BUSINESS ANALYTICS CLUB

Workshop Series 11.14

Sentiment Analysis:
Yelp & IMDB Reviews

Janet Ye
Shantanu Joshi

BAC

Set Up

- Download the zip file from bit.ly/bacdata "Sentiment Analysis" folder. In the zip file, you'll find:
    - Weka installer
        - PC: weka-3-6-13 jre.exe
        - Mac: Weka-3-6-12-oracle-jvm.app
    - Installation Guide
    - Data set

Learning Objective

1. Purpose/reasoning behind text analysis
2. Learn about Sentiment Analysis, NLP, and why software has difficulty working with language
3. Apply machine learning to language
4. Use Weka for implementing Sentiment Analysis
5. Clean and apply knowledge to Yelp & IMDB Datasets

Why text?

Information Retrieval: find documents from user queries, CRM, Medical records, etc…

Behavior/Communication: most communication online is via text, essential to better understand it for behavior

Text is particularly difficult to represent in software, and language constructs are even harder!

Text Classification Tasks

- Spam detection – email filters
- Sentiment analysis – positive/negative reviews
- New Stories – predicting stock market moves
- Social Media – understanding online interaction
- Search – natural language processor for searching
- CRM – classifying customer or user comments

Text is unstructured

- Documents vary in length
- Linguistic structure works well with humans terribly with computers
- Languages have many rules and many exceptions
    - Future tense usually has will+verb
    - Except in simple present tense where context can reveal future tense

“My train will arrive on Monday” = “My train arrives on Monday”

Text is dirty

- Imagine any tweet ever
  - Grammatically incorrect sentences
  - Spelling errors
  - Lack of punctuation

- Context is important

  "I hate classes, abhor clubs, and despise professors, but overall I love Stern"

What do we do with text?

Take a set of documents, transform the document into feature-vectors, and represent them in an individualistic way.

Terminology

Document: one instance of text data

Corpus: a collection of documents

Token: an individual term in a document (sometimes called a word)

Bag of Words

A representation model/framework:

- Treat each document as a collection of words
- Ignore punctuation, word order, sentence structure, etc.

Each instance is one document

Each attribute is one token from the corpus

For each document “1” if token is present and “0” otherwise

Let’s see an example.

Bag of Words Example

d1: jazz music has a swing rhythm

d2: swing is hard to explain

d3: swing rhythm is a natural rhythm

	a	explain	hard	has	is	jazz	music	natural	rhythm	swing	to
d1	1	0	0	1	0	1	1	0	1	1	0
d2	0	1	1	0	1	0	0	0	0	1	1
d3	1	0	0	0	1	0	0	1	1	1	0

Bag of Words Example

Benefits:

- Easy to generate
- Straightforward representation
- Many models use Bag of Words as a baseline and build on top if it

Expanding Bag of Words

Term Frequency

Tokens are more relevant to a given document if they appear in the document more than once.

ex) “Barrack Obama” in “State of the Union” vs. “SNL Hosts”

- Replace Bag of Word’s binary (0,1) marker with frequency markers

Term Frequency cont.

Documents have varying lengths and words can be in many or few documents

So we could normalize raw term frequencies:

- Ex) Divide term frequency by total # words in document

TF-IDF

Term Frequency – Inverse Document Frequency (TFIDF)

A very popular method of normalizing frequencies, words count higher if they occur more frequently within a specific document, and words are penalized for occurring across documents

$$\mathrm{TFIDF}(t,d) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t)$$

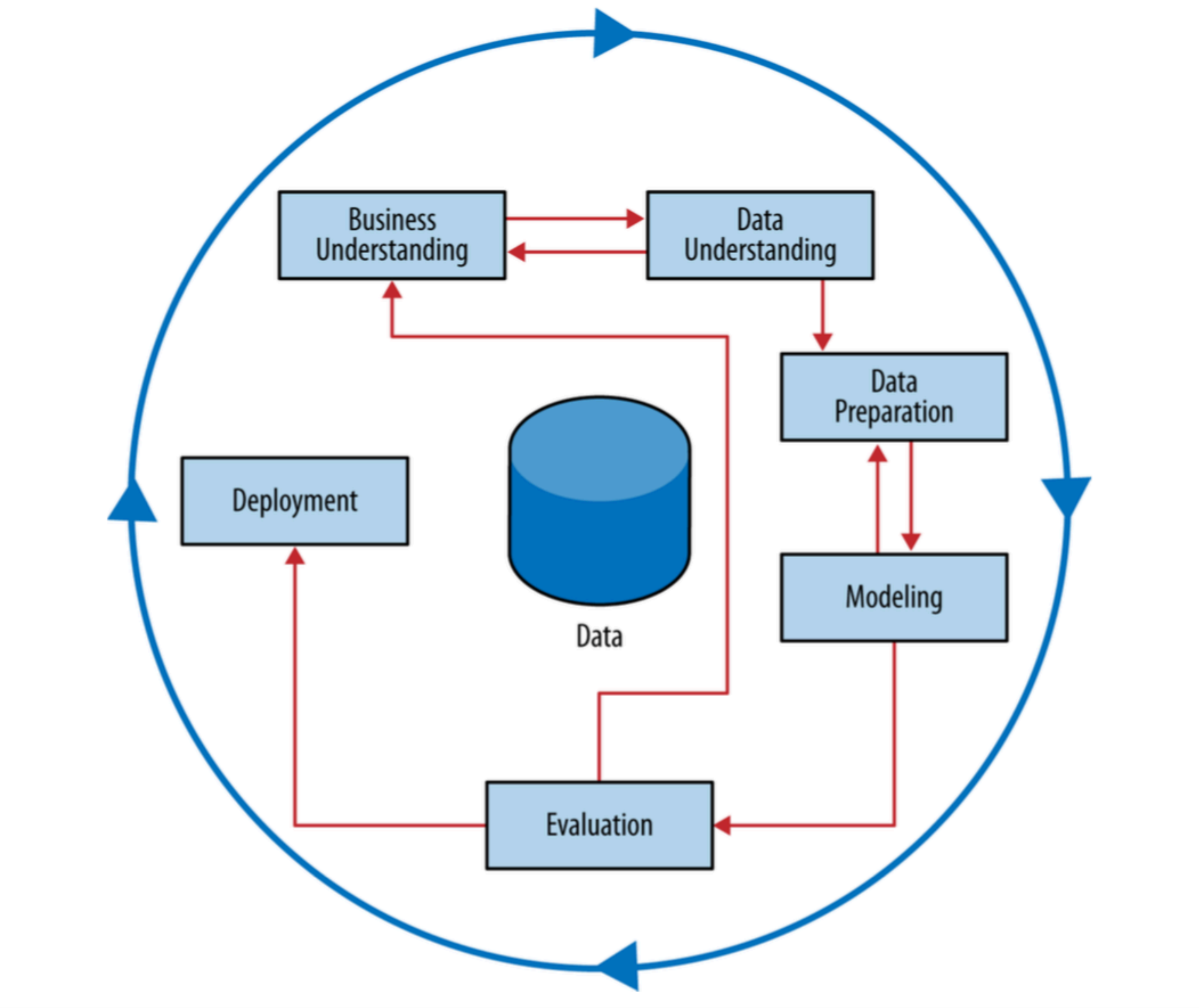
$$\text{Inverse Document Frequency}(t) =$$
$$1 + \log(\text{ Total Documents } / \text{ Documents with } t)$$

N-Grams

- Treat multiple words as one term
  - "New York", "San Diego"
  - "Department of Defense"

- Re-evaluate model with N-Grams instead of singular word terms

Data Mining Cycle



Weka

IMDB

- IMBD collects information about movies. It also lets users write their own reviews, and provide a rating from 1-10.
- Our dataset has: 12,500 positive and 12,500 negative reviews collected from IMDB.
  - Rating 7-10 is marked as Positive: “P”
  - Rating 1-4 is marked as Negative: “N”
  - We discard “neutral” reviews
- Goal: we will build a system that classifies which reviews are positive or negative.

Step 1: Load .arff file

- Under Applications, click on Explorer
- In the Preprocess tab, click on Open File, and load IMDB.arff

Step 2: Convert Text to Word

- Under Filter, click Choose -> unsupervised -> attribute -> StringToWordVector
- Click on the "tokenizer" parameter box, replace the default delimiter with: (.,;:'"?!@#$%^&* {}|□\<>/`~1234567890-=_+)
- Next, set useStoplist to False
- Click OK to close the dialogue box
- Don't forget to hit Apply
- Now, you should see a large list of words in the Attributes list. Scroll through and remove non-word (noise) attributes (usually at the bottom of the list).
- If you have weird words and symbols, you must have wrong delimiters. Click Undo and try again.

Step 3: Binarize Features

- Let's change the features to numeric features: 1 for "the word is present in this review" or 0 for "the word is absent". (Recall the jazz example earlier)
- Under Filter, click Choose -> unsupervised -> attribute -> NumericToBinary. Hit Apply
- Now, all the features in the attribute list should have _binarized appended
- Scroll through, and remove any unbinarized features in the attribute list (usually at the end of the list)

Step 4: Running an Algo

- To run a data mining model, go to the Classify tab
- Make sure (Nom) @class@ is selected in the dropdown menu. This indicates that the 0, 1 variables are our target
- Under Classifier, click Choose -> bayes -> NaïveBayes
- Select Percentage Split as your test option
- Click on More Options, and select Output Predictions
- …and hit Start!

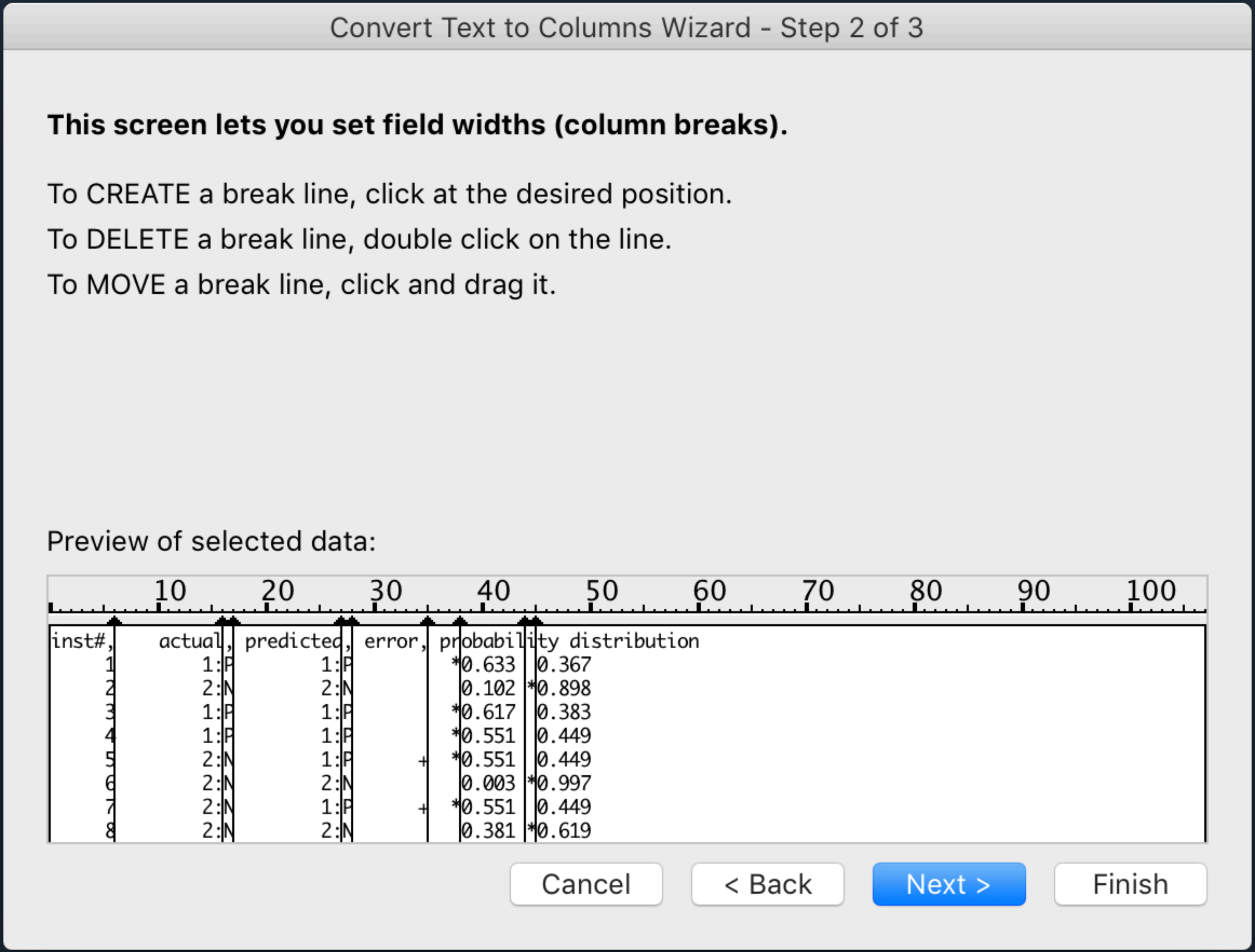
Interpretation and Evaluation

- It is important to use human judgement given the caveats of machine processing natural languages
- Open IMDB_EVALSET.csv, we have two columns – the review text and the target variable value.
- Now, go to your weka window, copy only the prediction part to a new Excel sheet. Clean up the data using Text to Columns.

[Demo]

Text to Columns

- Use Fixed Width



Convert Text to Columns Wizard - Step 2 of 3

This screen lets you set field widths (column breaks).

To CREATE a break line, click at the desired position.

To DELETE a break line, double click on the line.

To MOVE a break line, click and drag it.

Preview of selected data:

```
inst#,     actual, predicted, error, probability distribution
     1        1:P        1:P          *0.633  0.367
     2        2:N        2:N           0.102 *0.898
     3        1:P        1:P          *0.617  0.383
     4        1:P        1:P          *0.551  0.449
     5        2:N        1:P      +   *0.551  0.449
     6        2:N        2:N           0.003 *0.997
     7        2:N        1:P      +   *0.551  0.449
     8        2:N        2:N           0.381 *0.619
```

Interpretation and Evaluation

- Paste the Predicted Column from your new Excel sheet to IMDB_EVALSET.csv
- Let's look at a few reviews that may have misled the classifer. For example, instance number 4:

'Near the beginning after its been established that outlaw John Dillinger (Warren Oates) is an egomaniacal rapist another bandit of the 1930s is cornered in a farm house and surrounded by the FBI. Second-in-command Melvin Purvis (Ben Johnson) surveys the situations sticks a lighted cigar in his mouth picks up two loaded .45-caliber automatics and stalks off into the distant house alone. Bang bang bang. Purvis emerges alone from the house carrying the female hostage the miscreant dead. All in long shot. If youre enthralled by stories like Red Riding Hood this should have considerable appeal. Oh its as exciting as it is mindless. Pretty Boy Floyd meets his demise dramatically. Multiple violations of the civic code. Plenty of shoot outs with Tommy guns and pistols. Blood all over. As history it stinks. Few remember Melvin Purvis as an FBI hero partly I would guess because of his name. Melvin PURVIS? We all remember J. Edgar Hoover who fired Melvin Purvis because he was a rival in the quest for public attention though. The picture was written and directed by John Milius. Hes the guy who had it written into his contract that should any animals be shot and killed in the course of one of his productions he should be the designated shooter. Milius is the guy a compleat gun freak who had Teddy Roosevelts Rough Riders in the Spanish-American war shouting quotations from Henry V -- Saint Crispins Day and all that. Exciting yes and complete garbage. I knew Id never take him alive and I didnt try too hard neither. That is kill em all and let God sort them out. Youll just love it.'

Acknowledgement

Jessica Clark

PhD Candidate, Information,
Operations and Management Sciences