

Visualizing Textual Data



Matthew Renze

SOFTWARE CONSULTANT

@matthewrenze www.matthewrenze.com



Overview



Visualizing Textual Data

Demo



Textual Data Analysis

Body of text

Extract new information

Patterns and trends

Unstructured data

To be, or not to be: that is the question:
Whether 'tis nobler in the mind to suffer
The slings and arrows of outrageous fortune,
Or to take arms against a sea of troubles,
And by opposing end them? To die: to sleep;
No more; and by a sleep to say we end
The heart-ache and the thousand natural shocks
That flesh is heir to, 'tis a consummation
Devoutly to be wish'd. To die, to sleep;
To sleep: perchance to dream: ay, there's the rub;
For in that sleep of death what dreams may come

Hamlet, Act 3, Scene 1



Converting Unstructured Text Data

To be, or not to be: that is the question:
Whether 'tis nobler in the mind to suffer
The slings and arrows of outrageous fortune,
Or to take arms against a sea of troubles,
And by opposing end them? To die: to sleep;
No more; and by a sleep to say we end
The heart-ache and the thousand natural shocks
That flesh is heir to, 'tis a consummation
Devoutly to be wish'd. To die, to sleep;
To sleep: perchance to dream: ay, there's the rub;
For in that sleep of death what dreams may come

Hamlet, Act 3, Scene 1



Converting Unstructured Text Data

Convert to lower case

to be, or not to be: that is the question:
whether 'tis nobler in the mind to suffer
the slings and arrows of outrageous fortune,
or to take arms against a sea of troubles,
and by opposing end them? to die: to sleep;
no more; and by a sleep to say we end
the heart-ache and the thousand natural shocks
that flesh is heir to, 'tis a consummation
devoutly to be wish'd. to die, to sleep;
to sleep: perchance to dream: ay, there's the rub;
for in that sleep of death what dreams may come.

Hamlet, Act 3, Scene 1



Converting Unstructured Text Data

Convert to lower case

Remove punctuation

to be or not to be that is the question
whether 'tis nobler in the mind to suffer
the slings and arrows of outrageous fortune
or to take arms against a sea of troubles
and by opposing end them to die to sleep
no more and by a sleep to say we end
the heartache and the thousand natural shocks
that flesh is heir to 'tis a consummation
devoutly to be wish'd to die to sleep
to sleep perchance to dream ay there's the rub
for in that sleep of death what dreams may come

Hamlet, Act 3, Scene 1



Converting Unstructured Text Data

Convert to lower case

Remove punctuation

Remove stopwords

question
whether nobler mind suffer
slings arrows outrageous fortune
take arms sea troubles
opposing end die sleep
sleep say end
heartache thousand natural shocks
flesh heir consummation
devoutly wish'd die sleep
sleep perchance dream rub
sleep death dreams may come

Hamlet, Act 3, Scene 1



Converting Unstructured Text Data

Convert to lower case

Remove punctuation

Remove stopwords

Reduce to stem words

question
whether noble mind suffer
sling arrow outrag fortun
take arm sea troubl
oppos end die sleep
sleep say end
heartach thousand natur shock
flesh heir consumm
devout wish die sleep
sleep perchanc dream rub
sleep death dream may come

Hamlet, Act 3, Scene 1



Converting Unstructured Text Data

Convert to lower case

Remove punctuation

Remove stopwords

Reduce to stem words

Strip extra whitespace

question whether noble mind suffer sling arrow
outrag fortun take arm sea troubl oppos end die
sleep sleep say end heartach thousand natur shock
flesh heir consumm devout wish die sleep sleep
perchanc dream rub sleep death dream may come

Hamlet, Act 3, Scene 1



Converting Unstructured Text Data

Convert to lower case

Remove punctuation

Remove stopwords

Reduce to stem words

Strip extra whitespace

Convert to frequencies

Term	Frequency
sleep	5
die	2
dream	2
end	2
arm	1
arrow	1
come	1



Structured Text Data

Bag-of-words model

Frequency of terms

Additional variables

Term	Frequency
aaa	1
aang	1
aaron	2
aback	1
abagnale	1
abandon	1
abandoned	10



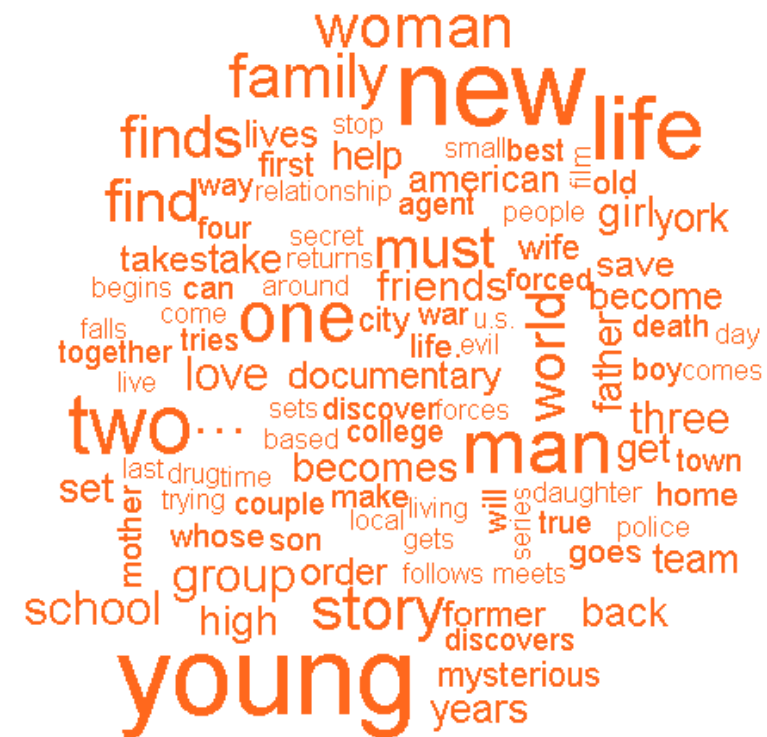
Frequency Word Cloud

Plot Keywords by Frequency

Visualization of terms

Sized by frequency

Random or ordered



Quantitative Word Cloud

Plot Keywords by Average Box Office Revenue

Sized by numeric variable

Add a main chart title



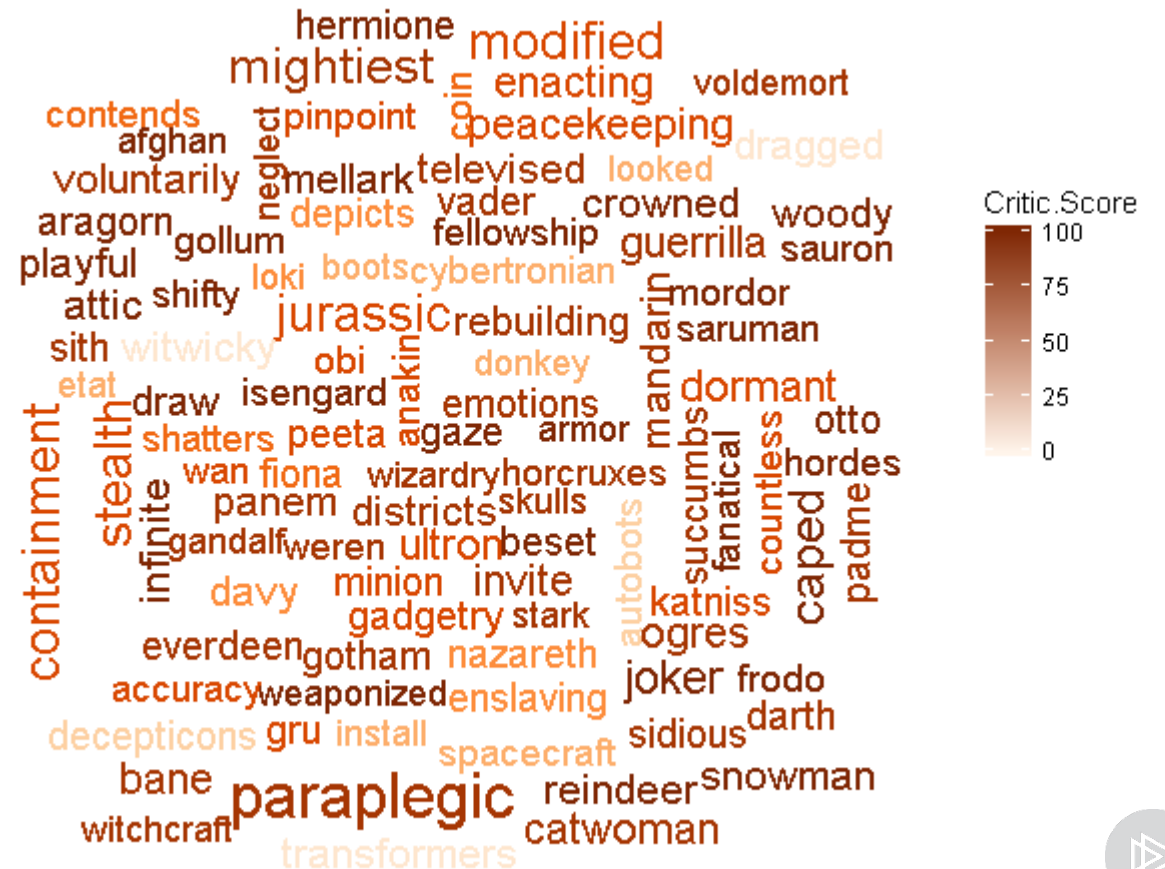
Colored Word Cloud

Plot Keywords by Average Box Office Revenue and Critic Score

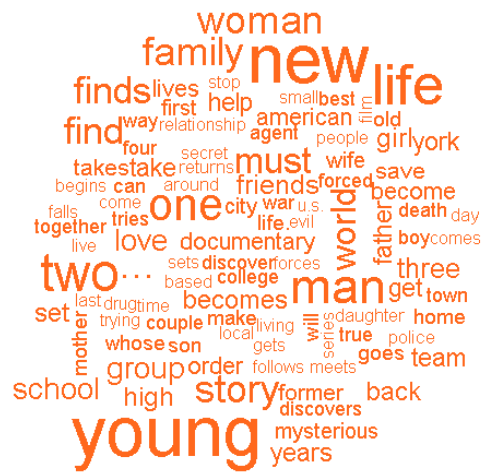
Sized by 1st variable

Colored by 2nd variable

Add title and color scale



Textual Data Visualization





1. What are the top keywords by frequency?
2. What are the top keywords by box office?
3. What are the top keywords by critic score?



Load the Plots Data



Create a Corpus



Transform the Corpus



Create a Frequency Word Cloud



Load the Words Data



Create Quantitative Word Cloud



Create a Color Palette



Create a Colored Word Cloud





Summary



Visualizing Textual Data

Demo

