MIS772

Predictive Analytics

Text Analytics

When text becomes numbers

Refer to your textbook by Vijay Kotu and Bala Deshpande, *Data Science: Concepts and Practice*, 2nd ed, Elsevier, 2018.

Text Analytics

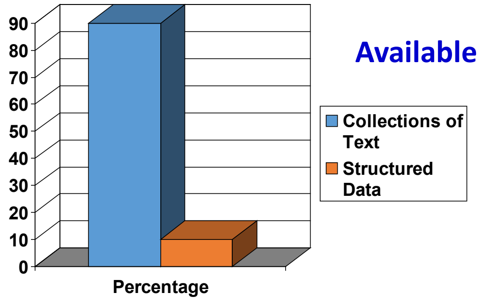
- **Concepts and examples**
- **Case: Skytrax airline reviews**
- **Text representation and parsing**
- **Text mining and predictive models**
- **Sentiment analysis and Segmentation**
- **Explaining and visualization of text mining results**

AACSB ACCREDITED

EFMD EQUIS ACCREDITED

DEAKIN BUSINESS SCHOOL

What is Text Analytics



Available data

What is text analytics?

Gartner:
the process of deriving information from text sources

SAS:
the process of providing structure to unstructured data

Applications

- Emails
- Insurance claims
- News articles
- Web pages
- Patent portfolios
- Contracts
- Technical documents
- Transcripts of phone calls
- Customer complaint letters

Challenges

- Information is in unstructured textual form
- Not readily accessible to standard computer programs
- Difficult to deal with huge collections of documents
- Issues in capturing context and semantics
- Text meaning has historical and cultural basis

The essence: convert text into structure

DEAKIN BUSINESS SCHOOL

Key Concepts in Text Processing

Parsing – breaking text into components, it may involve:

- ***Lexical Analysis*** **– analysis of words and their potential role in a sentence, e.g. articles (a, the), verbs, nouns, adjectives, etc.**
- ***Syntactic Analysis*** **– understanding relationships between words in a sentence, e.g. an adjective, noun and verb form a sentence**
- ***Semantic Analysis*** **– text analysis aiming at understanding the sentence meaning, e.g. a person acting on an object in a place**

On the mechanical level – suitable for analytics – it may involve:

- ***Tokenization*** **– splitting of text into meaningful terms, e.g. *words***
- ***Stemming*** **– process (algorithmic or dictionary based) of reducing words to their "stem", the base or root form, e.g. the words *stemming, stems* and *stemmed* can be replaced with *stem***
- ***Stop Words*** **– words that need to be ignored as they do not differentiate between documents (e.g. *the, here, him*)**
- ***Start Words*** **– words that are of special importance in a given domain, e.g. *products, services, jobs* and *transactions***
- ***Word Pairs, Vectors, Trees, Networks and Graphs*** **– complex relationships between concepts that assist text understanding**

Example Text

The data set consists of 41,396 reviews of air-travel and ***passenger recommendation*** of the airline based on their experience. Each review includes the name of the airline, the passenger / reviewer name and the country of their origin, date of travel, cabin class, route travelled, answers to the short quiz of the passenger experience, text of the review / praises / grievances, as well as, the final recommendation of the airline.

The data has been "wrangled" by *Quang Nguyen* from Skytrax web site.

The goal of this exercise is to use the text of included reviews to create:

1. **A predictive text mining model;**
2. **Extension: A model capable of predicting different aspects of air-travel experience from text.**

ExampleSet (100 examples, 0 special attributes, 20 regular attributes)

text "documents"

Filter (100 / 100 examples): all

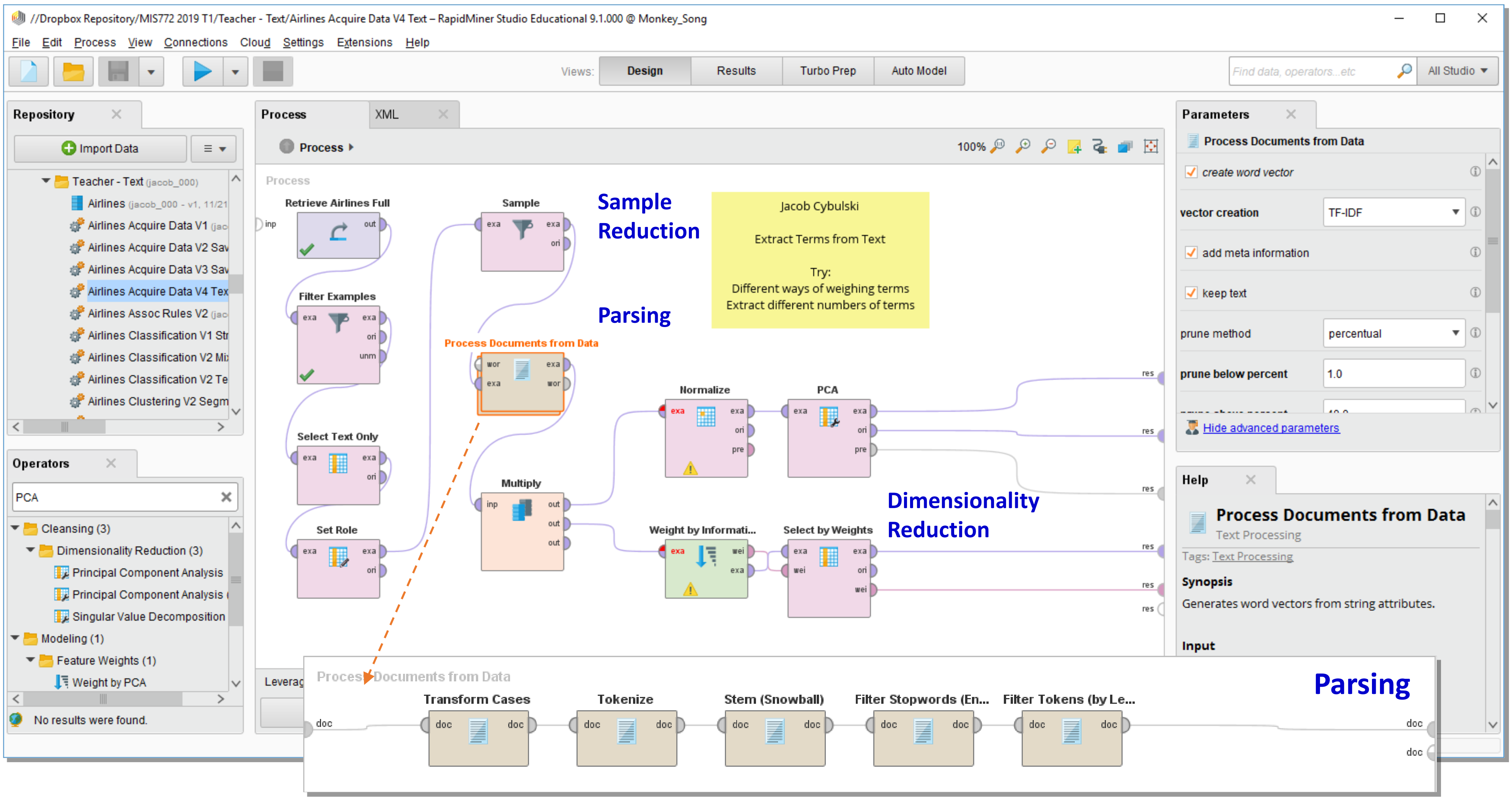
Row No.	airline_name	link	title	author	author_country	date	content	aircraft	type_traveller	cabin_flown	route	overall_rating	seat_comfo...
21	atlasjet-airlines	/airli...	Atlasglobal customer revi...	A J Bastien	United States	Nov 1, 2012	Oct 20 2012. Second flig...	?	?	Economy	?	8	2
22	austrian-airlines	/airli...	Austrian Airlines custome...	Matija Zeko	Croatia	Jun 26, 2015	Zagreb to Vienna we fle...	Dash Q...	Couple Leisu...	Economy	Zagreb to Rome ...	7	3
23	austrian-airlines	/airli...	Austrian Airlines custome...	Brooks Kathryn	Canada	Jan 11, 2015	We just arrived in Delhi f...	?	?	Business Cla...	?	10	5
24	austrian-airlines	/airli...	Austrian Airlines custome...	Bruno Lumpet	Switzerland	Dec 10, 2014	The seats in Business ...	?	?	Business Cla...	?	8	5
25	azul-linhas-aereas-bra...	/airli...	Azul Airlines customer rev...	Marcel van de...	Netherlands	Feb 20, 2013	FOR-REC v.v. with Azul. ...	?	?	Economy	?	9	?
26	british-airways	/airli...	British Airways customer ...	Ash Aryan	Ireland	Jul 8, 2015	I have been flying betwe...	E170	Business	First Class	Dublin to London...	3	2
27	british-airways	/airli...	British Airways customer ...	B Lakin	United Kingdom	Apr 10, 2015	LHR to Philadelphia but ...	?	?	First Class	?	9	4
28	bulgaria-air	/airli...	Bulgaria Air customer revi...	Stef Heathcote	United Kingdom	Sep 13, 2013	This airline lacks basic ...	?	?	Economy	?	1	2
29	cambodia-angkor-airlin...	/airli...	Cambodia Angkor Air cus...	Bassett Kevin	Australia	Jun 9, 2014	10/5/14 BKK-REP. We al...	?	?	Economy	?	7	4
30	british-airways	/airli...	British Airways customer ...	K Nicol	United Kingdom	Oct 31, 2014	LAX to LHR - 25 Oct 201...	?	?	Business Cla...	?	2	3
31	canjet-airlines	/airli...	CanJet Airlines customer...	C Wiebe	Canada	Mar 2, 2011	Kelowna - Puerto Vallart...	?	?	Economy	?	?	?
32	british-airways	/airli...	British Airways customer ...	P Harris	United Kingdom	May 7, 2014	Lanzarote to Gatwick on ...	?	?	Economy	?	1	3
33	china-southern-airlines	/airli...	China Southern Airlines c...	Yang Xi	New Zealand	Aug 18, 2014	CZ306 7 June from Auck...	?	?	Economy	?	10	3
34	cityjet	/airli...	CityJet customer review	Paul Cox	United Kingdom	Oct 12, 2014	Flew back from Dublin o...	?	?	Economy	?	10	4
35	cityjet	/airli...	CityJet customer review	Raynaud Fi	United Kingdom	Sep 6, 2014	Our flight Toulon - Lond...	?	?	Economy	?	1	1

Note that the original CSV file had a number of errors (such as spurious line breaks) and thus some data cleansing had to be undertaken. The examples also include many missing values that have been preserved, but which may need to be eliminated for certain tasks.

DEAKIN BUSINESS SCHOOL

During the acquisition of text we face a number of problems, i.e.

1) **How to represent unstructured textual in the form of numeric vectors (parsing);**
2) **How to reduce the potentially huge number of variables (weighing, PCA);**
3) **How to reduce huge volume of examples in a data set (sampling or clustering).**



Process of Text Parsing

1. Original text as a sequence of characters

Result History | Document (Process Documents from Data)

Document

The flight was terrible. The cabin was noisy. Flight attendants were helpless. I cancelled all future bookings.

The flight was terrible. The cabin was noisy. Flight attendants were helpless. I cancelled all future bookings.

Text

2. Text shifted to lower case

Result History | Document (Process Documents from Data)

Document

the flight was terrible. the cabin was noisy. flight attendants were helpless. i cancelled all future bookings.

The flight was terrible. The cabin was noisy. Flight attendants were helpless. I cancelled all future bookings.

3. Text split into terms, e.g. “attendants” and “cancelled”

Result History | Document (Process Documents from Data)

Document

the flight was terrible the cabin was noisy flight attendants were helpless i cancelled all future bookings

The flight was terrible. The cabin was noisy. Flight attendants were helpless. I cancelled all fu

4. Terms in a root form, e.g. “attend” and “cancel”

Result History | Document (Process Documents from Data)

Document

the flight was terribl the cabin was noisi flight attend were helpless i cancel all futur book

The flight was terrible. The cabin was noisy. Flight attendants were helpless. I cancelled all future bookings.

Once defined, the process of text processing can be reused across many text analytics workflows and models

Process

5. Information “poor” terms removed, e.g. “the”

Result History | Document (Process Documents from Data)

Document

flight terribl cabin noisi flight attend helpless i cancel futur book

The flight was terrible. The cabin was noisy. Flight attendants were helpless. I cancelled all future bookings.

6. Long or short terms removed, e.g. “I”

Result History | Document (Process Documents from Data)

Document

flight terribl cabin noisi flight attend helpless cancel futur book

The flight was terrible. The cabin was noisy. Flight attendants were helpless. I cancelled all future bookings.

Number Vectors

ExampleSet (3 examples, 1 special attribute, 23 regular attributes) Filter (3 / 3 examples): all

Row No.	text	attend	book	cabin	cancel	crew	flight	food	futur	good	great	help	helpless	mediocr	movi
1	the flight was great. the crew were very helpful. food was pretty good.	0	0	0	0	1	1	1	0	1	1	1	0	0	0
2	the flight was terrible. the cabin was noisy. flight attendants were helpless. i cancelled all future bookings.	1	1	1	1	0	2	0	1	0	0	0	1	0	0
3	it was all ok. food was mediocre. music and movies wer reasonable. seats were unconfortable.	0	0	0	0	0	0	1	0	0	0	0	0	1	1

7. Text as a vector of numbers, e.g. every term represented by the number of times it occurred in text

DEAKIN BUSINESS SCHOOL

Types of Term Vector Representation

Binary

ExampleSet (3 examples, 1 special attribute, 23 regular attributes) Filter (3 / 3 examples): all

Row No.	text	attend	book	cabin	cancel	crew	flight	food	futur	good	great
1	flight great crew veri help food pretti good	0	0	0	0	1	1	1	0	1	1
2	flight terribl cabin noisi flight attend helpless cancel futur book	1	1	1	1	0	1	0	1	0	0
3	food mediocr music movi wer reason seat unconfort	0	0	0	0	0	0	1	0	0	0

Binary representation of text, which indicates whether or not each term is present in a document / example (true / false or 1 / 0)

Term Occurrence

ExampleSet (3 examples, 1 special attribute, 23 regular attributes) Filter (3 / 3 examples): all

Row No.	text	attend	book	cabin	cancel	crew	flight	food	futur	good	great
1	flight great crew veri help food pretti good	0	0	0	0	1	1	1	0	1	1
2	flight terribl cabin noisi flight attend helpless cancel futur book	1	1	1	1	0	2	0	1	0	0
3	food mediocr music movi wer reason seat unconfort	0	0	0	0	0	0	1	0	0	0

Occurrence representation of text, which indicates how many times each term occurred in a document / example (0, 1, 2, …)

Term Frequency

ExampleSet (3 examples, 1 special attribute, 23 regular attributes) Filter (3 / 3 examples): all

Row No.	text	attend	book	cabin	cancel	crew	flight	food	futur	good	great
1	flight great crew veri help food pretti good	0	0	0	0	0.354	0.354	0.354	0	0.354	0.354
2	flight terribl cabin noisi flight attend helpless cancel futur book	0.289	0.289	0.289	0.289	0	0.577	0	0.289	0	0
3	food mediocr music movi wer reason seat unconfort	0	0	0	0	0	0	0.354	0	0	0

Frequency representation of text, which is a weighted and squared-scaled term frequency within a document

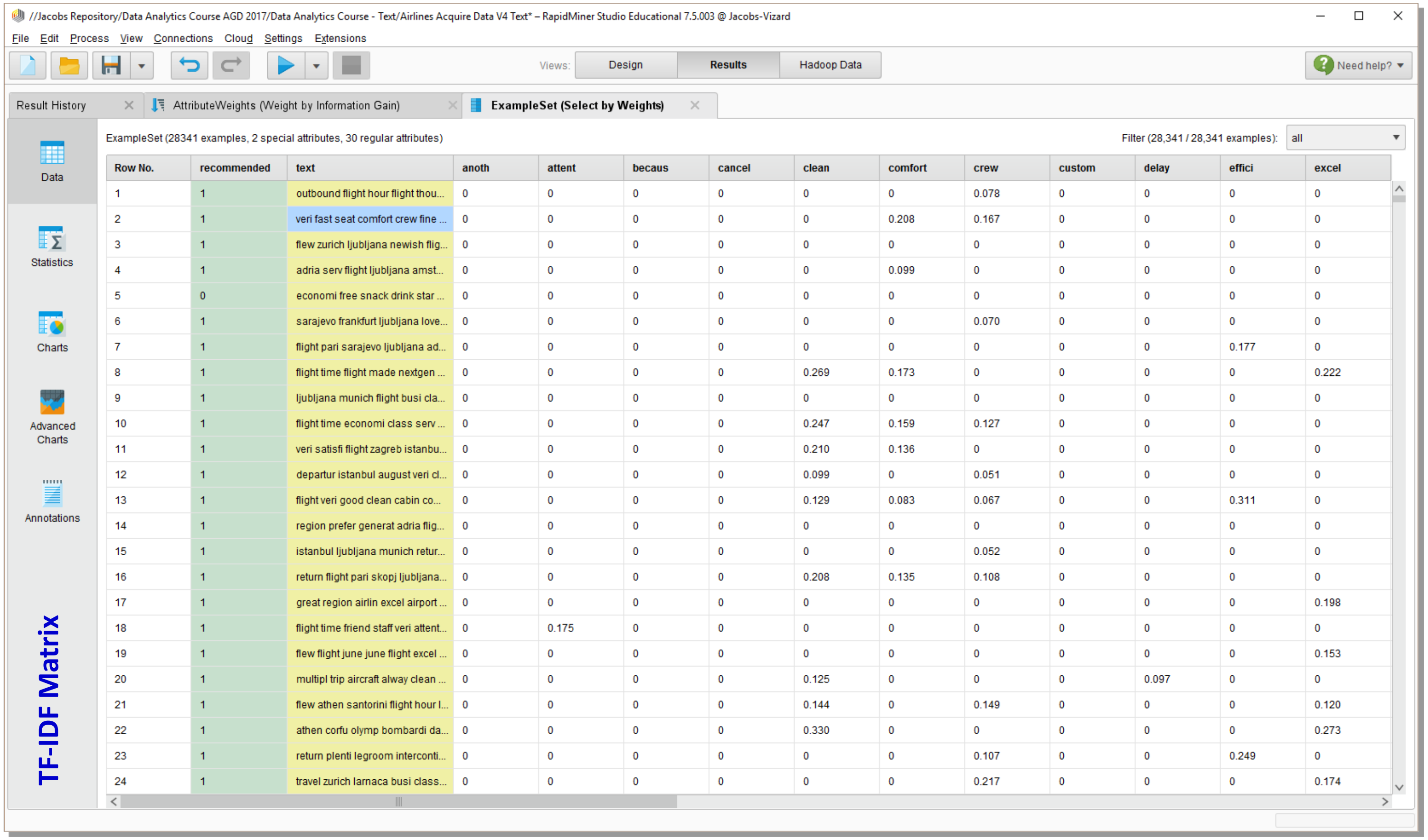
TF-IDF (most commonly used in practice)

ExampleSet (3 examples, 1 special attribute, 23 regular attributes) Filter (3 / 3 examples): all

Row No.	text	attend	book	cabin	cancel	crew	flight	food	futur	good	great
1	flight great crew veri help food pretti good	0	0	0	0	0.399	0.147	0.147	0	0.399	0.399
2	flight terribl cabin noisi flight attend helpless cancel futur bo...	0.342	0.342	0.342	0.342	0	0.253	0	0.342	0	0
3	food mediocr music movi wer reason seat unconfort	0	0	0	0	0	0	0.138	0	0	0

Term frequency – inverse document frequency, which weighs term frequency within a document against terms frequency across all documents and this way penalizes terms which occur often in all documents and thus do not differentiate between them

TF-IDF Text Representation and Dim Reduction



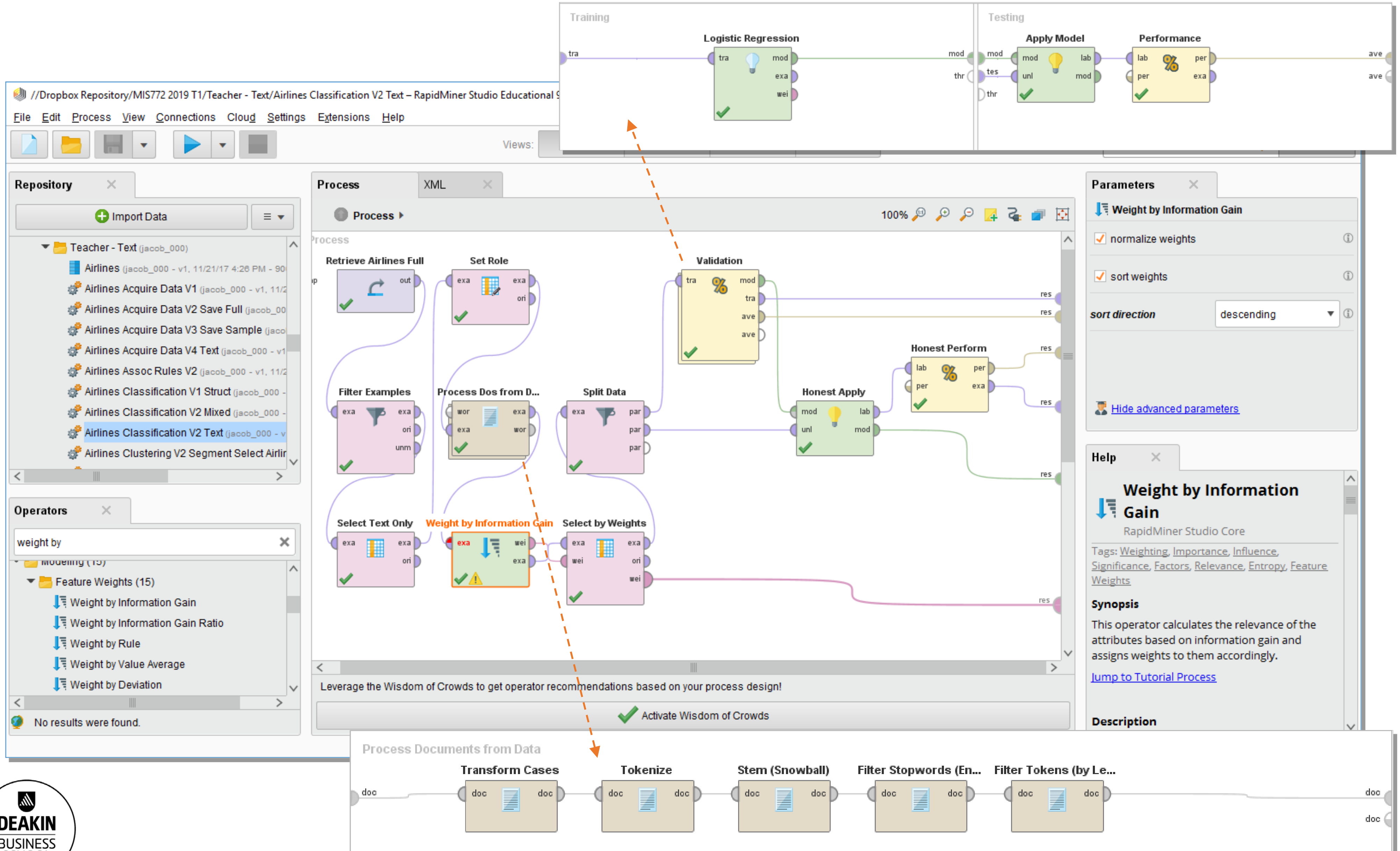
Row No.	recommended	text	anoth	attent	becaus	cancel	clean	comfort	crew	custom	delay	effici	excel
1	1	outbound flight hour flight thou...	0	0	0	0	0	0	0.078	0	0	0	0
2	1	veri fast seat comfort crew fine ...	0	0	0	0	0	0.208	0.167	0	0	0	0
3	1	flew zurich ljubljana newish flig...	0	0	0	0	0	0	0	0	0	0	0
4	1	adria serv flight ljubljana amst...	0	0	0	0	0	0.099	0	0	0	0	0
5	0	economi free snack drink star ...	0	0	0	0	0	0	0	0	0	0	0
6	1	sarajevo frankfurt ljubljana love...	0	0	0	0	0	0	0.070	0	0	0	0
7	1	flight pari sarajevo ljubljana ad...	0	0	0	0	0	0	0	0	0	0.177	0
8	1	flight time flight made nextgen ...	0	0	0	0	0.269	0.173	0	0	0	0	0.222
9	1	ljubljana munich flight busi cla...	0	0	0	0	0	0	0	0	0	0	0
10	1	flight time economi class serv ...	0	0	0	0	0.247	0.159	0.127	0	0	0	0
11	1	veri satisfi flight zagreb istanbu...	0	0	0	0	0.210	0.136	0	0	0	0	0
12	1	departur istanbul august veri cl...	0	0	0	0	0.099	0	0.051	0	0	0	0
13	1	flight veri good clean cabin co...	0	0	0	0	0.129	0.083	0.067	0	0	0.311	0
14	1	region prefer generat adria flig...	0	0	0	0	0	0	0	0	0	0	0
15	1	istanbul ljubljana munich retur...	0	0	0	0	0	0	0.052	0	0	0	0
16	1	return flight pari skopj ljubljana...	0	0	0	0	0.208	0.135	0.108	0	0	0	0
17	1	great region airlin excel airport ...	0	0	0	0	0	0	0	0	0	0	0.198
18	1	flight time friend staff veri attent...	0	0.175	0	0	0	0	0	0	0	0	0
19	1	flew flight june june flight excel ...	0	0	0	0	0	0	0	0	0	0	0.153
20	1	multipl trip aircraft alway clean ...	0	0	0	0	0.125	0	0	0	0.097	0	0
21	1	flew athen santorini flight hour l...	0	0	0	0	0.144	0	0.149	0	0	0	0.120
22	1	athen corfu olymp bombardi da...	0	0	0	0	0.330	0	0	0	0	0	0.273
23	1	return plenti legroom interconti...	0	0	0	0	0	0	0.107	0	0	0.249	0
24	1	travel zurich larnaca busi class...	0	0	0	0	0	0	0.217	0	0	0	0.174

attribute	weight
good	1
excel	0.846
told	0.621
comfort	0.596
friend	0.483
great	0.478
hour	0.453
worst	0.432
rude	0.387
poor	0.315
nice	0.299
uncomfort	0.290
custom	0.287
anoth	0.271
clean	0.268
pleasant	0.265
delay	0.261
attent	0.259
effici	0.253
final	0.246
cancel	0.234
wait	0.234
terribl	0.207
profession	0.193
plenti	0.179
crew	0.176
said	0.173
becaus	0.172
overal	0.166
roug	0.165

Top 30 terms in the order of their weights

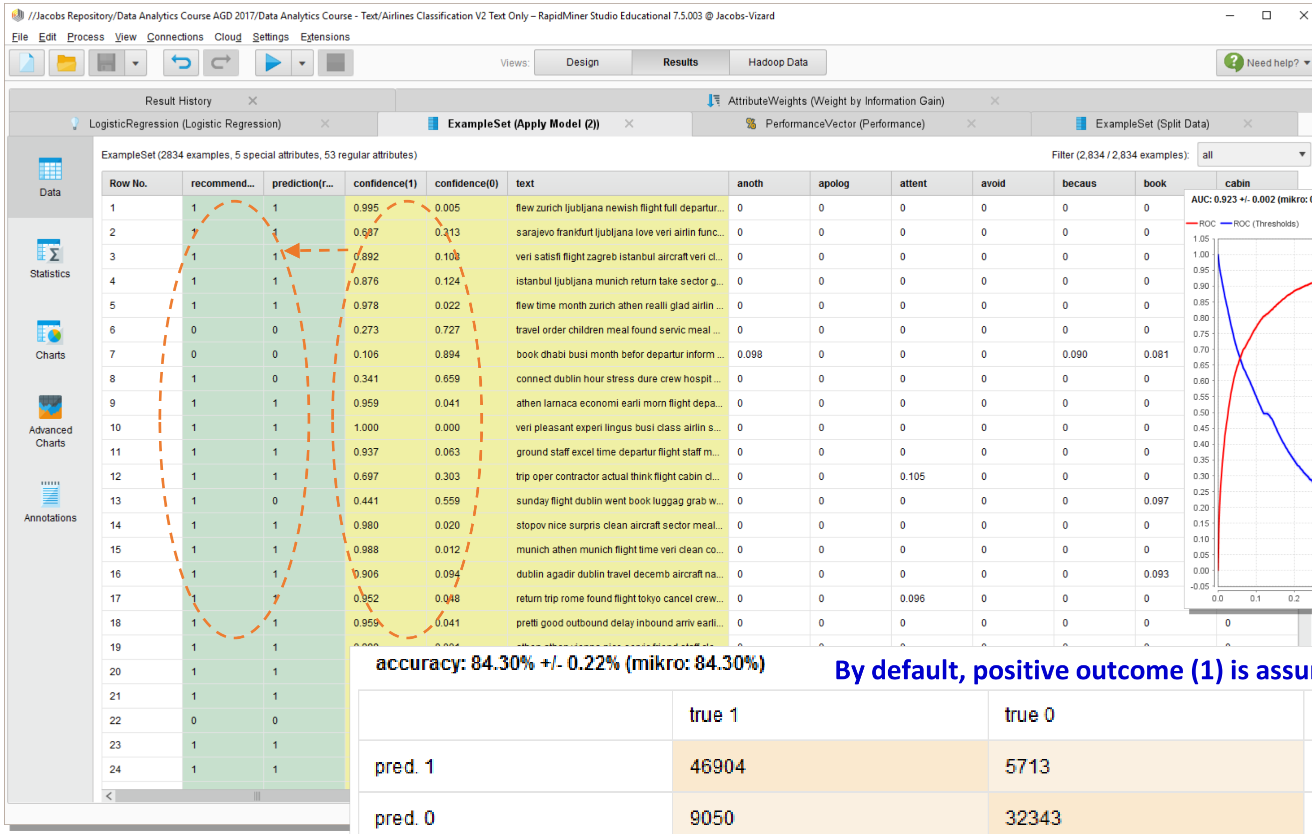
- **Document set is finally represented as a table of numbers**
- **Documents are rows - examples. Terms are columns - variables**
- **Each number indicates presence of a term in text (as TF-IDF)**
- **To reduce their number we need to decide which terms are most useful, e.g. by weighing them (by information gain / entropy) OR mathematically transforming a set of terms into a much smaller set of variables (using Principal Component Analysis)**

Text mining models aim to create new variables from text and then use them (often together with structured variables) to train models capable of predicting various aspects of the observed phenomena, e.g. sentiment but also passenger views on quality of meals, seat comfort or crew services.



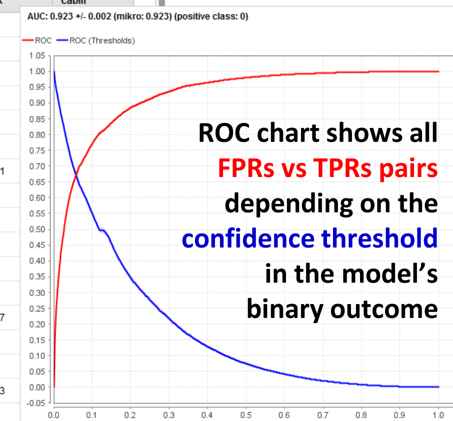
DEAKIN BUSINESS SCHOOL

Text Mining and Binomial Prediction



Classification

AUC / ROC Chart



ROC chart shows all **FPRs vs TPRs pairs** depending on the **confidence threshold** in the model's binary outcome

accuracy: 84.30% +/- 0.22% (mikro: 84.30%)

By default, positive outcome (1) is assumed for confidence > 0.5

	true 1	true 0	class precision
pred. 1	46904	5713	89.14%
pred. 0	9050	32343	78.14%
class recall	83.83%	84.99%	

Accuracy & Kappa

Text terms can be used as predictor attributes of a label attribute, which could represent a category (e.g. positive / negative), a Likert scale assessment of quality (survey answers 1-10), or monetary value (profit, loss, cost).

Once terms are turned into numeric attributes, any standard model can be used in the analytic workflow, e.g., in Classification: Decision Trees, Logistic Regression or k-NN.
Any standard performance measure can also be applied, e.g., Accuracy, Kappa, AUC.

Prediction models: Text only vars, Structured data only, mixed

DEAKIN BUSINESS SCHOOL

Sentiment Analysis

What is sentiment analysis?

- **Sentiment analysis is commonly used to analyse social media content for people's attitudes to products and services.**
- **The goal is to scan information to determine how people feel about an issue (e.g. brand or product), and what they will do about it.**
- **There are many software products and online services available to do so, e.g. from SAS, IBM, Microsoft, iSentia or Meltwater.**
- **The results of such analysis are indicative only and can be easily manipulated by companies themselves.**
- **A classification model that predicts sentiment is a sophisticated sentiment analysis system, *sentiment analysis = classification*.**

Simple sentiment analysis with word counts…

Sentiment lists can be obtained from the web, can be purchased or custom built for the specific application

1. Split a document into terms
2. Determine the *positive* and *negative* terms in sentences, paragraphs or documents
3. Calculate a *sentiment score* of the text based on the number of positive and negative terms, e.g. their proportion
4. Apply the sentiment score to all documents

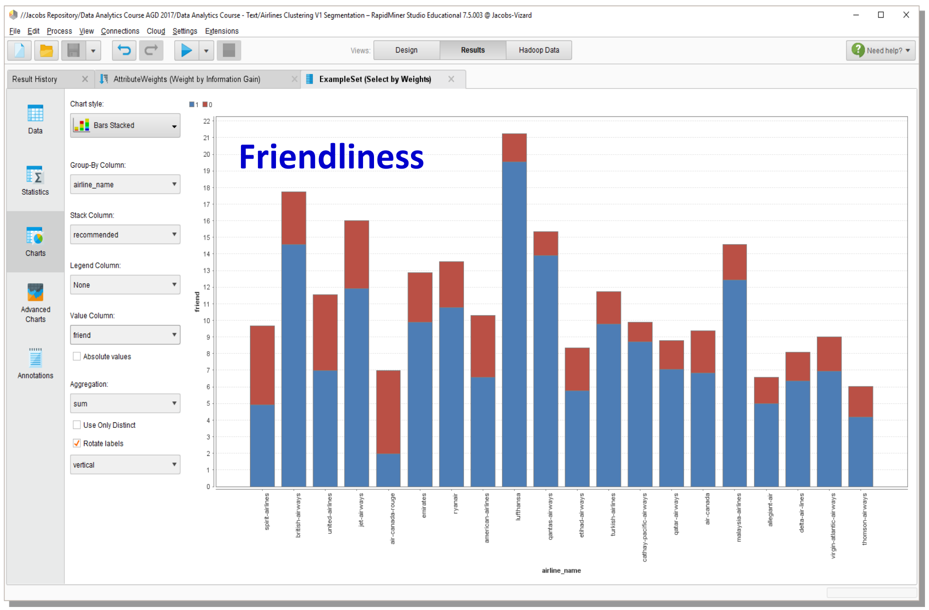
Sentiment analysis is available in many open source and commercial packages (e.g. in RM)

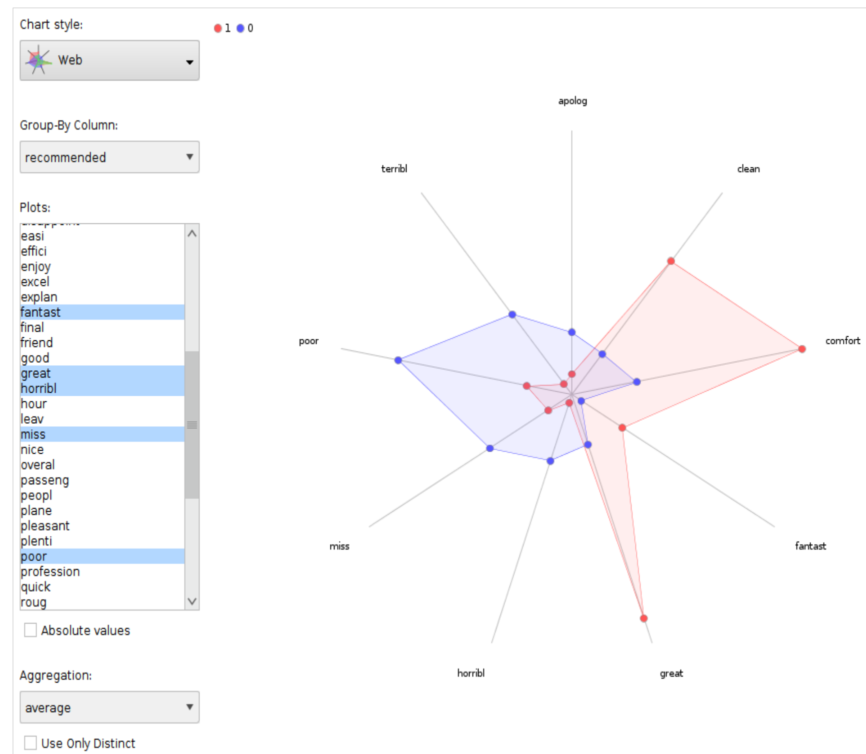
DEAKIN BUSINESS SCHOOL

Text Mining and Segmentation

Segmentation **(visualized using stacked bar charts, web/spider plots, etc.) is commonly performed on one of the nominal attributes, e.g., the label, to identify natural groups based on similarity of text descriptions, rather than any pre-existing categorisation.**

Such visualisations help qualify and explain sentiment in user terms and in relation to the existing categories of data.





Stacked column charts are amongst many tools used to understand data segmentation

DEAKIN BUSINESS SCHOOL

Summary / Review

- Why is text an important part of business analytics?
- What are the typical applications of text analytics?
- Explain why the objective of text analytics is to convert text into structured form?
- What is tokenization and stemming of text?
- How can dimensionality of text representation be reduced?
- Explain the use of PCA in text dimensionality reduction.
- What are the advantages and disadvantages of term weighing vs PCA in dimensionality reduction?
- How can text be visualized graphically? What for?
- What is sentiment analysis?
- How can sentiment analysis be performed?
- What the the shortcomings of simplistic sentiment analysis?

DEAKIN BUSINESS SCHOOL