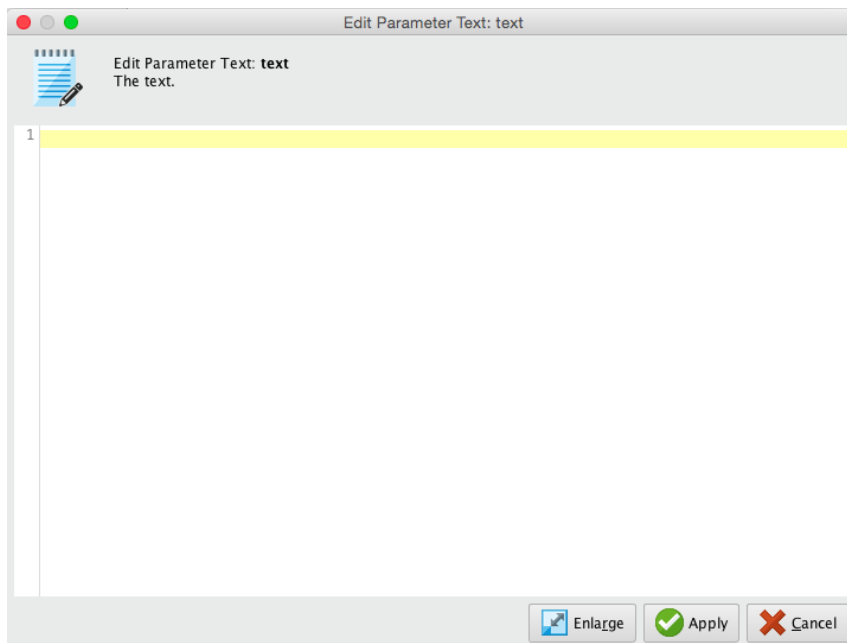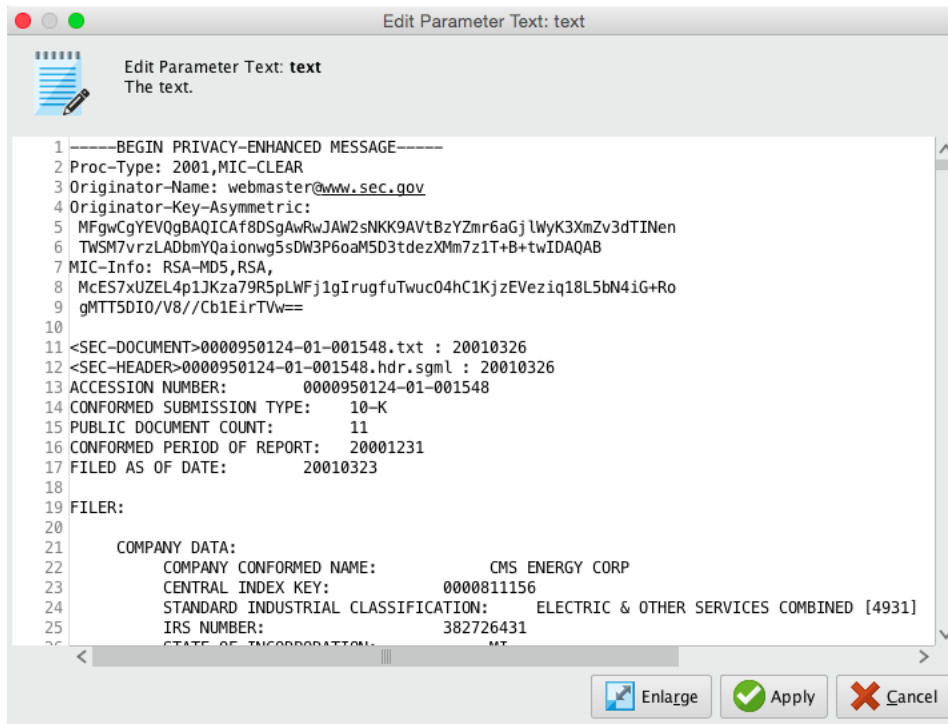
9. Follow these steps when adding in the other two 10-Ks to their appropriate operators.

10. Connect the "doc" ports to the "doc" ports in the "Analyze Sentiment" and then to the right side of the screen as well.
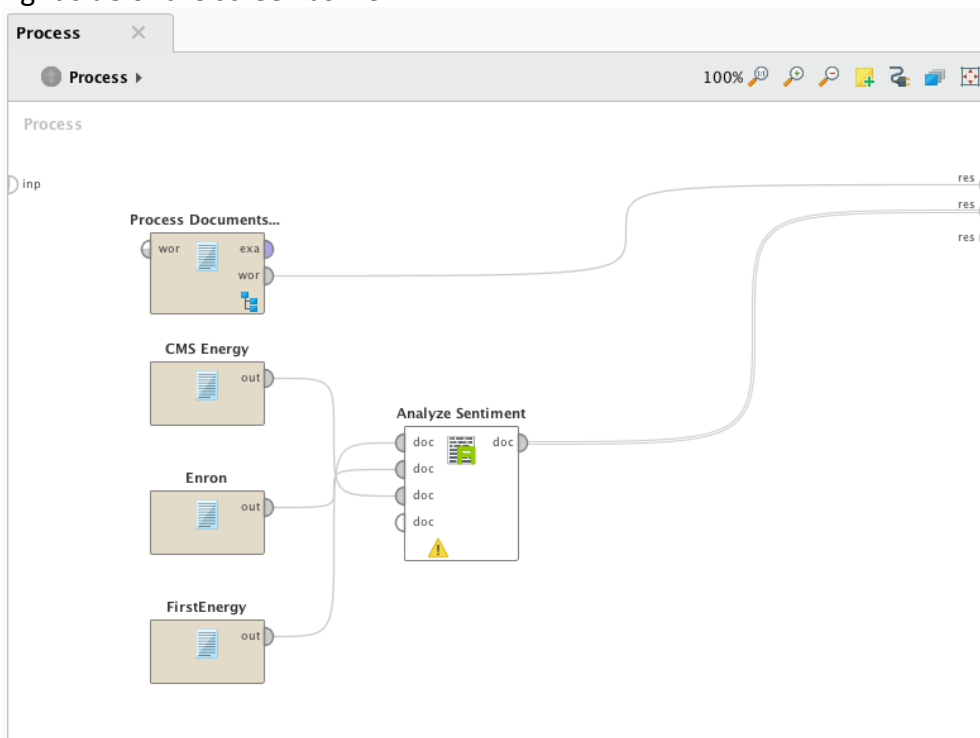
11. Click on the "Analyze Sentiment" operator and edit the parameters. The parameters should be changed. The connection should be the same that we had set up earlier so for this example it is named Text Analytics. Next, change the sentiment mode to "document".
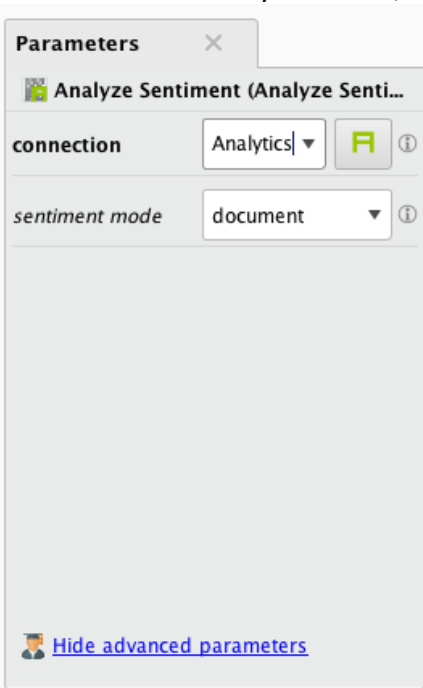
12. Now run the process by hitting the blue "play" button. The following results will pop up.
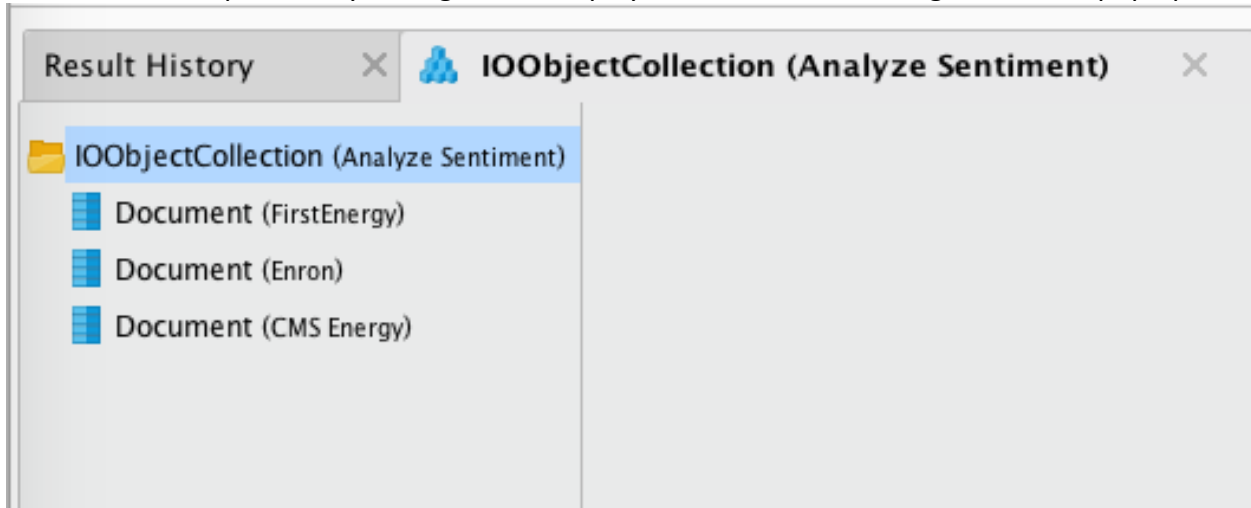
13. Now click on the First Energy document to get the sentiment results that we can see on the right side of the screen.

Click on the Enron document to see the results for that document:

Result History ✕ | 🔷 IOObjectCollection (Analyze Sentiment) ✕ | 📄 WordList (Process Documents from Files) ✕

📁 IOObjectCollection (Analyze Sentiment)
  📄 Document (FirstEnergy)
  📄 Document (Enron)
  📄 Document (CMS Energy)

Document ℹ

,――BEGIN PRIVACY-ENHANCED MESSAGE-----
Proc-Type: 2001,MIC-CLEAR
Originator-Name: webmaster@www.sec.gov
Originator-Key-Asymmetric:
 MFgwCgYEVQgBAQICAf8DSgAwRwJAW2sNKK9AVtBzYZmr6aGjlWyK3XmZv3dTINen
 TWSM7vrzLADbmYQaionwg5sDW3P6oaM5D3tdezXMm7z1T+B+twlDAQAB
MIC-Info: RSA-MD5,RSA,
 WVMqBDIh67AGAZHqEq9b+2C7mE5a9Dil4NiuXU+j8/U7MY4nL/lY9LZpo4elsZ/5
 uYSmX4MYDqjrU1O7b8UMYw==

<SEC-DOCUMENT>0001024401-01-500010.txt : 20010409
<SEC-HEADER>0001024401-01-500010.hdr.sgml : 20010409
ACCESSION NUMBER: 0001024401-01-500010
CONFORMED SUBMISSION TYPE: 10-K
PUBLIC DOCUMENT COUNT: 9
CONFORMED PERIOD OF REPORT: 20001231
FILED AS OF DATE: 20010402

FILER:

,――BEGIN PRIVACY-ENHANCED MESSAGE-----
Proc-Type: 2001,MIC-CLEAR
Originator-Name: webmaster@www.sec.gov
Originator-Key-Asymmetric:
 MFgwCgYEVQgBAQICAf8DSgAwRwJAW2sNKK9AVtBzYZmr6aGjlWyK3XmZv3dTINen
 TWSM7vrzLADbmYQaionwg5sDW3P6oaM5D3tdezXMm7z1T+B+twlDAQAB
MIC-Info: RSA-MD5,RSA,
 WVMqBDIh67AGAZHqEq9b+2C7mE5a9Dil4NiuXU+j8/U7MY4nL/lY9LZpo4elsZ/5
 uYSmX4MYDqjrU1O7b8UMYw==

<SEC-DOCUMENT>0001024401-01-500010.txt : 20010409
<SEC-HEADER>0001024401-01-500010.hdr.sgml : 20010409
ACCESSION NUMBER: 0001024401-01-500010
CONFORMED SUBMISSION TYPE: 10-K
PUBLIC DOCUMENT COUNT: 9
CONFORMED PERIOD OF REPORT: 20001231
FILED AS OF DATE: 20010402

FILER:

polarity	negative
subjecti...	unknown
polarity...	0.583
subjecti...	0

Click on the CMS Energy document to see the results for that document:

Result History ✕ | 🔷 IOObjectCollection (Analyze Sentiment) ✕ | 📄 WordList (Process Documents from Files) ✕

📁 IOObjectCollection (Analyze Sentiment)
  📄 Document (FirstEnergy)
  📄 Document (Enron)
  📄 Document (CMS Energy)

Document ℹ

-----BEGIN PRIVACY-ENHANCED MESSAGE-----
Proc-Type: 2001,MIC-CLEAR
Originator-Name: webmaster@www.sec.gov
Originator-Key-Asymmetric:
 MFgwCgYEVQgBAQICAf8DSgAwRwJAW2sNKK9AVtBzYZmr6aGjlWyK3XmZv3dTINen
 TWSM7vrzLADbmYQaionwg5sDW3P6oaM5D3tdezXMm7z1T+B+twlDAQAB
MIC-Info: RSA-MD5,RSA,
 McES7xUZEL4p1JKza79R5pLWFj1glrugfuTwucO4hC1KjzEVeziq18L5bN4iG+Ro
 gMTT5DIO/V8//Cb1EirTVw==

<SEC-DOCUMENT>0000950124-01-001548.txt : 20010326
<SEC-HEADER>0000950124-01-001548.hdr.sgml : 20010326
ACCESSION NUMBER: 0000950124-01-001548
CONFORMED SUBMISSION TYPE: 10-K
PUBLIC DOCUMENT COUNT: 11
CONFORMED PERIOD OF REPORT: 20001231
FILED AS OF DATE: 20010323

FILER:

-----BEGIN PRIVACY-ENHANCED MESSAGE-----
Proc-Type: 2001,MIC-CLEAR
Originator-Name: webmaster@www.sec.gov
Originator-Key-Asymmetric:
 MFgwCgYEVQgBAQICAf8DSgAwRwJAW2sNKK9AVtBzYZmr6aGjlWyK3XmZv3dTINen
 TWSM7vrzLADbmYQaionwg5sDW3P6oaM5D3tdezXMm7z1T+B+twlDAQAB
MIC-Info: RSA-MD5,RSA,
 McES7xUZEL4p1JKza79R5pLWFj1glrugfuTwucO4hC1KjzEVeziq18L5bN4iG+Ro
 gMTT5DIO/V8//Cb1EirTVw==

<SEC-DOCUMENT>0000950124-01-001548.txt : 20010326
<SEC-HEADER>0000950124-01-001548.hdr.sgml : 20010326
ACCESSION NUMBER: 0000950124-01-001548
CONFORMED SUBMISSION TYPE: 10-K
PUBLIC DOCUMENT COUNT: 11
CONFORMED PERIOD OF REPORT: 20001231
FILED AS OF DATE: 20010323

FILER:

polarity	positive
subjecti...	unknown
polarity...	0.576
subjecti...	0

14. Students should come to a conclusion about these results in a short paragraph.

APPENDIX A – Solution

Hasbro vs Mattel Using 2016 10-Ks

Passive Verb Results:

Word	Attribu...	Total O...	Docum...	Has... ↓	Mattel
related	related	339	2	225	114
consolid...	consolid...	304	2	195	109
based	based	279	2	137	142
compared	compared	164	2	107	57
ended	ended	118	2	103	15
included	included	128	2	93	35
offset	offset	115	2	68	47
united	united	78	2	67	11
expected	expected	116	2	62	54
increased	increased	86	2	60	26
deferred	deferred	97	2	57	40
recogni...	recogni...	103	2	48	55
recorded	recorded	78	2	45	33
continued	continued	51	2	44	7
required	required	65	2	42	23
dated	dated	42	2	41	1

Word	Attribu...	Total O...	Docum...	Hasbro	Mat... ↓
based	based	279	2	137	142
related	related	339	2	225	114
consolid...	consolid...	304	2	195	109
compared	compared	164	2	107	57
adjusted	adjusted	68	2	13	55
recogni...	recogni...	103	2	48	55
expected	expected	116	2	62	54
reported	reported	70	2	22	48
offset	offset	115	2	68	47
defined	defined	70	2	27	43
weighted	weighted	63	2	21	42
deferred	deferred	97	2	57	40
included	included	128	2	93	35
recorded	recorded	78	2	45	33
filed	filed	42	2	10	32

46

Uncertainty Results:

Word	Attribu...	Total O...	Docum...	Has... ↓	Mattel
approxi...	approxi...	186	2	80	106
risk	risk	80	2	35	45
intangib...	intangib...	40	2	32	8
anticipa...	anticipa...	35	2	31	4
intangible	intangible	66	2	29	37
assump...	assump...	54	2	22	32
believe	believe	23	2	19	4
believes	believes	55	2	18	37
risks	risks	40	2	18	22
contingent	contingent	19	2	15	4
indefinite	indefinite	15	1	15	0
anticipate	anticipate	16	2	13	3
depend...	depend...	17	2	11	6
differen...	differen...	19	2	11	8
indepen...	indepen...	15	2	11	4
fluctuati...	fluctuati...	24	2	10	14

Word	Attribu...	Total O...	Docum...	Hasbro	Mat... ↓
approxi...	approxi...	186	2	80	106
risk	risk	80	2	35	45
believes	believes	55	2	18	37
intangible	intangible	66	2	29	37
assump...	assump...	54	2	22	32
risks	risks	40	2	18	22
exposure	exposure	26	2	5	21
reasses...	reasses...	18	1	0	18
fluctuati...	fluctuati...	24	2	10	14
doubtful	doubtful	17	2	5	12
assumed	assumed	15	2	4	11
depend...	depend...	15	2	5	10
depends	depends	13	2	5	8
differ	differ	14	2	6	8
differen...	differen...	19	2	11	8
intangib...	intangib...	40	2	32	8

Sentiment Results:

10-K 1 d120433d10k.htm 10-K
Table of Contents
UNITED STATES SECURITIES AND EXCHANGE COMMISSION

Washington, D. C. 20549

Form 10-K

          x   ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES AND EXCHANGE ACT OF 1934
For the fiscal year ended December 27, 2015

Commission file number 1-6682

Hasbro, Inc.

(Exact Name of Registrant, As Specified in its Charter)

polarity	neutral
subjecti...	unknown
polarity...	0.448
subjecti...	0

10-K 1 mat1231201510-k.htm 10-K

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM 10-K
(Mark One)
ý
ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934

For the fiscal year ended December 31, 2015
or
¨
TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934

For the transition period from   to

polarity	neutral
subjecti...	unknown
polarity...	0.469
subjecti...	0

APPENDIX A – Solution

Wells Fargo, CitiBank, and Bank of America Using 2016 First Quarter 10-Qs

Wells Fargo

Passive Verbs:

Word	Attribu…	Total O…	Docum…	well… ↓	Citi Ba…	Bank o…
based	based	593	3	167	190	236
related	related	811	3	145	299	367
held	held	467	3	130	100	237
included	included	353	3	117	98	138
backed	backed	279	3	116	107	56
ended	ended	626	3	105	137	384
compared	compared	274	3	104	67	103
preferred	preferred	4930	3	104	84	4742
continued	continued	171	3	88	46	37
guarant…	guarant…	137	3	80	39	18
recorded	recorded	308	3	77	113	118
consolid…	consolid…	203	3	74	73	56
expected	expected	179	3	69	52	58
insured	insured	128	3	65	5	58
recogni…	recogni…	214	3	63	97	54
increased	increased	276	3	54	75	147

Uncertainty:

Word	Attribu…	Total O…	Docum…	Well… ↓	Citi Ba…	Bank o…	
risk	risk	1203	3	399	350	454	The colum
exposure	exposure	348	3	76	107	165	
assump…	assump…	105	3	39	29	37	
risks	risks	129	3	37	46	46	
variable	variable	101	3	35	37	29	
approxi…	approxi…	291	3	29	191	71	
exposur…	exposur…	186	3	21	82	83	
assump…	assump…	31	3	20	2	9	
believe	believe	59	3	20	1	38	
difference	difference	33	3	17	10	6	
unobser…	unobser…	56	3	17	23	16	
intangible	intangible	65	3	14	30	21	
differ	differ	34	3	13	11	10	
contingent	contingent	31	3	12	11	8	
intangib…	intangib…	43	3	12	2	29	

49

Sentiment Analysis:

Citi Bank:

Passive verbs:

Word	Attribu...	Total O...	Docum...	Wells F...	Citi ... ↓	Bank o...
related	related	811	3	145	299	367
consolid...	consolid...	558	3	93	255	210
based	based	593	3	167	190	236
ended	ended	626	3	105	137	384
offset	offset	241	3	41	118	82
recorded	recorded	308	3	77	113	118
sold	sold	241	3	57	113	71
weighted	weighted	215	3	48	112	55
backed	backed	284	3	118	110	56
purchas...	purchas...	300	3	46	102	152
held	held	467	3	130	100	237
included	included	353	3	117	98	138
recogni...	recogni...	214	3	63	97	54
issued	issued	378	3	53	87	238

Uncertainty:

Word	Attribu...	Total O...	Docum...	Wells F...	Citi ... ↓	Bank o...
risk	risk	1203	3	399	350	454
approxi...	approxi...	291	3	29	191	71
exposure	exposure	348	3	76	107	165
exposur...	exposur...	186	3	21	82	83
risks	risks	129	3	37	46	46
variable	variable	101	3	35	37	29
intangible	intangible	65	3	14	30	21
assump...	assump...	105	3	39	29	37
unobser...	unobser...	56	3	17	23	16
anticipa...	anticipa...	31	3	3	21	7
believes	believes	33	3	4	20	9
indepen...	indepen...	73	3	5	14	54
uncertai...	uncertai...	25	3	3	14	8
differen...	differen...	22	3	3	12	7
conting...	conting...	41	3	2	11	28

Analyze Sentiment:

	polarity	negative
	subjecti...	unknown
	polarity...	0.872
	subjecti...	0

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION WASHINGTON, D.C. 20549
FORM 10-Q
QUARTERLY REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE
Delaware
(State or other jurisdiction of incorporation or organization)
388 Greenwich Street, New York, NY
52-1568099
(I.R.S. Employer Identification No.)
For the quarterly period ended March 31, 2016 Commission file number 1-9924
Citigroup Inc.
(Exact name of registrant as specified in its charter)
(Address of principal executive offices)
(Registrant's telephone number, including area code)
10013
(Zip code)
(212) 559-1000
Indicate by check mark whether the registrant (1) has filed all reports required to be filed b
to file such reports), and (2) has been subject to such filing requirements for the past 90 d

Bank of America:

Passive Verbs:

Word	Attribu...	Total O...	Docum...	Wells F...	Citi Ba...	Ban... ↓
preferred	preferred	4989	3	106	85	4798
authoriz...	authoriz...	818	3	12	4	802
paid	paid	747	3	34	68	645
declared	declared	593	3	2	23	568
set	set	588	3	3	18	567
entitled	entitled	528	2	0	2	526
redeem...	redeem...	439	2	1	0	438
designa...	designa...	433	3	18	21	394
ended	ended	626	3	105	137	384
related	related	811	3	145	299	367
fixed	fixed	453	3	27	68	358
made	made	332	3	14	35	283
called	called	264	3	2	4	258

Uncertainty:

Word	Attribu...	Total O...	Docum...	Wells F...	Citi Ba...	Ban... ↓
risk	risk	1203	3	399	350	454
exposure	exposure	348	3	76	107	165
exposur...	exposur...	186	3	21	82	83
approxi...	approxi...	291	3	29	191	71
indepen...	indepen...	73	3	5	14	54
risks	risks	129	3	37	46	46
believe	believe	59	3	20	1	38
assump...	assump...	105	3	39	29	37
alteration	alteration	33	1	0	0	33
appear	appear	31	1	0	0	31
intangib...	intangib...	43	3	12	2	29
variable	variable	101	3	35	37	29
conting...	conting...	41	3	2	11	28
appearing	appearing	25	1	0	0	25

Analyze Sentiment:

Sources:

https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

https://sraf.nd.edu/textual-analysis/resources/#LM%20Sentiment%20Word%20Lists