GETTING STARTED with the IMDB data

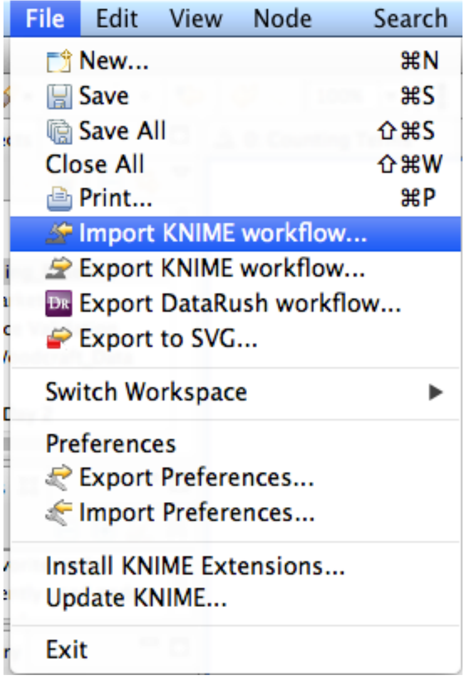
1) download the IMDB data located at http://abbottanalytics.com/data/textminingdata/IMDB%20Review%20Data.zip

UNZIP the data! In the class, we will be using the POS and NEG reviews with 100 reviews in each directory

2) download the baseline workflow, http://abbottanalytics.com/data/textminingdata/IntroTextMining_Baseline.zip

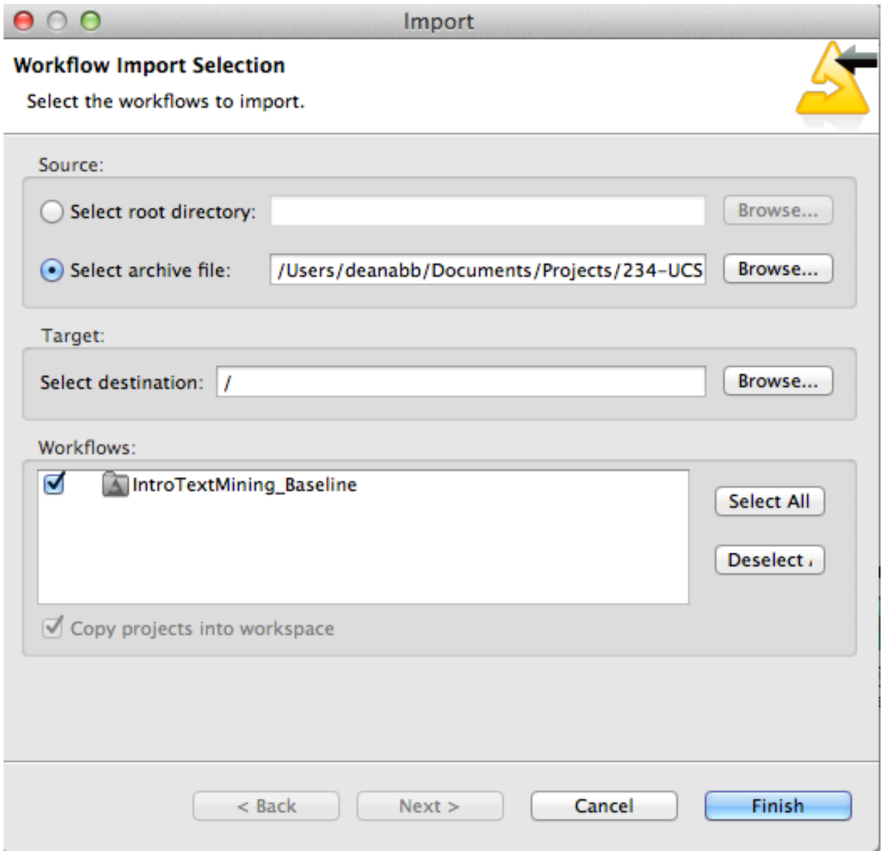
DO NOT UNZIP THIS FILE. You will be importing it as an **archive**

Once you have downloaded the IntroTextMining_Baseline.zip workflow, open the KNIME program, and import the KNIME workflow



After clicking on Import KNIME workflow…, you will see a window that allows you to select the workflow to import. Click on the "Select archive file:" radio button, find the workflow in your file system, and by default, you should see the "IntroTextMining_Baseline" workflow checked. Click "Finish". Leave the "Select destination" directly alone--this just pushes the workflow to your default workspace.

If you UNZIPped the workflow, you can import it as a "root directory" by pointing to the top level directly containing the unzipped workflow.



Import

Workflow Import Selection

Select the workflows to import.

Source:

○ Select root directory:

● Select archive file: /Users/deanabb/Documents/Projects/234-UCS

Target:

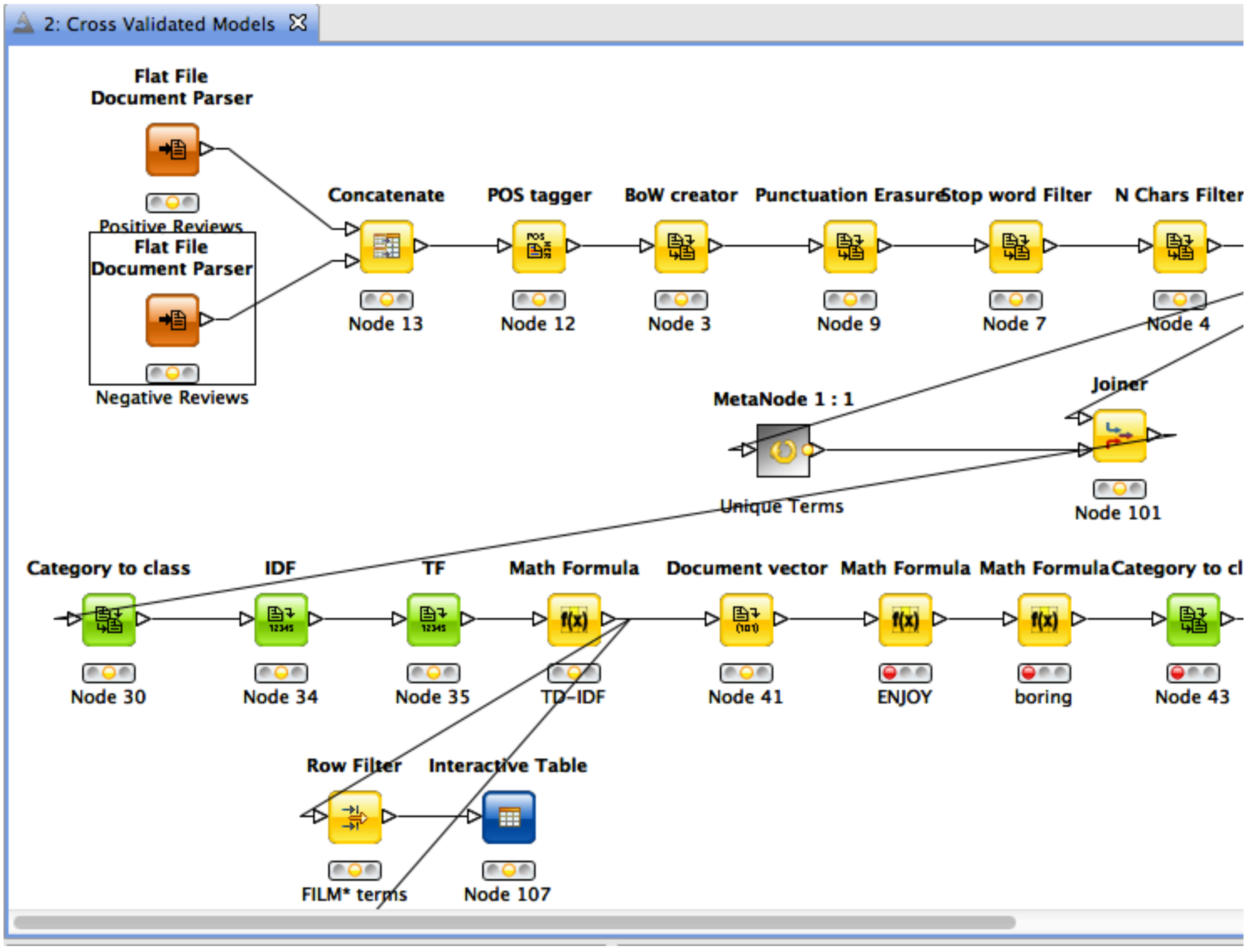
Select destination: /

Workflows:

☑ IntroTextMining_Baseline

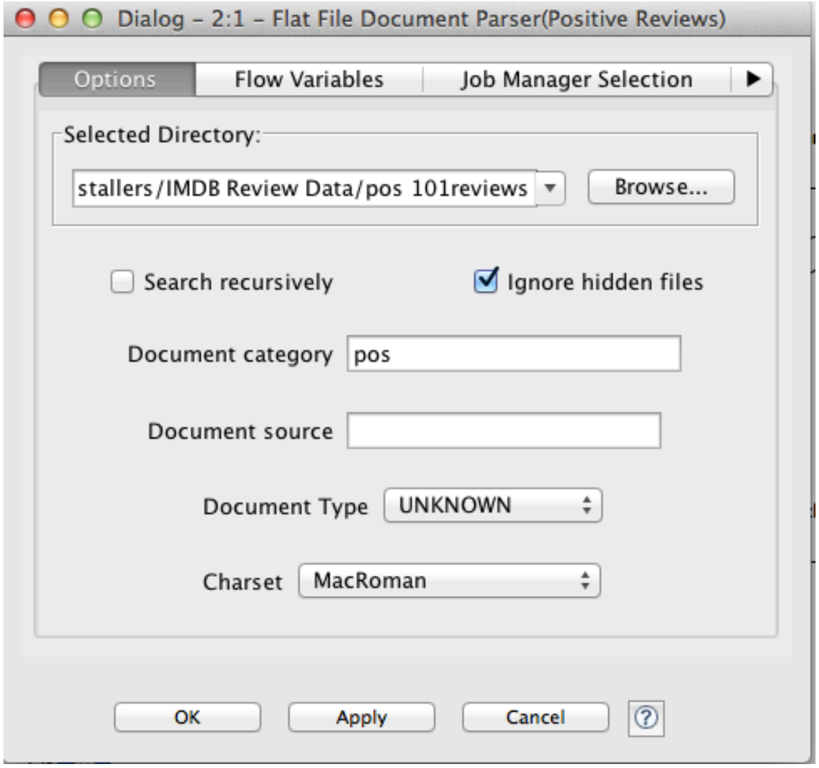
☑ Copy projects into workspace

You will see the workflow in the left side of your main KNIME window in the "Workflow Proejcts" area. Double-click on the workflow. It will load and look something like this:



If you haven't installed the extensions for TextProcessing and Math Expressions, you will see some icons that are generic. These will not execute. Please go back to the installation document and install these two extensions.

The configuration issue is pointing to the right data. Open the icon that says "Positive Reviews". Set the path to the data to point to the IMDB folder "pos 101reviews". Also make sure you set the Document category as "Pos". This merely labels the positive reviews with the label "Pos", indicating these are positive review. Do the same for the negative reviews in folder "neg 101reviews". Click OK.



Dialog – 2:1 – Flat File Document Parser(Positive Reviews)

Selected Directory: stallers/IMDB Review Data/pos 101reviews

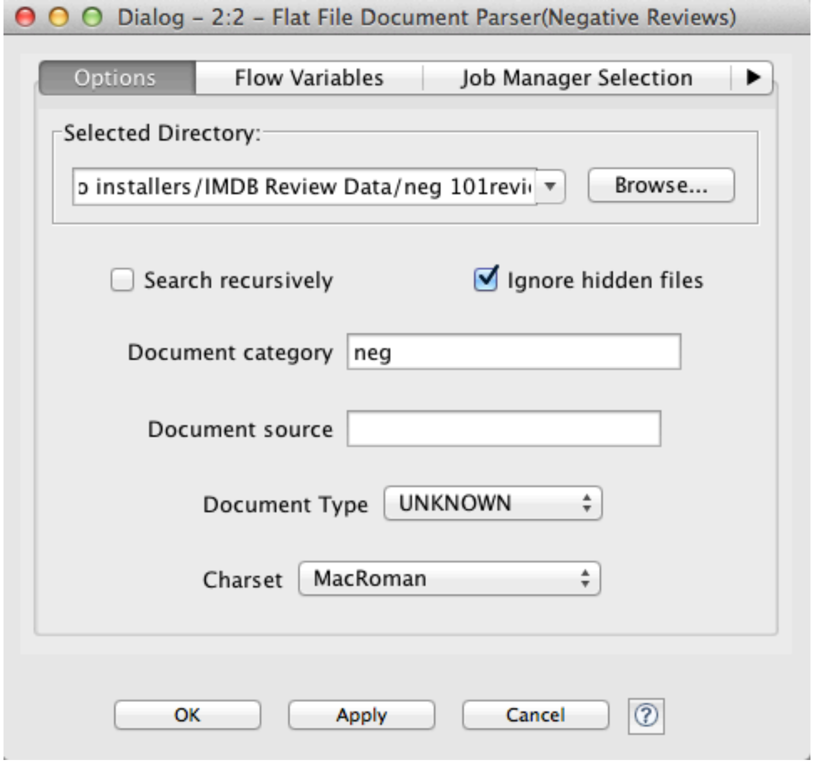
□ Search recursively ☑ Ignore hidden files

Document category: pos

Document source:

Document Type: UNKNOWN

Charset: MacRoman



Dialog – 2:2 – Flat File Document Parser(Negative Reviews)

Selected Directory: ɔ installers/IMDB Review Data/neg 101revi

□ Search recursively ☑ Ignore hidden files

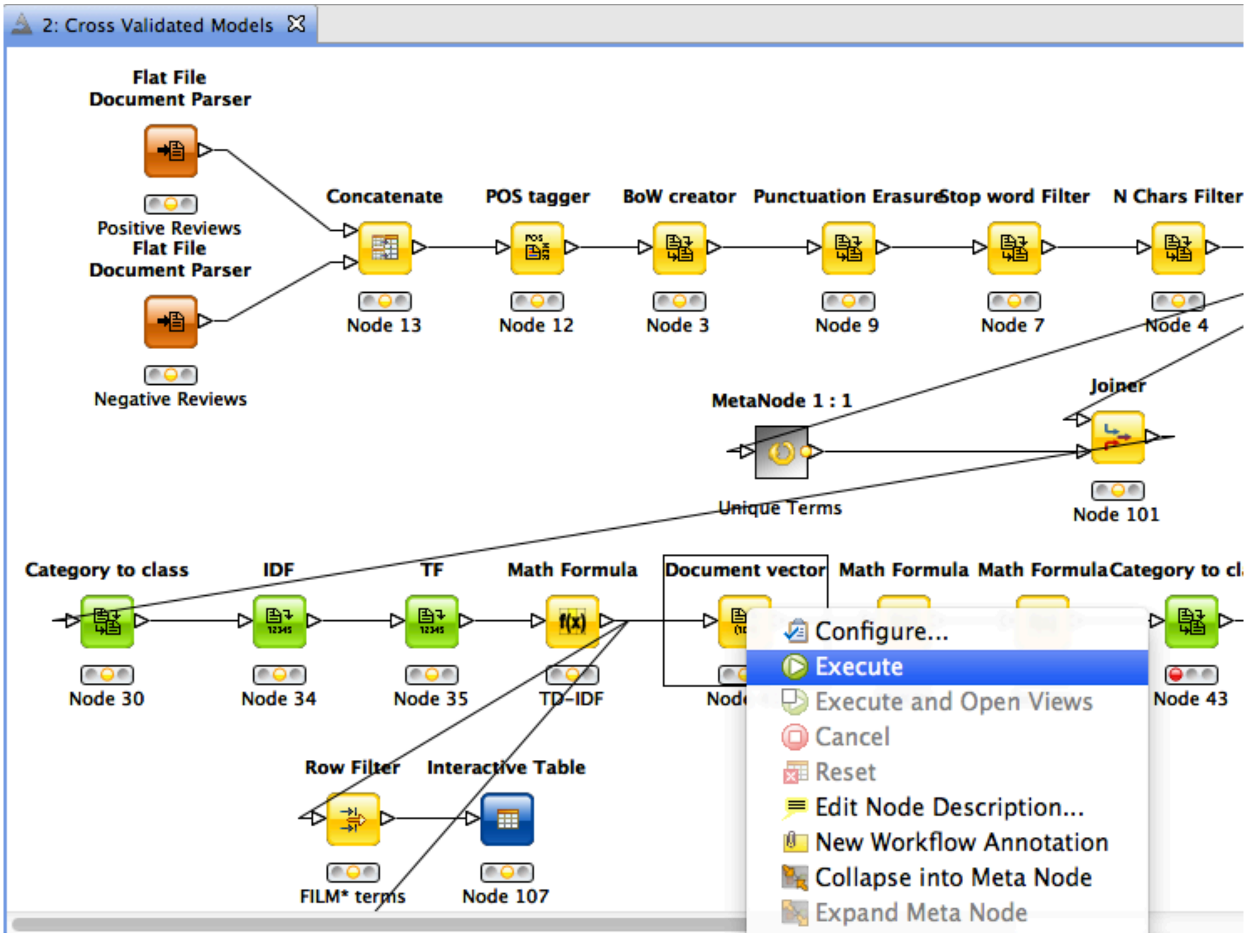
Document category: neg

Document source:

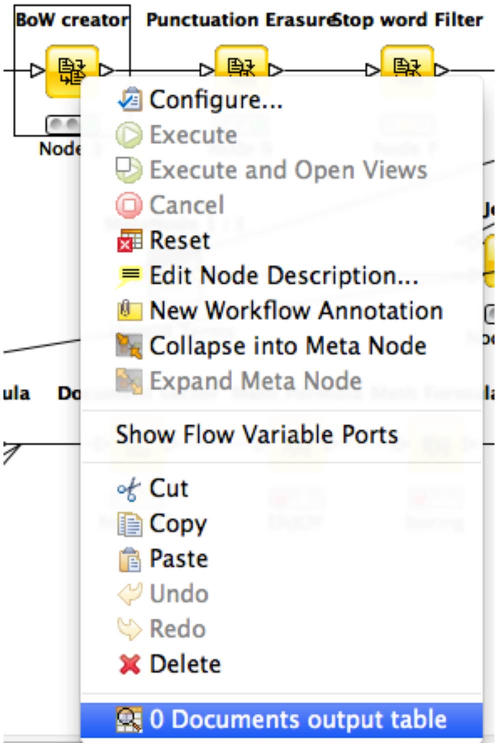
Document Type: UNKNOWN

Charset: MacRoman

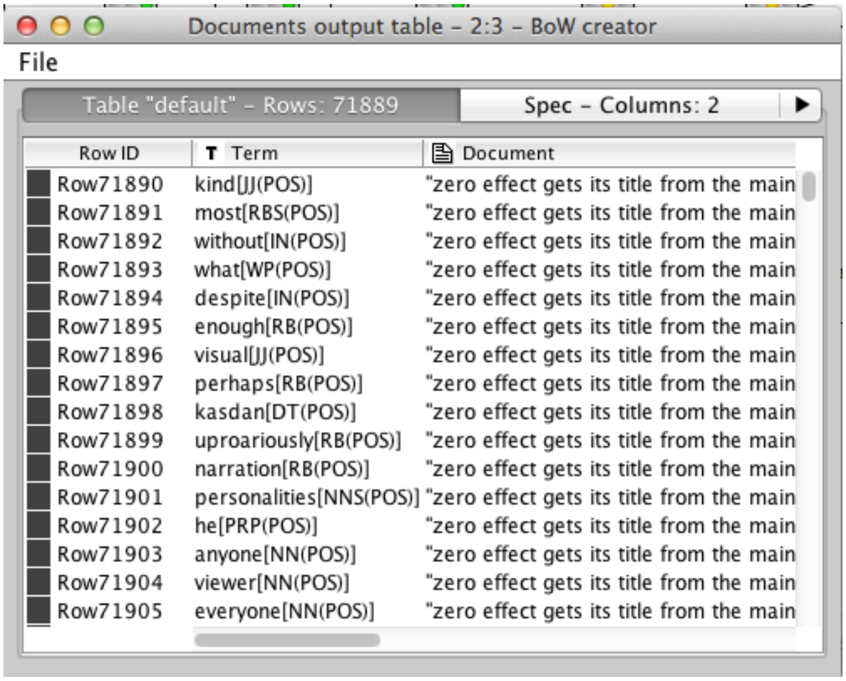
Once you do this, the node should have the yellow dot illuminated (it was red before). We will walk through each step in class, but to test the workflow, right-click on a node that is yellow, such as the "Document Vector" node and select "Execute". Each node as it completes will have its dot turn green.



At any time, you can still click on nodes. In fact, after a node is green, you can right-click on that node and select the bottom entry which is the table view of that node. For example, if you want to look at BoW creator results, you can right-click on that node and select the "output table" option.



In my data, the results look like this. The POS refers to the Part of Speech label. The first entry is an adjective, and that POS has been appended to the term itself.


Documents output table – 2:3 – BoW creator
File
Table "default" – Rows: 71889
Spec – Columns: 2

Row ID	Term	Document
Row71890	kind[JJ(POS)]	"zero effect gets its title from the main
Row71891	most[RBS(POS)]	"zero effect gets its title from the main
Row71892	without[IN(POS)]	"zero effect gets its title from the main
Row71893	what[WP(POS)]	"zero effect gets its title from the main
Row71894	despite[IN(POS)]	"zero effect gets its title from the main
Row71895	enough[RB(POS)]	"zero effect gets its title from the main
Row71896	visual[JJ(POS)]	"zero effect gets its title from the main
Row71897	perhaps[RB(POS)]	"zero effect gets its title from the main
Row71898	kasdan[DT(POS)]	"zero effect gets its title from the main
Row71899	uproariously[RB(POS)]	"zero effect gets its title from the main
Row71900	narration[RB(POS)]	"zero effect gets its title from the main
Row71901	personalities[NNS(POS)]	"zero effect gets its title from the main
Row71902	he[PRP(POS)]	"zero effect gets its title from the main
Row71903	anyone[NN(POS)]	"zero effect gets its title from the main
Row71904	viewer[NN(POS)]	"zero effect gets its title from the main
Row71905	everyone[NN(POS)]	"zero effect gets its title from the main