Text Mining Analytics

Getting Started

Agenda

Unstructured vs. Structured Data

Text Mining Analytics

-- concepts

-- cleaning data

-- word frequencies

-- hierarchical clustering

Structured v. Unstructured Data

Structured Data – Data is organized into a pre-defined structure, i.e., a database. We can have millions of rows, columns and tables, but a database is structured.

Unstructured Data – Data does not have a pre-defined data model. Think of a Twitter feed or a bunch of satellite images or the entire list of speeches from the British Parliament since 1803.

What is Text Analytics?

Gaining actionable insights through the analysis written communication

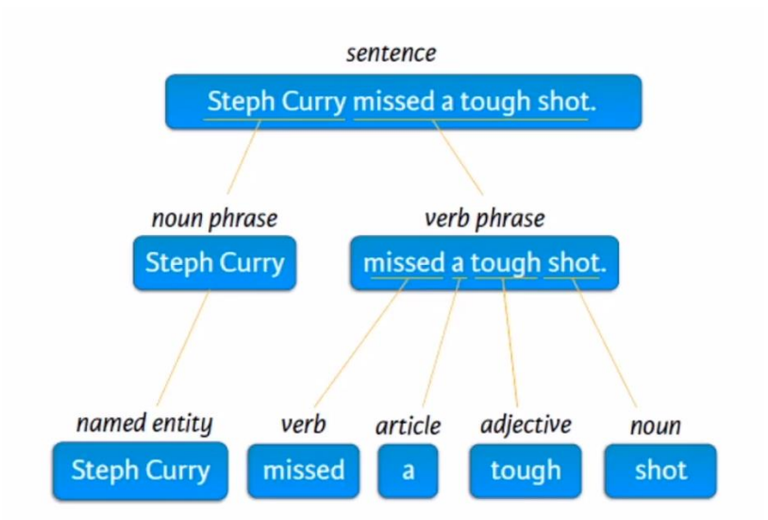
What is Text Analytics?

1. Sentiment Analysis
2. Search unstructured data
3. Spam filtering (characteristics of e-mails)
4. Social Media monitoring
5. Competitive intelligence (business, security…)
6. Translation
7. Simulation

Semantic v. Bag of Words

Semantic – the process of mapping a natural-language sentence into a formal representation of its meaning.

Bag of Words – the simplification of the text, disregarding grammar and word order.



What is a corpus?

Corpus (or corpora plural) – a large set of text

textAnalytics1.Rmd × | tidy_books ×

Filter

	book	linenumber	chapter	word
1	Sense & Sensibility	1	0	sense
2	Sense & Sensibility	1	0	and
3	Sense & Sensibility	1	0	sensibility
4	Sense & Sensibility	3	0	by
5	Sense & Sensibility	3	0	jane
6	Sense & Sensibility	3	0	austen
7	Sense & Sensibility	5	0	1811
8	Sense & Sensibility	10	1	chapter
9	Sense & Sensibility	10	1	1
10	Sense & Sensibility	13	1	the

Cleaning up the data

Stop Words – Common ‘useless’ words we generally want to remove for data analytics. Words like ‘the’, ‘an’, ‘a’ and ‘in’. The words are very common and can get in the way.

Stemming – Returning words in a text to their original stem. For instance, the words ‘chop’, ‘chopping’ and ‘chopped’ all simply become ‘chop’ when stemmed.

Cleaning up the data

TM Function	Description	Before	After
tolower()	Makes all text lowercase	Starbucks is from Seattle.	starbucks is from seattle.
removePunctuation()	Removes punctuation like periods and exclamation points	Watch out! That coffee is going to spill!	Watch out That coffee is going to spill
removeNumbers()	Removes numbers	I drank 4 cups of coffee 2 days	I drank cups of coffee days ago.
stripWhiteSpace()	Removes tabs and extra spaces	I like coffee.	I like coffee.
removeWords()	Removes specific words (e.g. "the", "of") defined by the data scientist	The coffee house and barista he visited were nice, she said hello.	The coffee house barista visited nice, said hello.

Cleaning up the data

Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense. Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbors.

Cleaning up the data

Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense. Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbors.

word <fctr>	freq <dbl>
the	4
and	3
dursley	3
they	3
very	3
was	3
were	3
mrs	2
much	2
neck	2

Cleaning up the data

Mr. Mrs. Dursley, number four, Privet Drive, proud say perfect normal, thank much. They last peopl expect involv anyth strang mysterious, just hold nonsense. Mr. Dursley director firm call Grunnings, made drills. He big, beefi man hard neck, although larg mustache. Mrs. Dursley thin blond near twice usual amount neck, came use spent much time crane garden fences, spi neighbors.

word <fctr>	freq <dbl>
dursley	3
mrs	2
much	2
neck	2
although	1
amount	1
anyth	1
beefi	1
big	1
blond	1

Cleaning up the data

word <fctr>	freq <dbl>
the	200
and	99
was	85
his	69
that	47
had	39
dursley	38
have	32
said	28
all	26

word <fctr>	freq <dbl>
dursley	47
said	28
mr.	23
professor	23
cat	19
peopl	19
mrs.	18
mcgonagal	17
look	16
back	14

Cleaning up the data

word <fctr>	freq <dbl>
harri	1326
ron	429
look	401
hagrid	370
hermion	269
back	267
one	267
get	235
know	219
like	213

We can also combine all of the chapters.

Or we could go through each chapter and try to detect a theme or style.

TDM v DTM

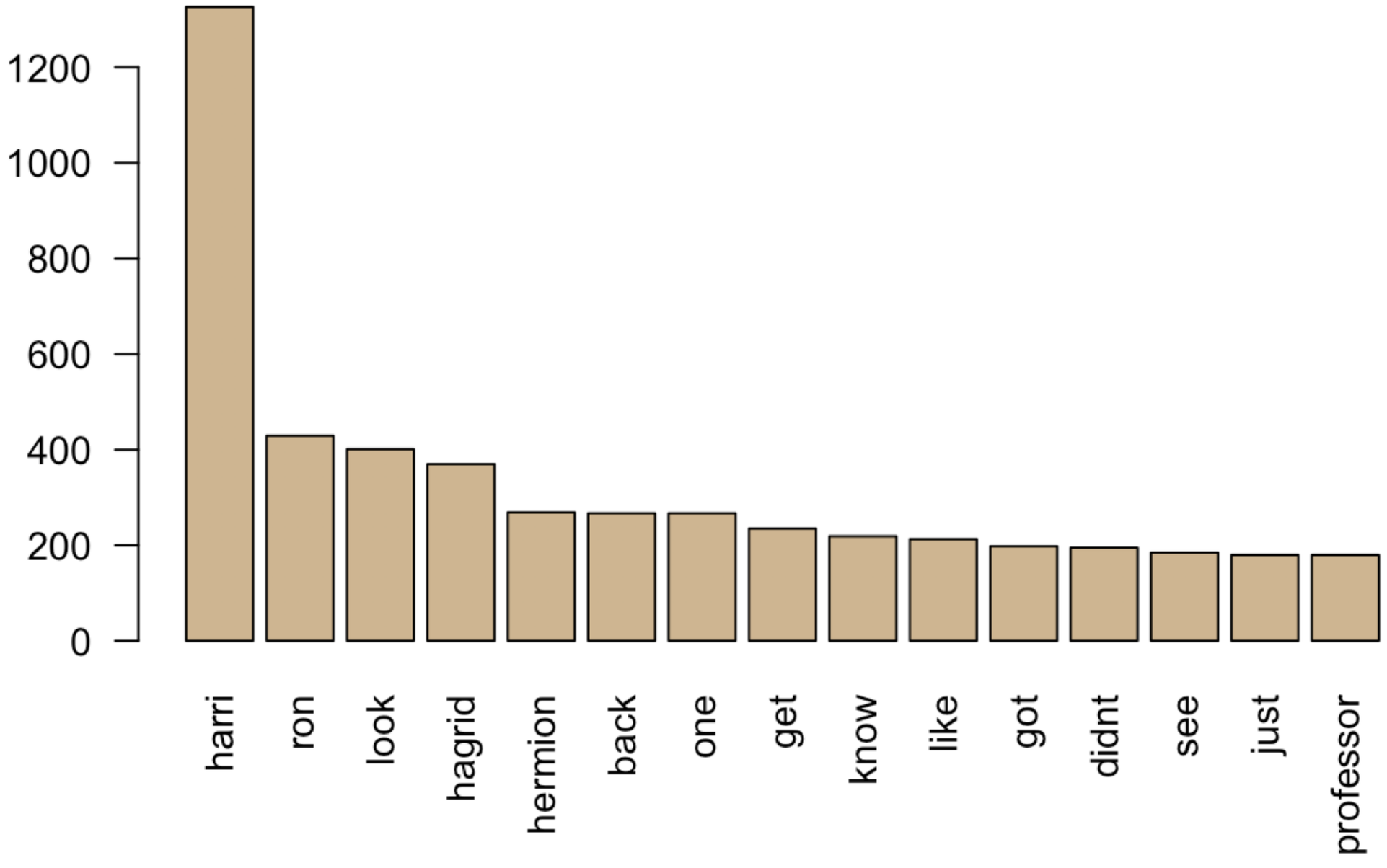
	Tweet 1	Tweet 2	Tweet 3	...	Tweet N
Term 1	0	0	0	0	0
Term 2	1	1	0	0	0
Term 3	1	0	0	0	0
...	0	0	3	1	1
Term M	0	0	0	1	0

Term Document Matrix (TDM)

	Term 1	Term 2	Term 3	...	Term M
Tweet 1	0	1	1	0	0
Tweet 2	0	1	0	0	0
Tweet 3	0	0	0	3	0
...	0	0	0	1	1
Tweet N	0	0	0	1	0

Document Term Matrix (DTM)

TDM v DTM


1200
1000
800
600
400
200
0
harri
ron
look
hagrid
hermion
back
one
get
know
like
got
didnt
see
just
professor