EUROPE DISCOVERY SUMMIT
EXPLORING DATA INSPIRING INNOVATION

Don’t judge a book by its cover

Bernard McKeown, SAS

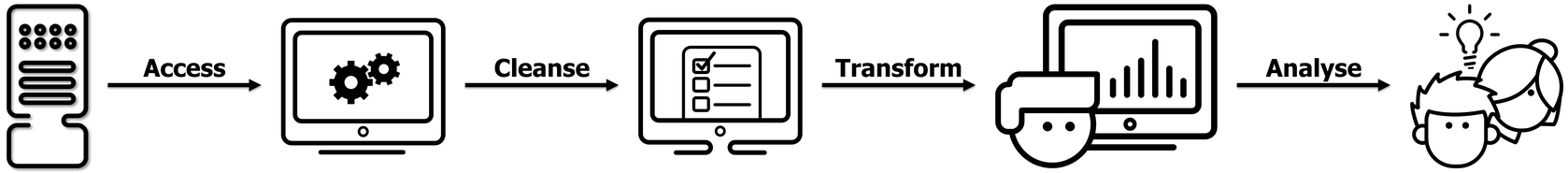
What aim of this poster?

I’m fascinated by words as well as numbers. And with JMP13’s new Text Explorer, I’ve a chance to look at both. I used Text Explorer to look at some of the leading works of all time to see if there is something that they tell us about the author or their writing.

What is text mining?

An estimated 80% of data is stored as text.

Text mining is the process of finding and exploiting useful patterns in text data.



Sources of text

- Emails from customers
- Notes from customer service, maintenance engineers, doctors, nurses
- Transcriptions, eg of customer service calls
- News articles
- Professional reports and journals
- And… books!

Text mining lexicon

- Term: character string separated by space or punctuation
- Tokenisation: turning words into Terms
- Stop Words: literally stop using the word in the analysis
- Stemming: categorising words with the same beginning
- Word pairs and phrases: some relationship between words is important (eg The Who)
- Document: itself! (a collection of words)
- Corpus: the collection of Documents being analysed

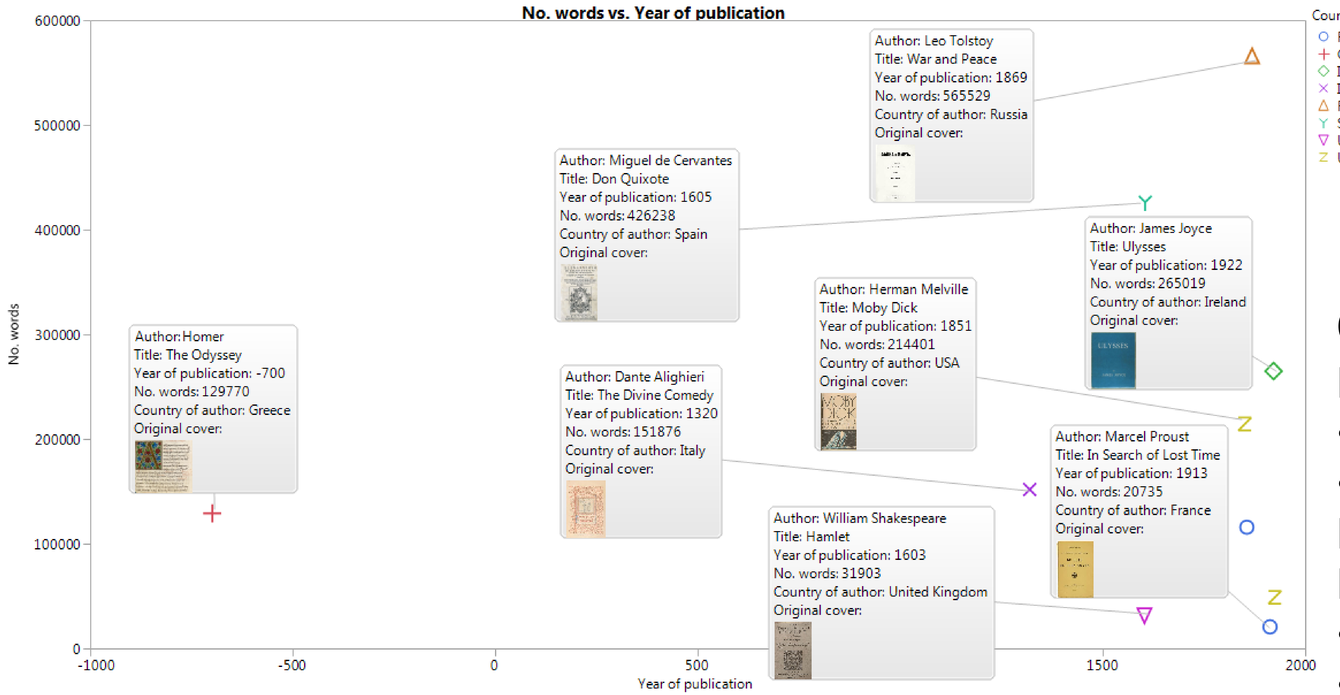
EUROPE DISCOVERY SUMMIT
EXPLORING DATA INSPIRING INNOVATION

Don’t judge a book by its cover

Bernard McKeown, SAS

Which books?

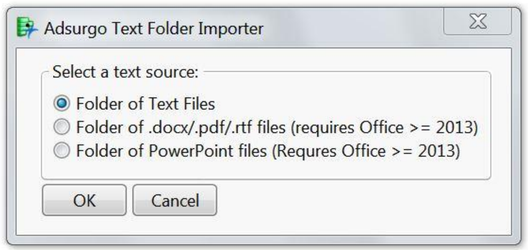
Greatest books survey of surveys:
http://thegreatestbooks.org/
Sourced from Project Gutenberg
https://www.gutenberg.org/

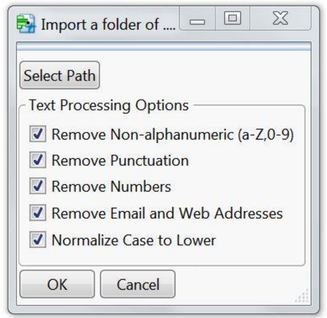


Accessing the data

Adsurgo Add-In (www.adsurgo.com) to access and cleanse:

- Extracts text from text files, pdf, Word, Powerpoint
- Removes non-alphanumeric, punctuation, numbers, email addresses
- Normalises data to lower case





Cleansing the data

Parse the data into Terms

- Bag of Words approach (converts Document into Terms)
- Regular expressions to parse text documents

Recoding to deal with spelling errors
Reducing words in Documents:

- Stop Words – literally “stop that word”
- Recoding to replace words with synonyms
- Stemming to capture words with similar stems

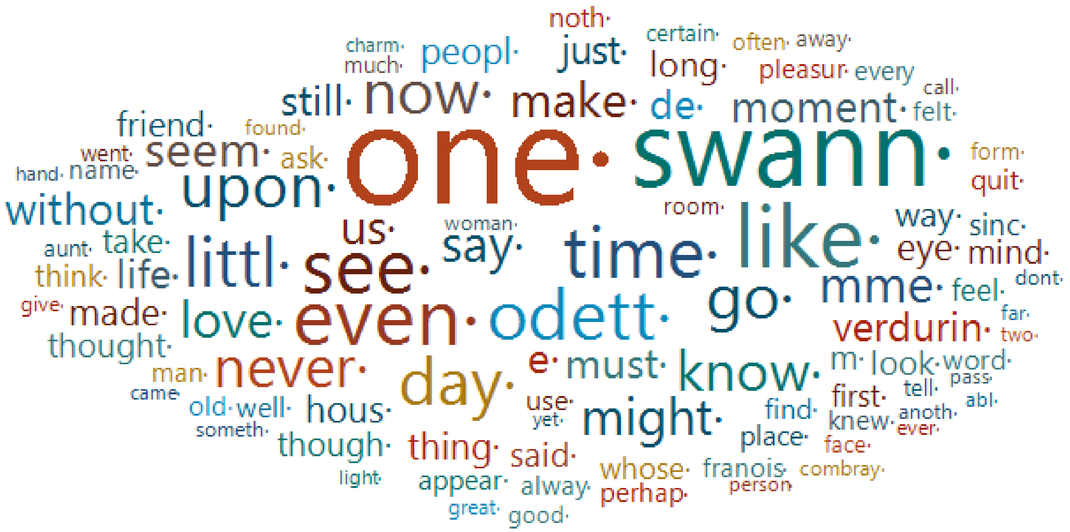
EUROPE DISCOVERY SUMMIT
EXPLORING DATA INSPIRING INNOVATION

Don’t judge a book by its cover

Bernard McKeown, SAS

Converting text to numbers

- Document summary including readability statistics
- Wordle or Word Cloud
- Document-Term Matrix
- Single-Value Decomposition (SVD): similar to PCA
- Naïve Bayesian Models: for classifying independent predictors
- Sentiment Analysis: how people feel about something



Where is text mining used?

- Classifying responses into codes
- Predicting customer response
- Maintenance planning
- Medical treatments
- Fraud detection
- Language translation
- Search engines

What I’ve learnt

Text mining follows same process as other mining

It has it’s own, unique lexicon

Requires similar documents

- The Document became the Corpus in this paper
- Readability stats probably more useful in this case

Useful where you want to find similarity between documents

- Used to prove Shakespeare is one person!

EUROPE
DISCOVERY SUMMIT
EXPLORING DATA
INSPIRING INNOVATION

Don’t judge a book by its cover

Bernard McKeown, SAS

#7 The Odyssey (Homer, 700 BC)

ulysses one said house men man us now went go e son telemachus upon tell may see suitors home ship made shall father came good let take much back sea took great gods way make people still like jove also heaven set till old answered long minerva among hands round two even away though know give day many never thus therefore saw time first left put must might another told every say wine water without brought country ithaca penelope well yet gave soon end going heard hand place find land stranger return wife nothing mother side far others ships spoke

#9 The Divine Comedy (Dante, 1320)

v thou one thy thus thee now c canto yet see us may first whose hath o shall eyes made light like love ye upon said still forth e whence spirit th well saw words doth man good I turnd came two een round time way spake great ii much son world guide many place sun god st far life p long art might name part near earth ere sight know side soon heard iii unto left spirits

#5 Hamlet (Shakespeare, 1603)

ham lord king queen o shall hor thou hamlet good now let pol thy well like e know sir tis may go us enter love hath ill laer speak give must oph thee make upon ros say man father much clown one think laertes see horatio heaven play tell time yet thus exit death mother exeunt look polonius take ay guil mar soul life dead ghost hear might night whose indeed made nothing guildenstern ophelia dear god leave within

#2 Ulysses (Joyce, 1922)

said bloom like mr one stephen old says now see man two time o back yes eyes know good hand well street little first father way say us never day just round right long face go night head must sir god name dont e put j mrs im let going thing look came john life still hes young thats asked poor woman last made went away ill something voice might tell three always dedalus make though house give hat course left much white ever love saw hands mulligan world lord want new behind black told think take bit every took