**Week 13 Presentation**

**Week 12 Assignment:**

1)Post your, one or two page, coding and analysis steps from Week 11 assignment in the discussion board. Use a word document format. Do not post the company and analysis, just the analysis steps and any coding performed.

2)Post midterm project presentation or word document and annotated code in discussion board.

**Definition of Text Analytics:**

The terms text analytics, text data mining, and text mining will be used synonymously in this course.

Text analytics uses algorithms for turning free-form text into data that can then be analyzed by applying statistical and machine learning methods, as well as natural language processing techniques.

Text analytics encompasses many sub areas - pattern discovery or exploratory analysis and predictive modeling, as it pertains to text analytics. We will discuss several topics in these areas.

*Often the most challenging part of the data mining process is obtaining and preprocessing the data…*

**Text Mining:**

Text mining as presented here has the following characteristics:

operates with respect to a ***corpus*** of documents

creates a ***dictionary*** or ***vocabulary*** to identify relevant terms

accommodates a variety of ***metrics*** to quantify the contents of a document within the corpus

derives a ***structured vector*** of measurements for each document relative to the corpus

uses ***analytical methods*** that are applied to the structured vector of measurements based on the goals of the analysis (for example, groups documents into segments)

The concept of a dictionary can be thought of as a *vocabulary*. The document collection has a vocabulary that is the union of all the terms contained in each document. Consequently, text mining uses ***dictionary*** or ***vocabulary*** to refer to the collection of terms that are used in the analysis.

Terms not in the dictionary are ignored, except possibly for use in determining the relative frequencies of terms in each document. ***Zipf’s Law***, discussed in a later chapter, helps identify terms in a dictionary that should be included in an analysis.

Text mining works with a collection of documents, ***corpus***. The collection can be dynamic, that is, documents can be added to the collection. You can use the collection to train a model, and you can apply the model to new documents coming into the collection.

New documents are ***scored*** relative to how they compare to the original documents in the collection. If a new document contains a new term, then text mining is ignorant of this new term until that document is used in a new training step.

Many commercial text mining products have strong text-analytics capabilities, most lack data mining capabilities beyond text analytics. The ability to score new documents using a decision tree or a neural network presents new opportunities to improve text mining outcomes (for example, making it possible to use variables derived from text analytics in predictive models).

**Data Mining – two broad areas:**

Pattern Discovery/Exploratory Analysis (Unsupervised Learning): There is no target variable, and some form of analysis is performed to do the following:

identify or define homogeneous groups, clusters, or segments

find links or associations between entities, as in market basket analysis

Prediction (Supervised Learning): A target variable is used, and some form of predictive or classification model is developed.

input variables are associated with values of a target variable, and the model produces a predicted target value for a given set of inputs.

*Data mining analysts know how a predictive model scores new data. However, some analysts might be unaware that unsupervised learning models (that is, data without a known, available target) can also generate scores, and new data can be scored using the model. For example, a new document is scored by calculating the probability of membership in each cluster, and then it is assigned to the cluster associated with the highest probability.*

**Text Mining Applications – Unsupervised:**

**Information retrieval**

finding documents with relevant content of interest

used for researching medical, scientific, legal, and news documents such as books and journal articles

**Document categorization for organizing**

clustering documents into naturally occurring groups

extracting themes or concepts

**Anomaly detection**

identifying unusual documents that might be associated with cases requiring special handling such as unhappy customers, fraud activity, and so on

**Text Mining Applications – Supervised:**

Many typical predictive modeling or classification applications can be enhanced by incorporating textual data in addition to traditional input variables.

churning propensity models that include customer center notes, website forms, e-mails, and Twitter messages

hospital admission prediction models incorporating medical records notes as a new source of information

insurance fraud modeling using adjustor notes

sentiment categorization from customer comments

stylometry or forensic applications that identify the author of a particular writing sample

**Text Mining Signal versus Noise:**

Psychologists know that human beings might react differently to the same stimulus if sufficient time elapses between exposures. On Monday, when you are hungry at lunchtime, you eat a sandwich. Yet, on Tuesday when you are hungry, you opt for a salad. This tendency for different outcomes to occur with similar inputs is attributed to noise, which is unpredictable. You can predict with almost certainty that you will eat lunch next Thursday, but you cannot predict what you will eat with the same certainty.

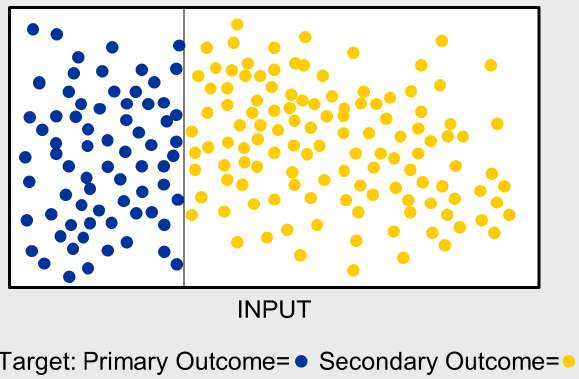
Analytic experts expect errors in prediction related to noise, so methods are developed to minimize errors in the presence of noise. The incremental value that text mining can provide predictive models should be assessed by comparing the quality of a model without incorporating text mining to that achieved after text mining is added.

Target = Signal + Noise

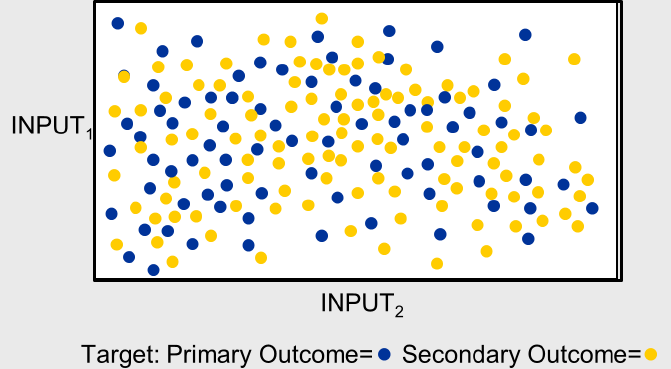
Signal = Systematic Variation = Predictable

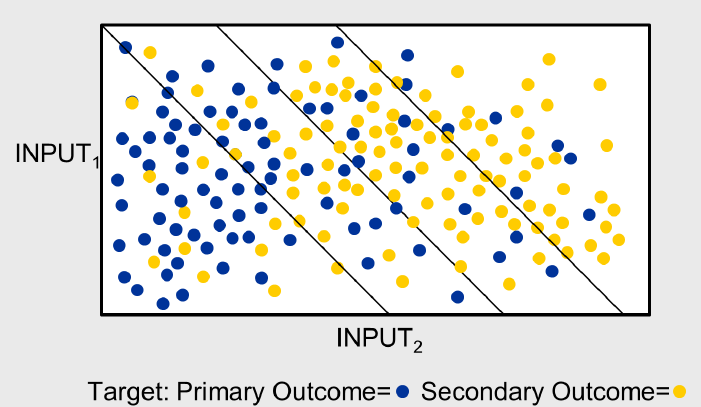
Noise = Random Variation = Unpredictable

The graphic illustrates the pure signal situation. In this case, the training data can be perfectly separated into primary or secondary outcomes using a linear decision boundary. You rarely expect to see this in practice.

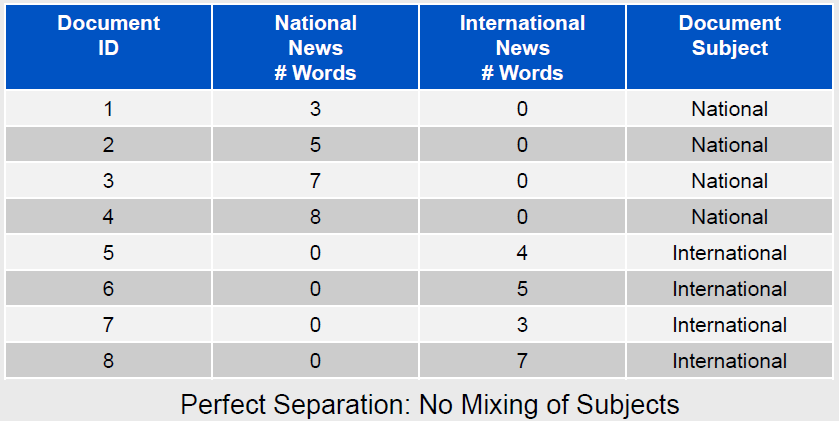


At the other extreme is the pure noise situation. In this case, the training data appears to have no patterns upon which to base a model that can separate the primary outcomes from the secondary outcomes. This situation is more common than you might like. Although pure signal is very rare, pure noise can actually occur in practice.



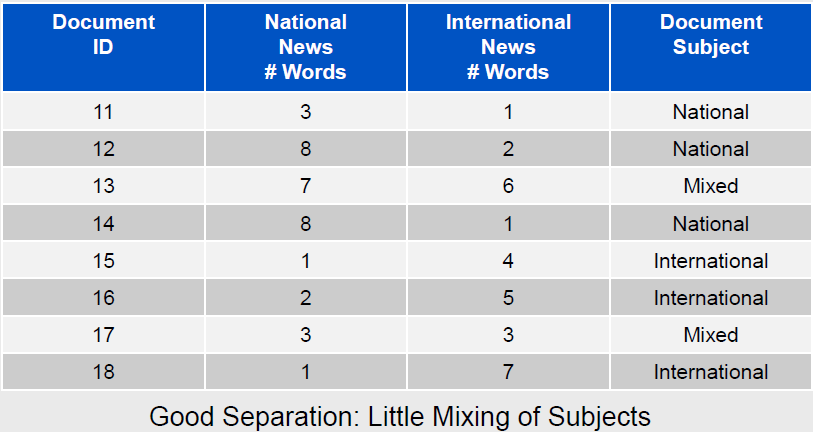
The most common situation in practice is a mixture of signal and noise. You can predict more accurately than randomly guessing. How well you predict depends on whether data is dominated by systematic variation or random variation.

**Text Mining – Perfect Separation:**

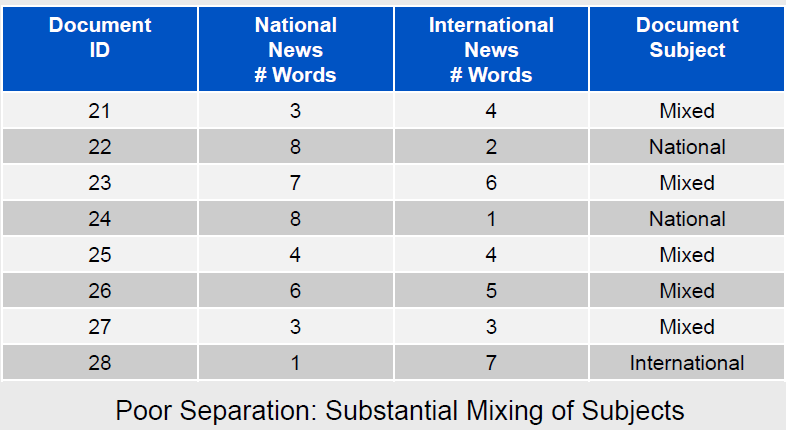
Some document collections are well separated for analytic purposes. The hypothetical example shows eight documents, with four that describe national news items exclusively, and the remaining four describing international news items exclusively. Suppose that you could identify a set of terms that are associated with national news and another set of terms associated with international news. These terms could then be used to classify the documents in the corpus.

**Text Mining – Imperfect Separation:**

With the same topic and analytic objective, another document collection has documents that might mention a heterogeneous set of news articles. You still get good separation, but noise creeps in due to the fact that a document can include multiple subjects.



**Text Mining – Poor Separation:**

Finally, the example shows that if you have a collection of documents that mention many topics and mixes topics, then trying to classify documents into clean categories is difficult.

**Text Analytics References in R and SAS**

Using R - Preparation for our next assignment:

http://handsondatascience.com/TextMiningO.pdf

https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

http://cran.us.r-project.org/doc/Rnews/Rnews\_2008-2.pdf

https://rstudio-pubs-static.s3.amazonaws.com/31867\_8236987cf0a8444e962ccd2aec46d9c3.html

http://www.r-bloggers.com/intro-to-text-analysis-with-r/

Using Base SAS - Preparation for our next assignment:

http://support.sas.com/resources/papers/proceedings12/133-2012.pdf

*SAS related FYI as we don’t have access to the modules:*

https://support.sas.com/resources/papers/proceedings14/1288-2014.pdf

https://support.sas.com/resources/papers/Benchmark\_R\_Mahout\_SAS.pdf

https://support.sas.com/resources/papers/proceedings12/137-2012.pdf

**tm package overview:**

Initiate package: *> library(tm)*

Chracteristics:

Create a corpus – a collection of text documents

Provide various preprocessing operations e.g., stemDoc(), stripWhitespace(), tmTolower()

Create a Document-Term matrix

Inspect / manipulate the Document-Term matrix (e.g. convert into a data frame needed by classifiers)

Train a classifier on pre-classified Document-Term data frame

Apply the trained classifier on new text documents to obtain class predictions and evaluate performance

*Reference: http://web.letras.up.pt/bhsmaia/EDV/apresentacoes/Bradzil\_Classif\_withTM.pdf*

**Introducing Python**

http://en.wikipedia.org/wiki/Python\_(programming\_language):

Python was created by *Guido Van Rossem* in 1991 and emphasizes productivity and code readability. Programmers that want to delve into data analysis or apply statistical techniques are some of the main users of Python for statistical purposes.

The closer you get to working in an engineering environment, the more likely it is you might prefer Python. It’s a flexible language that is great to do something novel, and given its focus on readability and simplicity, its learning curve is relatively low.

Similar to R, Python has packages as well. PyPi (https://pypi.python.org/pypi) is the Python Package index and consists of libraries to which users can contribute. Just like R, Python has a great community but it is a bit more scattered, since it’s a general purpose language. Nevertheless, Python for data science is rapidly claiming a more dominant position in the Python universe: the expectations are growing and more innovative data science applications will see their origin here.

**Getting started with Python**

You can use Python when your data analysis tasks need to be integrated with web apps or if statistics code needs to be incorporated into a production database. Being a fully fledged programming language, it’s a great tool to implement algorithms for production use.

While the infancy of Python packages for data analysis was an issue in the past, this has improved significantly over the years. Make sure to install NumPy /SciPy (scientific computing) and pandas (data manipulation) to make Python usable for data analysis. Also have a look at matplotlib to make graphics, and scikit-learn for machine learning.

Unlike R, Python has no clear “winning” IDE. We recommend you to have a look at Spyder,IPython Notebook and Rodeo to see which one best fits your needs.

**Python Pros and Cons**

***Pro: IPython Notebook:*** The IPython Notebook makes it easier to work with Python and data. You can easily share notebooks with colleagues, without having them to install anything. This drastically reduces the overhead of organizing code, output and notes files. This will allow you to spend more time doing real work.

***Pro: A general purpose language:*** Python is a general purpose language that is easy and intuitive. This gives it a relatively flat learning curve, and it increases the speed at which you can write a program. In short, you need less time to code and you have more time to play around with it!

Furthermore, the Python testing framework is a built-in, low-barrier-to-entry testing framework that encourages good test coverage. This guarantees your code is reusable and dependable.

***Pro: A multi purpose language:*** Python brings people with different backgrounds together. As a common, easy to understand language that is known by programmers and that can easily be learnt by statisticians, you can build a single tool that integrates with every part of your workflow.

***Pro/Con: Visualizations:*** Visualizations are an important criteria when choosing data analysis software. Although Python has some nice visualization libraries, such as Seaborn, Bokeh and Pygal, there are maybe too many options to choose from. Moreover, compared to R, visualizations are usually more convoluted, and the results are not always so pleasing to the eye.

***Con: Python is a challenger:*** Python is a challenger to R. It does not offer an alternative to the hundreds of essential R packages. Although it’s catching up, it’s still unclear if this will make people give up R?