Some more application examples

Thiemo Fetzer

University of Chicago & University of Warwick

May 18, 2017

Plan

Building your own classifier

Helping you code data

Sentiment Analysis

Building your own classifier

- We talked about a wide range of different methods to build classifiers.
- In that process, there are a lot of decisions to be made.
- There is "no classifier" to dominate them all.
- There are appealing features to be considered.
- A lot will involve *trial and error*
- Most common approach taken is that of "ensemble agreement".

Another Minimum Working Example

```r
library(e1071)
happy <- readLines("R/happy.txt")
sad <- readLines("R/sad.txt")
happy_test <- readLines("R/happy_test.txt")
sad_test <- readLines("R/sad_test.txt")

tweet <- c(happy, sad)
tweet_test <- c(happy_test, sad_test)
tweet_all <- c(tweet, tweet_test)
sentiment <- c(rep("happy", length(happy)), rep("sad", length(sad)))
sentiment_test <- c(rep("happy", length(happy_test)), rep("sad", length(sad_test)))
sentiment_all <- as.factor(c(sentiment, sentiment_test))
```

Another Minimum Working Example

```
library(RTextTools)

# naive bayes
mat <- create_matrix(tweet_all, language = "english", removeStopwords = FALSE, removeNumbers = TRUE,
    stemWords = FALSE, tm::weightTfIdf)

mat <- as.matrix(mat)

classifier <- naiveBayes(mat[1:160, ], as.factor(sentiment_all[1:160]))
predicted <- predict(classifier, mat[161:180, ])

predicted

##  [1] sad   happy sad   happy happy sad   happy sad   happy happy sad   sad   sad   sad
## [15] sad   sad   sad   happy happy happy
## Levels: happy sad

table(sentiment_test, predicted)

##               predicted
## sentiment_test happy sad
##          happy     6   4
##          sad       3   7

## better than a coin toss../
recall_accuracy(sentiment_test, predicted)

## [1] 0.65

## better than estimated prior
table(sentiment)

## sentiment
## happy   sad
##    80    80
```

Another Minimum Working Example

```
mat <- create_matrix(tweet_all, language = "english", removeStopwords = FALSE, removeNumbers = TRUE,
    stemWords = FALSE, tm::weightTfIdf)

container <- create_container(mat, as.numeric(sentiment_all), trainSize = 1:160, testSize = 161:180,
    virgin = FALSE)

models <- train_models(container, algorithms = c("MAXENT", "SVM", "BAGGING", "RF", "TREE"))
results <- classify_models(container, models)
table(as.numeric(as.numeric(sentiment_all[161:180])), results[, "FORESTS_LABEL"])

##
##      1  2
##   1 10  0
##   2  1  9

recall_accuracy(as.numeric(as.numeric(sentiment_all[161:180])), results[, "FORESTS_LABEL"])

## [1] 0.95
```

Analytics Post Training/ Prediction

RTextTools provides a range of post training analytics through the `create_analytics` functionality.

`analytics@algorithm_summary` Summary of precision, recall, f-scores, and accuracy sorted by topic code for each algorithm

`analytics@label_summary` Summary of label (e.g. Topic) accuracy

`analytics@document_summary` : Raw summary of all data and scoring

`analytics@ensemble_summary` : Summary of ensemble precision/coverage. Uses the n variable passed into `create_analytics()`

The `@` operator is used to access so-called "slots" of S3 Objects.

Exploring the Analytics Object

```r
# formal tests
analytics <- create_analytics(container, results)

head(analytics@algorithm_summary)
```

```
##   SVM_PRECISION SVM_RECALL SVM_FSCORE BAGGING_PRECISION BAGGING_RECALL BAGGING_FSCORE
## 1          0.91        1.0       0.95              0.91            1.0           0.95
## 2          1.00        0.9       0.95              1.00            0.9           0.95
##   FORESTS_PRECISION FORESTS_RECALL FORESTS_FSCORE TREE_PRECISION TREE_RECALL TREE_FSCORE
## 1              0.91            1.0           0.95              1           1           1
## 2              1.00            0.9           0.95              1           1           1
##   MAXENTROPY_PRECISION MAXENTROPY_RECALL MAXENTROPY_FSCORE
## 1                 0.91               1.0              0.95
## 2                 1.00               0.9              0.95
```

```r
head(analytics@label_summary)
```

```
##   NUM_MANUALLY_CODED NUM_CONSENSUS_CODED NUM_PROBABILITY_CODED PCT_CONSENSUS_CODED
## 1                 10                  11                    11                 110
## 2                 10                   9                     9                  90
##   PCT_PROBABILITY_CODED PCT_CORRECTLY_CODED_CONSENSUS PCT_CORRECTLY_CODED_PROBABILITY
## 1                   110                           100                             100
## 2                    90                            90                              90
```

```r
head(analytics@document_summary)
```

```
##   MAXENTROPY_LABEL MAXENTROPY_PROB SVM_LABEL SVM_PROB BAGGING_LABEL BAGGING_PROB
## 1                1               1         1    0.999             1            1
## 2                1               1         1    0.999             1            1
## 3                1               1         1    0.975             1            1
## 4                1               1         1    0.971             1            1
## 5                1               1         1    0.982             1            1
## 6                1               1         1    0.944             1            1
##   FORESTS_LABEL FORESTS_PROB TREE_LABEL TREE_PROB MANUAL_CODE CONSENSUS_CODE
## 1             1        0.910          1         1           1              1
## 2             1        0.895          1         1           1              1
## 3             1        0.930          1         1           1              1
## 4             1        0.930          1         1           1              1
```

Ensemble Agreement

```
# Ensemble Agreement
analytics@ensemble_summary
```

```
##        n-ENSEMBLE COVERAGE n-ENSEMBLE RECALL
## n >= 1                1.00              0.95
## n >= 2                1.00              0.95
## n >= 3                1.00              0.95
## n >= 4                1.00              0.95
## n >= 5                0.95              1.00
```

Cross Validation

```
# Cross Validation
N <- 3
cross_SVM <- cross_validate(container, N, "SVM")

## Fold 1 Out of Sample Accuracy = 0.982
## Fold 2 Out of Sample Accuracy = 0.966
## Fold 3 Out of Sample Accuracy = 0.954

cross_MAXENT <- cross_validate(container, N, "MAXENT")
```

Plan

Building your own classifier

Helping you code data

Sentiment Analysis

Using Classifiers to (help) code data

- I want to illustrate another use for classifiers to code data
- Asset declarations of politicians or disclosure often times changes format.
- Classification of types is something that could be done manually, but it also is super scalable for machine learning...
- You can save a lot of RA time with that...
- Sometimes, the data you are working with already provides the training data you need.

Asset Declarations of Brasilian Politicians

Asset delarations available from TSE
`http://divulgacandcontas.tse.jus.br`.

Asset Declarations of Brasilian Politicians

Declaração de bens apresentada à Justiça Eleitoral (2016)

Tipo de bem	Descrição do bem	Valor (R$)
Outros bens imóveis	FAZENDA SÃO FRANCISCO - GUAPIAÇU/SP	140.000,00
Loja	100% CAPITAL SOCIAL DA ALEXANDER RIO PRETO CONFECCÇÕES	18.000,00
Loja	100% CAPITAL - CONSTRUTORA BEM BRASIL - GUAPIAÇU	80.000,00
Terreno	01 TERRENO RESIDENCIAL BEM BRASIL EM GUAPIAÇU/SP	3.800,00
Loja	100% DO CAPITAL SOCIAL DO SUPERMERCADO BEM BRASIL - GUAPIAÇU/SP	150.000,00
Terreno	RESIDENCIAL DAMHA IV - SÃO JOSÉ DO RIO PRETO	107.485,12
Terreno	03 TERRENOS NO RESIDENCIAL BEM BRASIL - QUADRA 04	3.000,00
Loja	100% DO CAPITAL SOCIAL - EMPRESA BEM BRASIL PRODUTOS AGROPECUARIOS LTDA - GUAPIAÇU/SP	75.000,00
Veículo automotor terrestre: caminhão, automóvel, moto, etc.	VEICULO MARCA HONDA - CITY	43.283,20

Asset delarations available from TSE

`http://divulgacandcontas.tse.jus.br`.

Asset Declarations of Brasilian Politicians

Cristovam Buarque (2006)

Dados pessoais do candidato

Nome completo:	Cristovam Ricardo Calvacanti Buarque
CPF:	223.641.291-68 *
Data de nascimento:	20/02/1944
Idade ao final de 2006:	62
Município de nascimento:	Recife /PE
Nacionalidade:	Brasileira
Município de residência:	Brasília
Sexo:	Masculino
Estado Civil:	Casado(A)
Grau de Instrução:	Superior Completo
Ocupação principal declarada:	Professor De Ensino Superior

* Saiba como checar o CPF dos políticos e sua situação fiscal

Dados eleitorais do candidato

Cargo disputado:	Presidente
Nome na urna:	Cristovam Buarque
Número eleitoral:	12
Nome do partido:	Partido Democrático Trabalhista
Sigla/ número do partido:	PDT /12
Nome do vice:	Jefferson Peres
Partido do vice:	PDT /12

Apuração de votos

1º turno	2.538.844 votos	2,64%

Situação eleitoral: Não eleito

Asset delarations available from TSE `http://divulgacandcontas.tse.jus.br`.

Asset Declarations of Brasilian Politicians

Declaração de bens apresentada à Justiça Eleitoral

Descrição do bem	Valor do bem
Apartamento 22 Local. No 2º Andar. Ou 3. Pavimento Do Cond. Ed. Del Nero, Situado Na R. Vanderlei Nr. 527, No 19. Subdistrito Pe	R$ 160.000,00
Apartamento Sqs 215, Bloco K, Apto 603 - Brasília (Df), Adquirido Em 1980 Pelo Sfh	R$ 125.000,00
Banco Do Brasil Agencia 3603 Conta 375.482-0	R$ 6.492,13
Carro Volkswagen Sedan 1983, Adquirido Em 1994, Ba9628	R$ 2.800,00
Linha Telefônica Instalada Na Residência Do Declarante, Nº 32721604, Adquirida Em 1985	R$ 904,00
Meia Parte Do Apartamento Na Av. 17 De Agosto, 301 Adquirido Em 1987 Pela Família Da Pessoa, Onde Vive Uma Cunhada.	R$ 14.711,00
Sala 1015 Da Quadra 02 Bloco D Modulo B Sc/Norte, Centro Empresarial Encol, Adquirida De Tatiana De Sausa Dualibe, Cpf 334061381	R$ 21.883,40
Sala Comercial No Mo Lote 02 Da Quadra 01 Do Sau/Sul, 7.0 Andar Nº 711, Adquirida De Gfs Software E Consultoria Ltda, Cnpj 24692	R$ 90.000,00
Saldo No Banco Brasil S.A Agencia 3603-X, Conta 375482-0	R$ 2.176,86
Vaga De Garagem Situada No Segundo Subsolo Do Sau Quadra 01 Lote 02, Adquirida De Construtora Lider Ltda, Cnpj 17.429.010/0007-3	R$ 22.000,00
Terreno Urbano Em Caldas Nova - Adquirido Em 1983 - Goias	R$ 5.964,00
Telefone Nº 33499529	R$ 904,00
Saldo No Banco National West Nr. 54000645	R$ 1.043,70
Saldo No Banco Do Brasil S.A. Agencia 2636-0 Conta 9895-7	R$ 1.248,65
Saldo No Banco Do Brasil	R$ 16.436,96
Saldo No Banco Interamericano De Desenvolvimento - Credit Union Washington Dc, Onde O Declarante Trabalhou No Periodo De 1973-19	R$ 19.880,00
Sala Comercial Localizada Na Scln 213 Lote 04 Nr. 101, Adquirida De Talento Engenharia Ltda, Cnpj 04422795/0001-87, Por R$ 68000	R$ 68.000,00
Patrimonio Em Obras De Artes E Livros	R$ 180.200,00
Lnha Telefonica Nº 32682505, Em Recife, Adquirida Em 1973 Para Uso Na Residência Da Mãe Do Declarante.	R$ 1.000,00
Linha Telefônica Instalada Na Residência Do Declarante, Em Brasília, 32731730 Adquirida Em 1979	R$ 904,00
Dois Lotes Loteamento S. Antonio - Df, Fora De Brasília, Adquirido Em 1981, Area Total De 4,5 Ha. Doado Junto A Fundação Educaci	R$ 6.700,00

Asset delarations available from TSE
`http://divulgacandcontas.tse.jus.br`.

Asset Declarations of Brasilian Politicians

	DS_TIPO_BEM_CANDIDATO	N	TYPE
1	ne	883060	NOTHING
2	apartamento	33877	REAL ESTATE
3	casa	184618	REAL ESTATE
4	terra nua	33058	REAL ESTATE
5	outros bens imÃ³veis	51954	REAL ESTATE
6	veÃculo automotor terrestre c	331267	CAR
9	depÃ³sito bancÃ¡rio em conta c	44363	FINANCIAL ASSETS
11	terreno	133371	REAL ESTATE
12	outras participaÃ§Ãµes societÃ	11870	FINANCIAL ASSETS
13	quotas ou quinhÃµes de capital	44972	FINANCIAL ASSETS
14	outros fundos	9124	FINANCIAL ASSETS
16	aplicaÃ§Ã£o de renda fixa cdb r	14504	FINANCIAL ASSETS
24	prÃ©dio comercial	10405	REAL ESTATE
25	crÃ©dito decorrente de alienaÃ	464	FINANCIAL ASSETS
26	ouro ativo financeiro	660	FINANCIAL ASSETS

Asset delarations available from TSE
`http://divulgacandcontas.tse.jus.br`.

Looping over Sparsity Measure

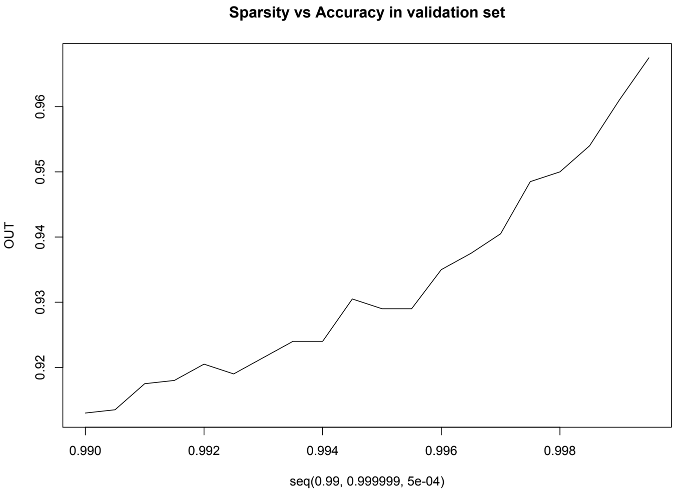
```
OUT<-NULL
k = 1
for(i in c(0.999,0.9995,0.9999,.99995,0.99999,0.999999)) {
cat(i, " ")
DOC<-create_matrix(c(BEM.TRAIN[,paste(BEMDETAIL,sep=" ")]),removeStopwords=FALSE,
                   removeNumbers=TRUE,stemWords=FALSE,removePunctuation=TRUE,removeSparseTerms=i)

DOCCONT<-create_container(DOC,BEM.TRAIN$TYPENUM, trainSize=1:(nrow(BEM.TRAIN)-testsize),testSize=(nrow(BEM.TRA
MOD <- train_models(DOCCONT, algorithms=c("SVM","MAXENT"))
RES <- classify_models(DOCCONT, MOD)
analytics <- create_analytics(DOCCONT, RES)
res<-data.table(analytics@document_summary)

VALID<-cbind(BEM.TRAIN[validation==1],res)

OUT[[k]] <- sum(diag(3) * table(VALID$CONSENSUS_CODE,VALID$TYPE))/nrow(VALID)
k = k+1
}
```

Trade-Off Sparsity vs Accuracy



Plan

Building your own classifier

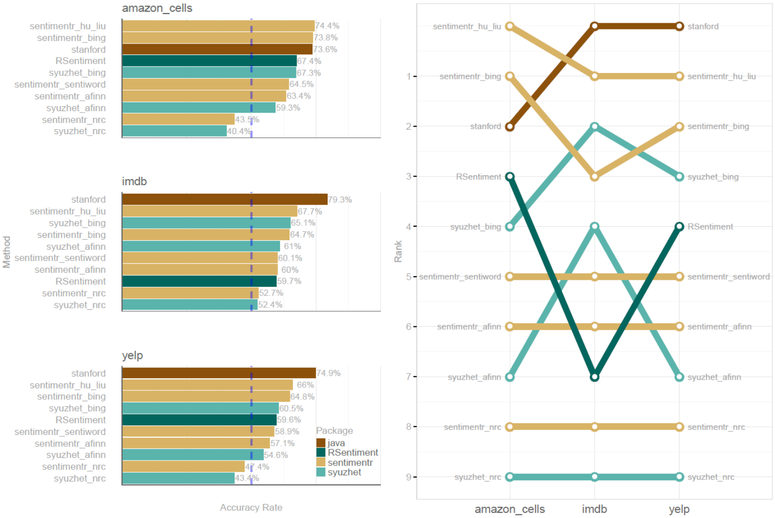
Helping you code data

Sentiment Analysis

Pre-Produced Packages for Sentiment Analysis

- There exist a range of packages in R to do sentiment analysis.
- Most packages work of sentiment dictionaries that simply perform a lookup exercise, nothing fancy or Bayesian at all.
- For large scaling, these approaches may achieve reasonable accuracy.
- We can think of annotated dictionaries as being vocabulary that have been extracted from a training set - i.e. they are features that have a discriminative Bayes score.
- The packages are robust to non-overlapping vocabulary: if the text you are classifying contains no features, then the priors are used for classification.
- It could be that the priors are bad though...

Performance of pre-produced packages for Sentiment Analysis



Sourcing Twitter data: Individual user level

```
library(twitteR)
# setup_twitter_oauth(consumer_key, consumer_secret, access_token=NULL, access_secret=NULL)
set.seed(12122016)
tw.user <- userTimeline("realDonaldTrump", n = 3200)
tw.user.df <- data.table(twListToDF(tw.user))

save(tw.user.df, file = "../../Data/trumpstweets.rdata")
```

Trump Tweets

```
load(file = "../../Data/trumpstweets.rdata")

head(tw.user.df$text)

## [1] "With millions of dollars of negative and phony ads against me by the establishment, my numbers contin
## [2] ".@EWErickson got fired like a dog from RedState\nand now he is the one leading opposition against me."
## [3] "Senator @LindseyGrahamSC made horrible statements about @SenTedCruz  and then he endorsed him. No won
## [4] "Lyin' Ted Cruz lost all five races on Tuesday-and he was just given the jinx - a Lindsey Graham endor
## [5] "Join us in Salt Lake City, Utah- tonight!\n#MakeAmericaGreatAgain #Trump2016\nhttps://t.co/1cJ7OFbQiz"
## [6] "Hillary Clinton has been involved in corruption for most of her professional life!"
```

Cleaning Tweets

```
library(quanteda)
load(file = "../../Data/trumpstweets.rdata")
# remove retweet entities
tw.user.df$text <- gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", tw.user.df$text)
# remove at people
tw.user.df$text <- gsub("@\\w+", "", tw.user.df$text)
# remove punctuation
tw.user.df$text <- gsub("[[:punct:]]", "", tw.user.df$text)
# remove numbers
tw.user.df$text <- gsub("[[:digit:]]", "", tw.user.df$text)
# remove html links
tw.user.df$text <- gsub("http\\w+", "", tw.user.df$text)
# remove unnecessary spaces
tw.user.df$text <- gsub("[ \t]{2,}", "", tw.user.df$text)
tw.user.df$text <- gsub("^\\s+|\\s+$", "", tw.user.df$text)
tw.user.df <- tw.user.df[!is.na(text)]
head(tw.user.df$text)

## [1] "With millions of dollars of negative and phony ads against me by the establishment my numbers continue
## [2] "got fired like a dog from RedState\nand now he is the one leading opposition against me"
## [3] "Senatormade horrible statements aboutand then he endorsed him No wonder nobody trusts politicians"
## [4] "Lyin Ted Cruz lost all five races on Tuesdayand he was just given the jinxa Lindsey Graham endorsement
## [5] "Join us in Salt Lake City Utah tonight\nMakeAmericaGreatAgain Trump"
## [6] "Hillary Clinton has been involved in corruption for most of her professional life"

# build dfm
trump.dfm1 <- dfm(tw.user.df$text)
trump.dfm1

## Document-feature matrix of: 2,328 documents, 5,341 features (99.7% sparse).
```

Using quanteda to clean tweets

```
library(quanteda)
library(operator.tools)
load(file = "../../Data/trumpstweets.rdata")

trump <- corpus(tw.user.df$text, docvars = tw.user.df[, names(tw.user.df) %!in% "text", with = F])

trump.dfm2 <- dfm(trump, removeTwitter = TRUE)
trump.dfm2
```

```
## Document-feature matrix of: 2,328 documents, 6,250 features (99.7% sparse).
```

NRC accessible through

Home | About Saif | Research | Publications | Word Association Lexicons | Invited Talks | Contact

NRC Word-Emotion Association Lexicon (aka EmoLex)

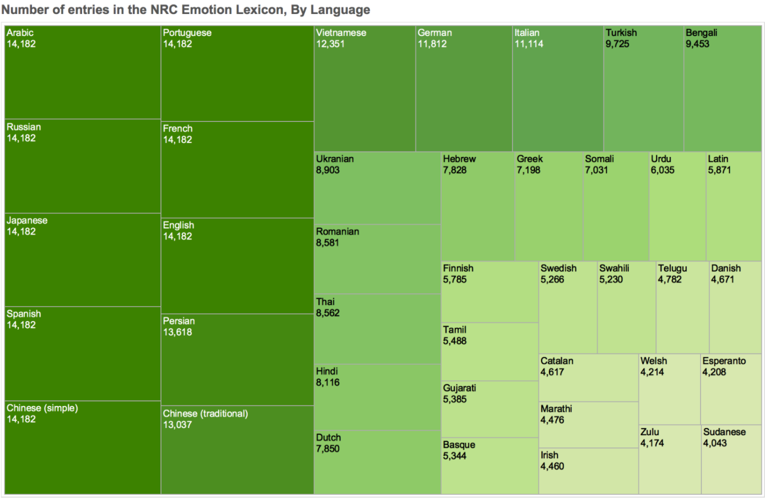
The NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The annotations were manually done by crowdsourcing.

Email: saif.mohammad@nrc-cnrc.gc.ca

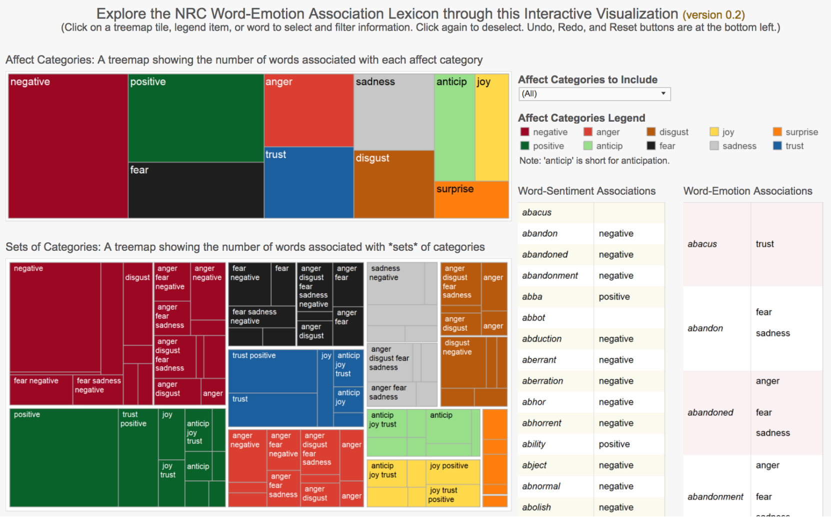
Follow @SaifMMohammad

Association Lexicon	Version	# of Terms	Categories	Association Scores	Method of Creation	Papers
Word-Emotion and Word-Sentiment Association Lexicon						
NRC Word-Emotion Association Lexicon (also called EmoLex) README	0.92 (2010)	14,182 unigrams (words)	**sentiments:** negative, positive **emotions:** anger, anticipation, disgust, fear, joy, sadness, surprise, trust	0 (not associated) or 1 (associated)	**Manual:** By crowdsourcing on Mechanical Turk. **Domain:** General	**Crowdsourcing a Word-Emotion Association Lexicon**, Saif Mohammad and Peter Turney, *Computational Intelligence*, 29 (3), 436-465, 2013. Paper (pdf) BibTeX **Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon**, Saif Mohammad and Peter Turney, In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, June 2010, LA, California. Abstract Paper (pdf) Presentation
		~25,000 senses*		not associated, weakly, moderately, or strongly associated		

NRC accessible through



NRC accessible through



MPQA Subjectivity Lexicon

subjclueslen1-HLTEMNLP05.tff

File Path : ~/Downloads/subjectivity_clues_hltemnlp05/subjclueslen1-HLTEMNLP05.tff

subjclueslen1-HLTEMNLP05.tff

```
type=weaksubj len=1 word1=abandoned pos1=adj stemmed1=n priorpolarity=negative
type=weaksubj len=1 word1=abandonment pos1=noun stemmed1=n priorpolarity=negative
type=weaksubj len=1 word1=abandon pos1=verb stemmed1=y priorpolarity=negative
type=strongsubj len=1 word1=abase pos1=verb stemmed1=y priorpolarity=negative
type=strongsubj len=1 word1=abasement pos1=anypos stemmed1=y priorpolarity=negative
type=strongsubj len=1 word1=abash pos1=verb stemmed1=y priorpolarity=negative
type=weaksubj len=1 word1=abate pos1=verb stemmed1=y priorpolarity=negative
type=weaksubj len=1 word1=abdicate pos1=verb stemmed1=y priorpolarity=negative
type=strongsubj len=1 word1=aberration pos1=adj stemmed1=n priorpolarity=negative
type=strongsubj len=1 word1=aberration pos1=noun stemmed1=n priorpolarity=negative
type=strongsubj len=1 word1=abhor pos1=anypos stemmed1=y priorpolarity=negative
type=strongsubj len=1 word1=abhor pos1=verb stemmed1=y priorpolarity=negative
type=strongsubj len=1 word1=abhorred pos1=adj stemmed1=n priorpolarity=negative
type=strongsubj len=1 word1=abhorrence pos1=noun stemmed1=n priorpolarity=negative
type=strongsubj len=1 word1=abhorrent pos1=adj stemmed1=n priorpolarity=negative
type=strongsubj len=1 word1=abhorrently pos1=anypos stemmed1=n priorpolarity=negative
type=strongsubj len=1 word1=abhors pos1=adj stemmed1=n priorpolarity=negative
type=strongsubj len=1 word1=abhors pos1=noun stemmed1=n priorpolarity=negative
```

Reading in the MPQA Lexicon

```
MPQA<-data.table(read.csv2(file="R/subjclueslen1-HLTEMNLP05.tff",sep=" "))
MPQA[, priorpolarity := str_extract(priorpolarity, "([a-z]+)$") ]
MPQA[, word1 := str_extract(word1, "([a-z]+)$") ]
MPQA[, pos1 := str_extract(pos1, "([a-z]+)$") ]
MPQA[, stemmed1 := str_extract(stemmed1, "([a-z]+)$") ]
MPQA[, type := str_extract(type, "([a-z]+)$") ]
MPQA<-MPQA[priorpolarity %in% c("negative","positive","neutral")]

head(MPQA)

##          type   len       word1   pos1 stemmed1 priorpolarity
## 1:   weaksubj len=1   abandoned    adj        n      negative
## 2:   weaksubj len=1 abandonment   noun        n      negative
## 3:   weaksubj len=1     abandon   verb        y      negative
## 4: strongsubj len=1       abase   verb        y      negative
## 5: strongsubj len=1   abasement anypos        y      negative
## 6: strongsubj len=1       abash   verb        y      negative
```

Sentiment Analysis Lexicons

Sentiment Analysis (Opinion Mining) lexicons

- MPQA Subjectivity Lexicon
- Bing Liu and Minqing Hu Sentiment Lexicon
- SentiWordNet (Included in NLTK)
- VADER Sentiment Lexicon
- SenticNet
- LIWC (not free)
- Harvard Inquirer
- ANEW