THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Big Data with ADAMS

Tweet, tweet, tweet

Peter Reutemann

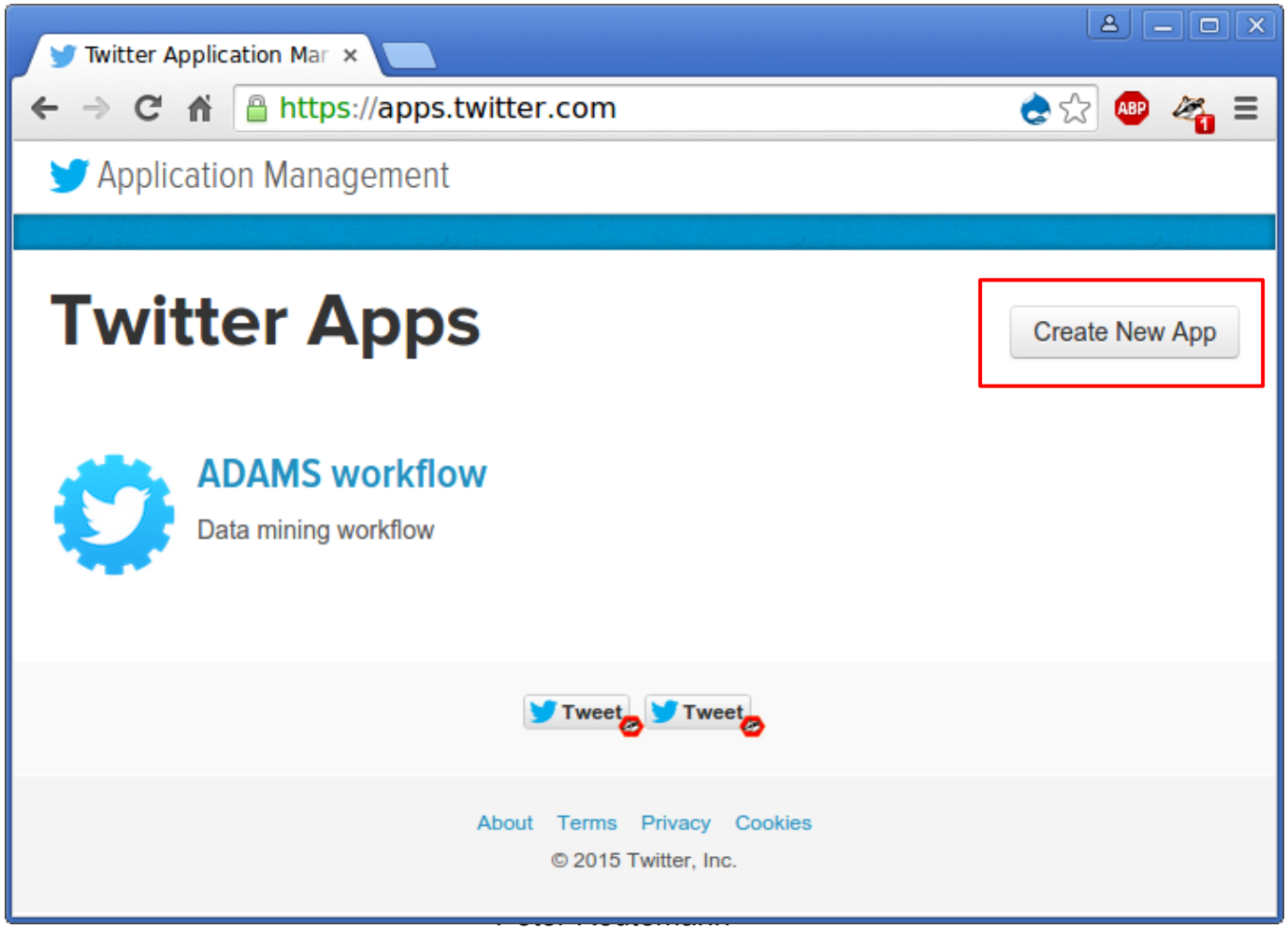
Collecting tweets

- Twitter allows searches of public tweets
- Twitter offers access to tweets in real time
- ADAMS uses twitter4j.org to access Twitter
- Requires setting up an App

  https://apps.twitter.com
- Don't worry, it's not that hard...

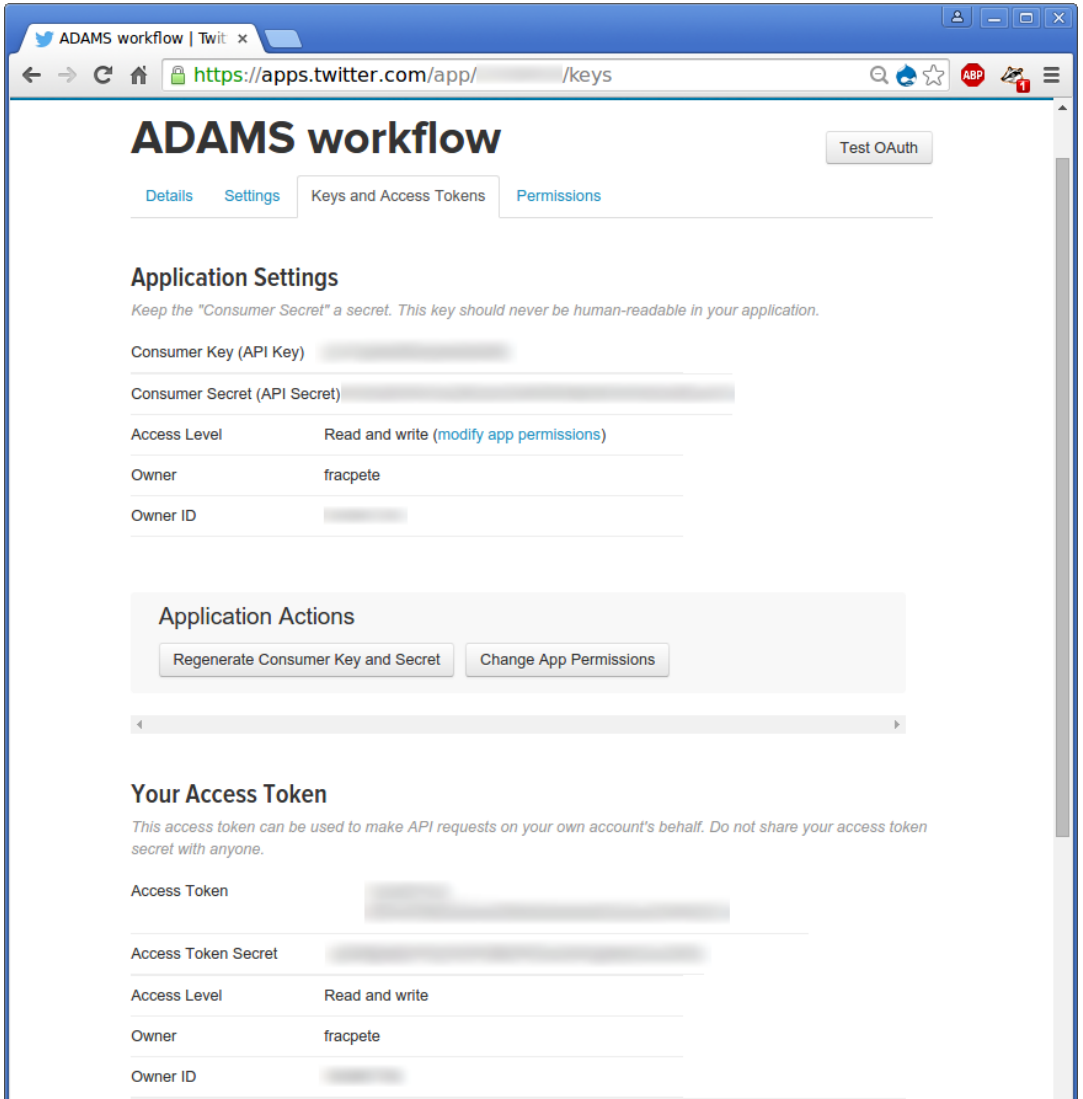
Settings things up

- Create an app



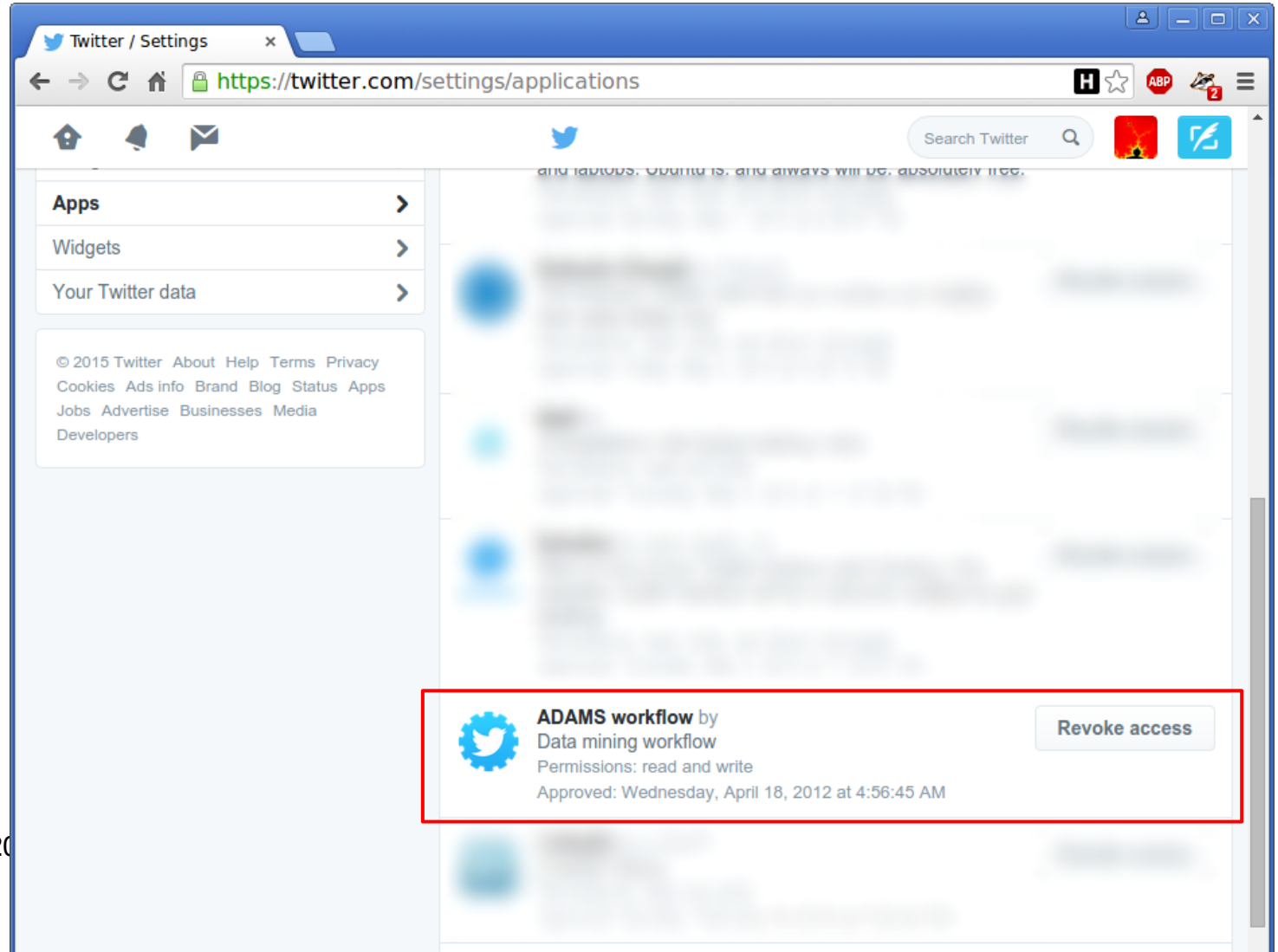
Settings things up (2)

- Set up tokens (consumer and access)



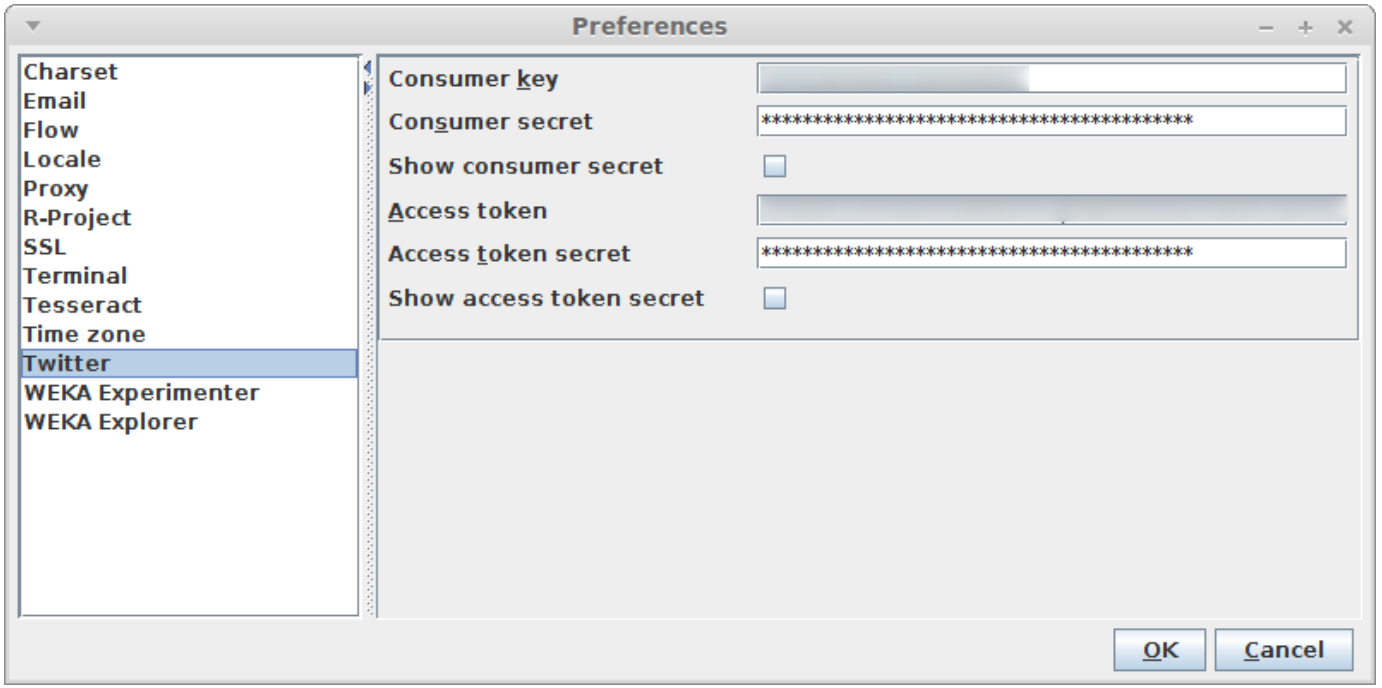
Settings things up (3)

- App should show up in your profile settings



Settings things up (4)

- Finally, fill in Twitter preferences in ADAMS



User queries

- You can query twitter using something like:

  movie -scary :)

  [tweets containing “movie” and “:)” but not “scary”]
- Actors to use
  - TwitterQuery
  - TwitterConverter
- Use case?
  - when looking for specific keywords

User queries

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

connection settings; can override global preference settings

send query

convert tweet into textual format

Flow editor [twitter_1_query.flow -- /home/fracpete/documents/presentations/adams/kuching-

File Edit Run View Window

twitter_1_query

Flow
Example flow for querying the Twitter API.
For query examples, see the following web page:
https://dev.twitter.com/rest/public/search

- TwitterConnection
- TwitterQuery query: "movie -scary :)", max tweets: 200
- TwitterConverter TextConverter (fields: TEXT, separator: \t, quote)
- Count min=-1, max=-1, interval=-1
  - Convert AnyToString
  - StringReplace replace '$' with ':'
  - SetVariable @{count} [REPLACE]
- StringReplace replace '^' with '@{count}'
- Display X:left, Y:top, W:640, H:480, font: Mon

Flow.TwitterQuery

Parameters

Name

adams.flow.source.TwitterQuery

Parameters

- query
  movie -scary :)
- maxTweets
  200

query string, maximum to return

Flow.Display

File Edit View

```
170:"watching #DetroitRockCity. one of my fave\'s. used to have this movie on vhs. :-) @
171:"RT @ASPRO_PANGERANO: 3 hari lg votting ditutup jgn lupa SA1 RICKY HARUN SA9 PANGERA
172:"18 years old and im seeing the peanuts movie with my mom :-)"
173:"So, I\'m watching The Pajama Game play (couldn\'t find the full movie online), and
174:"@IanMcKellen You\'re wonderful as Mr. Holmes! The movie is great and the cast is fa
175:"Congrats to @Joe_Sugg and @Caspar_Lee on the premiere. I heard it went well and I c
176:"Movie :) (at @PlazaTorresLind in Gustavo A. Madero, DF) https://t.co/l49enYfFkR"
177:"Clueless is my all time favorite movie :)"
178:"What\'s the last movie you watched? — افواه و ارانب :D https://t.co/yR98aYNYWq"
179:"RT @justinbieber: Last night was a movie. Love my peoples :)"
180:"@1975INTHEAM Danielle Campbell :) she was in the movie Starstruck and plays Davina
181:"@reynantem Thanks! Learned a lot.\nBTW I highly recommend that movie. Best movie Au
182:"@BenBassOfficial I just heard that you are doing a movie with @H_Combs  I\'m so exc
183:"RT @HeeYun_hojub: Watched the movie he watched, ate what he ate... but what I reall
184:"RT @nhan127: A great movie :D\nhttps://t.co/W4gIGKCdAE"
185:"@blm849 3 of my favorite things->That song, that movie, & the incredibly bright & g
186:"RT @nhan127: A great movie :D\nhttps://t.co/W4gIGKCdAE"
187:"RT @Ashton5SOS: I hope one day we can make a 5SOS movie :).... yeah dat would be co
188:"RT @Kev_h80: Here is my interview featuring the amazing @LadyLunaWolf from yesterda
189:"My ultimate idol @chriscolfer has a new movie! Please support him for another excit
190:"movie recommendation, PLEASE? :)"
191:"@DoveCameron #LivAndMaddie #Descendants I LOVE BOTH SHOWS AND MOVIE AND I LOVE YOU
192:"@Movie_Tv_fans thank u for saying my fanfic is great btw :)"
193:"RT @Laralee1836: The New LAURELS have arrived!  :D  #film #filmmaking #indiefilm #s
194:"RT @NicoleFaria9: https://t.co/4daTr4Ajeu movie :)"
195:"@mainedcm continue writing who knows it will become a full blown book and movie lat
196:"@casseracingnc @DisneyPixar Agreed! I rank Toy Story as my favorite movie of all-ti
197:"RT @tanu_insan: @Gurmeetramrahim #MSG2MSG2inHaryana Thnx papa ji fr giving us a lif
198:"@NickelodeonTV how i feel right now about the new Hey Arnold TV movie :) https://t.
199:"@jumblejim I absolutely love that movie :)"
200:"@KingSargon26 haha this is a new movie coming out soon :)"
```

Listening

- Rather than posting queries, listen to tweets in real time
- Public access to 1% sample of tweets

  “garden hose” vs “fire hose”
- Sample bias?

  http://arxiv.org/abs/1212.1684
- Use case?
  - create tweet archives for repeatable experiments
  - capture “the moment”
- Actor
  - TwitterListener

Listening

retrieve a maximum of 100 tweets

convert tweets to text and display them

convert tweets into spreadsheet format and save to CSV **“archive tweets”**

Flow editor [twitter_2_listen-simple.flow -- /home/fracpete/documents/presentations/adams/k

File Edit Run View Window

twitter_2_listen-simple

Flow
Example flow for listening to the sample stream of Twitter status updates.
Requires actual twitter user credentials to work.

TwitterConnection
TwitterListener 100 status updates
Branch parallel, threads: #cores
display
TwitterConverter TextConverter (fields: CREATED, USER_NAME, T
Display X:left, Y:top, W:640, H:480, font: Monospaced-PLAIN-12
file
TwitterConverter SpreadSheetConverter (fields: ID, CREATED, US
SpreadSheetFileWriter CsvSpreadSheetWriter, file: ${TMP}/twitt

Flow.Branch.display.TwitterConverter

Parameters

Name

adams.flow.transformer.TwitterConverter

Parameters

- converter
  adams.data.twitter.TextConverter -field CREATED -field USER_NAME -field TEXT -separator \\t

conversion to text using only some fields

Flow.Branch.display.Display

File Edit View

```
2015-11-24 15:12:43 新着チケットオークション情報        【新着情報】https://t.co/lpi2kTyNqz イ
2015-11-24 15:12:43 たっちー@オズ余韻→11/25胎盤        RT @yuna_____rw: 11月28日、RADとミスチ
・11月25日の21時までに支払い
・ファミマ支払いできる方
・1Fスタンディング
定価でお譲りします。希望者多数の場合、抽選で選ばせていただきます。
締め切りは今日の22時半までで！…
2015-11-24 15:12:43 Isabele Soares      RT @OgroAmarelo: "Por um mundo onde as pessoas
2015-11-24 15:12:43 cool      @HMUCLAYTON @bruhitszach I hope everything is okay ♥
2015-11-24 15:12:43 parker    SAY THIS OUT LOUD https://t.co/edh4JZ2syR
2015-11-24 15:12:43 ■コスモディアドットコム!!          ■物産展Webマガジン-『コスモディア・ブッサ
2015-11-24 15:12:43 ポチョム菌@うっちー追加 意識高い系ライバー https://t.co/Mz5ZvA6rER
2015-11-24 15:12:43 Fajhria Budi        RT @noellastrna: 😉😉😉 https://t.co/NAAM5swQro
2015-11-24 15:12:43 HMVエソラ池袋        【#超特急×HMV】15日から開始！のHMV限定トレカプレゼン
2015-11-24 15:12:43 #BuyPrettyGirls     bd05324 #MTVStars Demi Lovato ▯ Britney Spears
2015-11-24 15:12:43 adriana de la rosa  @marleen_la olis mana saludos. https://t.co/DfT
2015-11-24 15:12:43 BamJubb   RT @nuttyinlove46:[illegible] 😂#markbam
2015-11-24 15:12:43 Дарья Бахтина       утро добрым не бывает, знаете ли (
2015-11-24 15:12:44 karleigh  RT @ChrisOflyng: Hey! I love you. Why? Cause I said so.
2015-11-24 15:12:44 だいちゃん  RT @mu7mu: 維新W当選か。また都構想言い出すぞ。もうええっちゅうね
2015-11-24 15:12:44 bad for you         RT @bby_laana: Okay 😩▯ https://t.co/3KYR1U6ri4
2015-11-24 15:12:44 [illegible]  RT @_zpeachz: [illegible] 25-29 [illegible]
2015-11-24 15:12:44 3 DIAS 1D!!         @harrypasivo Feliz  Cumpleaños!!! 😊😊 bebe me
2015-11-24 15:12:44 rudy      RT @_sonlar: I miss soccer season so much :(
2015-11-24 15:12:44 Lizeth Rico         RT @SensualGifsES: Las buenas noches deberían s
2015-11-24 15:12:44 Fernando  RT @FVallejoQuotes: Hay que amar con locura, sufrir con l
```

Continuous archiving

- “Fleeting” nature of tweets makes it hard to repeat experiments
- Archival and replay of tweets solves this
- Next flow shows
  - continuous archival
  - compressed CSV spreadsheets
  - one archive per day

Continuous archival

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

create filename based on current day

check “queue” whether new spreadsheet to compress

listen till flow gets stopped

“queue” new spreadsheet to be gzip'ed

create spreadhsheet row and append file

Flow editor [twitter_3_listen-continuous.flow -- /home/fracpete/documents/presentations/adams/kuc

File Edit Run View Window

twitter_3_listen-continuous

- Flow
  - TwitterConnection
  - CallableActors
    - filename — Here the CSV filename is generated
    - Progress — X:left, Y:top, W:640, H:480, font: Monospaced-PLAIN-12
  - HashSetInit — storage: csvfiles
  - Events
    - Cron — 0 * * * * ? — Checking every minute for new CSV files to gzip
      - HashSet — storage: csvfiles
      - SpreadSheetVariableRowIterator — rows: first-last, cols: first-last
      - gzip file
      - remove file from "to-zip" list
  - TwitterListener — unlimited status updates
  - initial filename — @{variable}
  - Count — min=-1, max=-1, interval=1000
  - ConditionalTrigger — min=-1, max=-1, interval=1000
    - CallableSource — filename
    - ConditionalTee — "@{filename}" != "X" — New CSV file started?
      - old filename — @{filename_old} = @{filename} [REPLACE]
      - new filename — @{filename} [REPLACE]
      - queue CSV file for gzipping
        - Variable — @{filename_old}
        - HashSetAdd — storage: csvfiles, col: first
  - TwitterConverter — SpreadSheetConverter (fields: ID, USER_ID, USER_NAME, SC
  - SpreadSheetFileWriter — CsvSpreadSheetWriter, file: @{filename}

Parameters

Name

adams.flow.transformer.TwitterConverter

Parameters

- **converter**
  adams.data.twitter.SpreadSheetConverter -field ID -field USER_ID -field USER_NAME -field SOURCE -field TEXT -field CREATED -field FAVORITED -field RETWEET -field RETWEET_COUNT -field RETWEET_BY_ME -field POSSIBLY_SENSITIVE -field GEO_LATITUDE -field GEO_LONGITUDE -field LANGUAGE_CODE -field PLACE -field PLACE_TYPE -field PLACE_URL -field STREET_ADDRESS -field COUNTRY -field COUNTRY_CODE -field IN_REPLY_TO_STATUS_ID -field IN_REPLY_TO_USER_ID -field RETWEETED_STATUS_ID -field RETWEETED_STATUS_USER_ID -field RETWEETED_STATUS_CREATED -field EXPANDED_URLS -data-row-type adams.data.spreadsheet.DenseDataRow -spreadsheet-type adams.data.spreadsheet.SpreadSheet

collect as much meta-data as possible

Flow.TwitterConverter

Filtering archives

- At time of writing ~3 million tweets per day
- Experiments can operate on subsets to speed things up considerably

  → pre-filter spreadsheet archives
- Actors to use
  -  SpreadSheetFileReader
  -  SpreadSheetRowFilter
  -  SpreadSheetFileWriter

Filtering by country

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

select archive files to filter

create output filename based on input file

leave only rows with value “MY” in column “COUNTRY”

Flow editor [twitter_4_extract_my.flow -- /home/fracpete/documents/presentations/adams/kuching-2

File Edit Run View Window

twitter_4_extract_my

- Flow
  Filters out tweets with country code MY
  - SelectFile directory: ${CWD}
  - SetVariable @{infile} [REPLACE]
  - Tee
    - Convert ReplaceFileExtension
    - SetVariable @{outfile} [REPLACE]
  - SpreadSheetFileReader reader: CsvSpreadSheetReader
  - SpreadSheetRowFilter ByValue (col: COUNTRY_CODE, regexp: MY)
  - SpreadSheetFileWriter CsvSpreadSheetWriter, file: @{outfile}

Parameters

Name

adams.flow.transformer.SpreadSheetFileReader

Parameters

- **reader**
  adams.data.io.input.CsvSpreadSheetReader -data-row-type adams.data.spreadsheet.DenseDataRow -spreadsheet-type adams.data.spreadsheet.SpreadSheet -missing \"\" -chunk-size 10000

read spreadsheet in chunks, limit memory usage

Flow.SpreadSheetFileReader

Filtering by content

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

filter column "TEXT" using a regular expression

Flow editor [twitter_5_extract_tweets_with_emoticons.flow -- /home/fracpete/documents/presentatio

File Edit Run View Window

twitter_5_extra.._with_emoticons

- Flow
  Filters out tweets with emoticons
  - SelectFile directory: ${CWD}
  - SetVariable @{infile} [REPLACE]
  - Tee
    - Convert ReplaceFileExtension
    - SetVariable @{outfile} [REPLACE]
  - SpreadSheetFileReader reader: CsvSpreadSheetReader
  - SpreadSheetRowFilter ByValue (col: TEXT, regexp: .*(😊|😂|😩|😫).*)
  - SpreadSheetFileWriter CsvSpreadSheetWriter, file: @{outfile}

Parameters

Name

adams.flow.transformer.SpreadSheetRowFilter

Parameters

- **finder**
  adams.data.spreadsheet.rowfinder.ByValue -att-index TEXT -reg-exp .*(😊|😂|😩|😫).*

tweets must contain these four emoticons (reg exp)

Flow.SpreadSheetRowFilter

Replaying archives

- Replaying archives is excellent for repeatable experiments
- Use cases
  - further filtering
  - feature extraction
  - plotting
- Actor
  - TweetReplay

New Years “Happiness”

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

plots happy vs sad

select archives to replay

replay current archive, creating “fake” tweets from archived data

create plot containers each minute

count “happy” and “sad” tweets

Flow editor [twitter_6_new_years_happiness.flow -- /home/fracpete/documents/presentations/adams/

File Edit Run View Window

twitter_6_new_years_happiness

Flow
Demonstrates how to replay archived tweets (generated with ADAMS).
Plots the distribution of happy and sad smileys per minute.
CallableActors
SequencePlotter X:left, Y:top, W:800, H:350
initial timestamp @{timestamp}
SelectFile directory: ${CWD} [stops flow if canceled]
SetVariable @{archive} [REPLACE]
replay
TweetReplay replay: SpreadSheetConverterArchive (archive: @{archive})
determine timestamp
TwitterConverter TextConverter (fields: CREATED, separator: \t)
timestamp changed? "X" != "@{timestamp}"
plot? "@{timestamp}" != "" and len("@{timestamp}") > 1
Start
timestamp to double
plot happy
plot sad
reset happy @{happy} = 0 [REPLACE]
reset sad @{sad} = 0 [REPLACE]
SetVariable @{timestamp} [REPLACE]
happy text:"😊" or text:"😂"
IncVariable @{happy}, INTEGER
sad text:"☹" or text:"😭"

Flow.replay.happy

Parameters

Name

adams.flow.control.ConditionalTee

Parameters

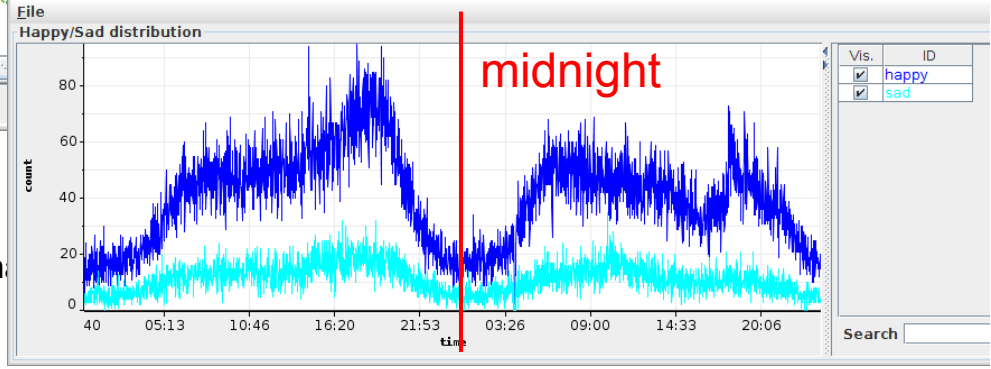
- name
  happy
- condition
  adams.flow.condition.bool.TwitterFilterExpression -expression \"text:\\\"😊\\\" or text:\\\"😂\\\"\"

“conditional” tee, which requires “happy” tweets

Flow.CallableActors.SequencePlotter

File

Happy/Sad distribution



Create dataset

- Build dataset from tweets as basis for predictive model
- Use happy vs sad as the “state” of the tweet (= class attribute)
- Add class attribute column using
  - SpreadSheetInsertColumn

Create dataset

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Flow editor [twitter_7_happiness_csv.flow -- /home/fracpete/documents/presentations/adams/kuching

File Edit Run View Window

twitter_7_happiness_csv

Flow
Demonstrates how to replay archived tweets (generated with ADAMS).
Creates CSV file from archives, labeling happy and sad tweets.
- Start
- select output file
  - SelectFile directory: ${CWD} [stops flow if canceled]
  - SetVariable @{outfile} [REPLACE]
- process archives
  - SelectFile directory: ${CWD} [stops flow if canceled]
  - SetVariable @{archive} [REPLACE]
  - replay
    - TweetReplay replay: SpreadSheetConverterArchive (archive: @{archive}
    - Continue ! text:"😊" or text:"😂" or text:"😞" or text:"😭" allow only happy/sad tweets
    - happy or sad?
      - IfThenElse text:"😊" or text:"😂"
        - then @{state} = happy [REPLACE]
        - else @{state} = sad [REPLACE]
    - TwitterConverter SpreadSheetConverter (fields: ID, CREATED, TEXT, GE
    - SpreadSheetInsertColumn header: 'State', after: last, insert: @{stat
    - SpreadSheetFileWriter CsvSpreadSheetWrit file: @{outfile}

Parameters

Name

adams.flow.control.Continue

Parameters

- **annotations**
  allow only happy/sad tweets
- **condition**
  adams.flow.condition.bool.Not -condition \" adams.flow.condition.bool.TwitterFilterExpression -expression \\\" text:\\\\\\\"😊\\\\\\\" or text:\\\\\\\"😂\\\\\\\" or text:\\\\\\\"😞\\\\\\\" or text:\\\\\\\"😭\\\\\\\"\\\" \"

boolean tweet filter expression

only let happy/sad tweets pass through

determine “state” of tweet set in variable

insert column with “state”

orer [blah.csv -- /home/fracpete/temp]

Associate | Select attributes | Visualize | Forecast
Preprocess | Classify | Cluster

Filter
Choose | AllFilter | Apply

Current relation
Relation: Attributes: 8
Instanc... Sum of weights: 139047

Attributes
All | ... | I... | ...

No.	Name
1	ID
2	CREATED
3	TEXT
4	GEO_LATITUDE
5	GEO_LONGITUDE
6	LANGUAGE_CODE
7	COUNTRY_CODE
8	State

Remove

Selected attribute
Name: GEO_LATITU... Type: ..
Missi... 1... Distinct: Unique: ..

Statistic	Value
Minimum	-42.845
Maximum	65.816
Mean	33.67
StdDev	16.135

Class: State (N... | Visualize All

-42.85 11.49 65.82

Create features

- Using WEKA's StringToWordVector to create features from tweet text
- Apply token cleaners to remove unwanted content
  - users
  - URLs
  - hashtag
  - emoticons

Create features

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

create output filename

remove some attribute

generate features from tweet text

Flow editor [twitter_8_create_arff.flow -- /home/fracpete/documents/presentations/adams/kuching-2

File Edit Run View Window

twitter_8_create_arff

- Flow
  - SelectFile directory: ${CWD} [stops flow if canceled]
  - output file
    - Convert ReplaceFileExtension
    - SetVariable @{outfile} [REPLACE]
  - WekaFileReader automatic (DATASET)
  - WekaFilter-1 RemoveByName -E (TEXT|State) -V
    remove unnecessary attributes
  - WekaFilter StringToWordVector -R first -W 1000 -pru...
    create features
  - Branch parallel, threads: #cores
    - WekaInstancesDisplay X:left, Y:top, W:640, H:480
    - WekaFileWriter @{outfile}

Flow finished.

Parameters

Name

adams.flow.transformer.WekaFilter

Parameters

filter setup

- **annotations**
  create features
- **filter**
  weka.filters.unsupervised.attribute.StringToWordVector -R first -W 1000 -prune-rate -1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -stopwords-handler weka.core.stopwords.Null -M 1 -tokenizer \"weka.core.tokenizers.TwitterNLPTokenizer -use-lower-case -cleaner \\\"weka.core.tokenizers.cleaners.MultiCleaner -cleaner weka.core.tokenizers.cleaners.NormalizeDuplicateChars -cleaner \\\\\\\"weka.core.tokenizers.cleaners.RemoveEmoticons -model /home/fracpete/documents/presentations/adams/kuching-2015/4_twitter/flows/data/AFINN-emoticon-8.txt\\\\\\\" -cleaner weka.core.tokenizers.cleaners.RemoveNonWordCharTokens\\\"\"

Flow.Branch.WekaInstancesDisplay

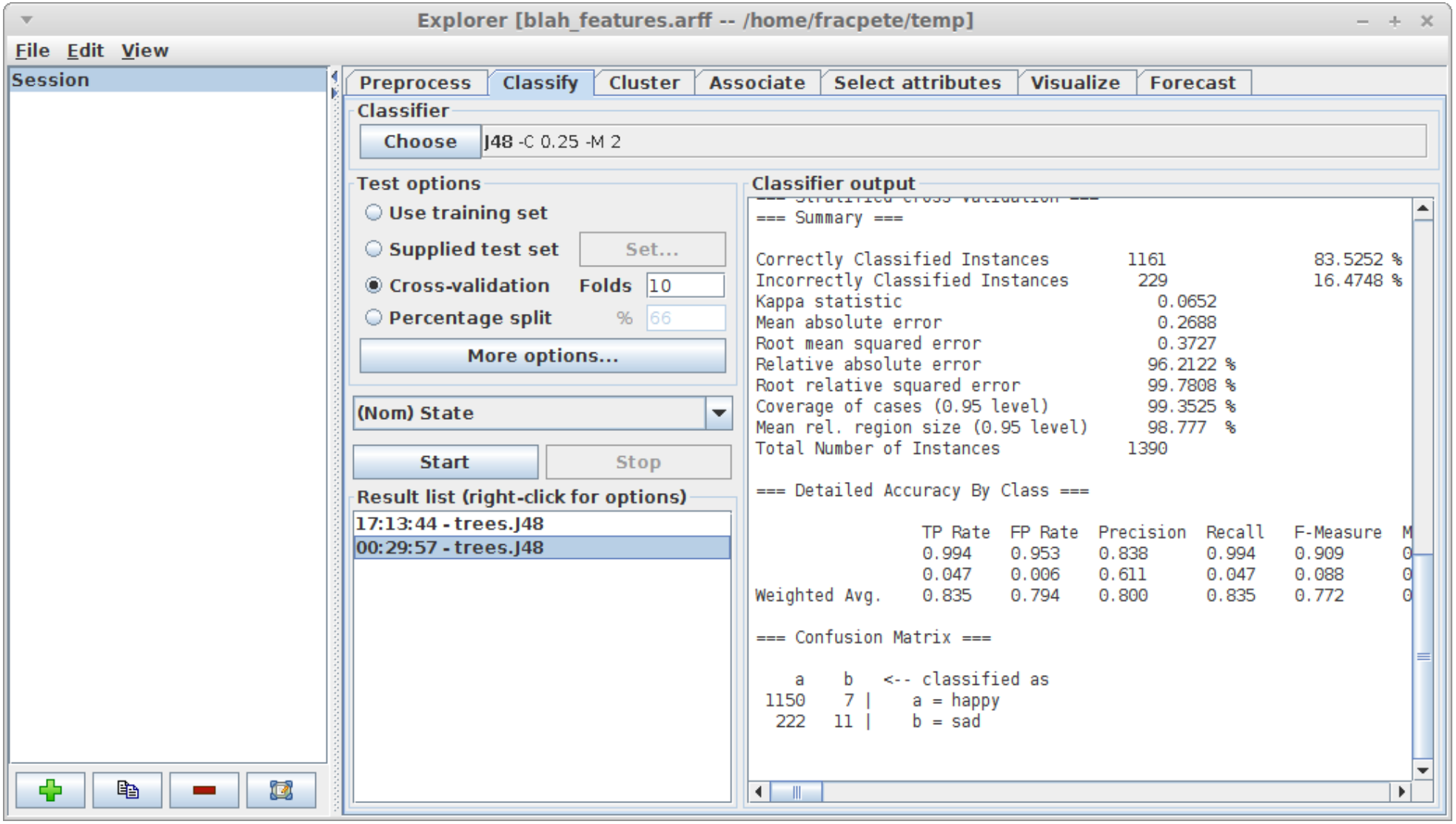
File

Relation: adams-weka.filters.unsupervised.attribute.Remove-V-R3,8-weka.filters.unsu...

No.	1: State Nominal	2: 1 Numeric	3: 10 Numeric	4: 12 Numeric	5: 15 Numeric	6: 1st Numeric	7: 2 Numeric	8: 20 Numeric	9: 2013 Numeric	10: 2014 Numeric	11: 3 Numeric	12: Num
1	happy	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2	happy	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
3	sad	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0
4	happy	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
5	sad	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
6	sad	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
7	happy	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
8	happy	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
9	happy	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
10	happy	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
11	happy	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
12	happy	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0

Create features

- using J48 on 1% percent sub-sample



Label tweet stream

- Build model on training set
  - FilteredClassifier: StringToWordVector and J48
- Predict labels for new tweets coming through
- Actors to use
  - WekaTrainClassifier
  - WekaClassifying

Label tweet stream

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

setup generates features from tweet content

train model on training set

create Weka dataset and add class attribute

set class attribute and output row by row

apply trained model output row with filled in class value

Flow editor [twitter_9_label_tweet_stream.flow -- /home/fracpete/documents/presentations/adams/ku

File Edit Run View Window

twitter_9_label_tweet_stream

- Flow
- CallableActors
  - WekaClassifierSetup FilteredClassifier -F "weka.filters.uns
  - model model
- Start
- train model
  - SelectFile directory: ${CWD} [stops flow if canceled]
  - WekaFileReader automatic (DATASET)
  - WekaClassSelector last, name: .*
  - WekaTrainClassifier WekaClassifierSetup
  - ContainerValuePicker Model [outputs switched]
  - SetStorageValue model
- label tweets
  - TwitterListener 100 status updates
  - TweetReplay replay: SpreadSheetConverterArchive (archive if off-line
  - TwitterConverter SpreadSheetConverter (fields: TEXT)
  - Convert SpreadSheetToWekaInstances
  - WekaFilter Add -T NOM -N Class -L positive,neutral,...
  - WekaClassSelector last, name: .*
  - WekaInstanceBuffer INSTANCES_TO_INSTANCE, in
  - WekaClassifying model
  - Display X:left, Y:top, W:640, H:480, font: Monospaced-P

Parameters

Name

adams.flow.source.WekaClassifierSetup

Parameters

StringToWordVector, tokenizer, cleaners

- **classifier**
  weka.classifiers.meta.FilteredClassifier -F \"weka.filters.unsupervised.attribute.StringToWordVector -R first -W 1000 -prune-rate -1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -stopwords-handler weka.core.stopwords.Null -M 1 -tokenizer \\\"weka.core.tokenizers.TwitterNLPTokenizer -use-lower-case -cleaner \\\\\\\"weka.core.tokenizers.cleaners.MultiCleaner -cleaner weka.core.tokenizers.cleaners.NormalizeDuplicateChars -cleaner weka.core.tokenizers.cleaners.RemoveNonWordCharTokens -cleaner \\\\\\\\\\\\\\\"weka.core.tokenizers.cleaners.RemoveEmoticons -model /home/fracpete/documents/presentations/adams/kuching-2015/4_twitter/flows/data/AFINN-emoticon-8.txt\\\\\\\\\\\\\\\"\\\\\\\"\\\"\" -W weka.classifiers.trees.J48 -- -C 0.25 -M 2

Flow.label tweets.Display

File Edit View

```
co/f3dbGdY8gu',neutral
tofollow #teamfollowback #mustfollow #followall #autofollow',neutral
era policial , melhoor sonho da minha vida <33',neutral
enso que me voy que levantar',neutral

P exploit only @  http://t.co/EsZBTGSPW9',positive
 اذا ريتويت\n◆3-تابع من يسوي رتويت\n◆4-التزم بالشروط لتستفيد \n 268',neutral

m',neutral
top you from being the best person you can be.',neutral
',neutral

AGg3",positive
```

Competition: Twitter

- Build a good classification model
- Use the “sanders.arff” dataset
- Perform cross-validation
- Beat ZeroR: 67.1409 %
- Hints
  - Use FilteredClassifier with StringToWordVector
  - Use TwitterNLPTokenizer
  - Use token cleaners

Questions?

https://adams.cms.waikato.ac.nz/

@TheAdamsFlow