Text as Data
Text Analytics

Robert Stine
Department of Statistics
Wharton School of the University of Pennsylvania

www-stat.wharton.upenn.edu/~stine

Introduction

Why look at text as data?

Why look at text?

Interesting

How does ETS they score the written SAT? Diagnose autism?

What gives away how a justice on the Supreme Court will vote?

Opportunity to augment classical data

“How can I use these written comments?”

Connections to modern statistical modeling

Issues of big data, neural networks/deep learning, and variable/model selection

Examples of text data

Medical data combine lab measurements with clinical evaluations

Open-ended survey responses (e.g., ANES)

Written employment applications

Ad click prediction based on search text

Illustrative Applications

Two types: supervised and unsupervised

Supervised have a known response to guide analysis

Unsupervised don’t (think cluster analysis)

Unsupervised examples

Are Facebook posts about my company positive or negative?

What topics dominate articles written in science?

Supervised

Does the content of a speech indicate political leaning?

Can you anticipate popularity of a movie from initial review?

Does text improve models or proxy for numerical data?

Lecture Schedule

Plan

Monday	Introduction
	A deep dive, then back to fundamentals
Tuesday	Sentiment analysis, vector space models
	Latent semantic analysis
Wednesday	Generative probability models
	Naive Bayes and hierarchical topic models
Thursday	Overflow, deep learning
	Language models

Style

First hour of lecture, some computing

Second hour more focused on R computing

Further Topics in Text

Not covering everything!

Emphasize problems with statistics connection

Some things you will want to learn more about

Linguistics, structure of language

- Parts of speech, named entities. Make a friend of a linguist!

Language modeling, translation

- Sequence to sequence modeling needs even more data

Text manipulations using regular expressions

- Get a copy on-line of egrep_for_linguists.pdf

Books

Manning and Schütze (1999) Foundations of Statistical NLP

Jurasfsky and Martin (2008) Speech and Language

Software

Comparison to Mosteller & Wallace analysis

They studied authoship of the Federalist papers “by hand”

Mosteller and Wallace (1963). Inference in an authorship problem.

JMP, SAS

Text tools now found in mainstream packages

R

Reproducible research: Scripting versus point and click

tm (text miner) supplemented by tidytext

Supporting package: dplyr, ggplot2, stringr, readr

Alternative: NLTK and python

But then you have to move to R for the analysis

Overview Example

Questions and Data

Wine tasting notes

Can you distinguish a red wine from a white wine using a brief note that describes its taste and aroma?

classification

Can you recognize the variety of red wine?

Cabernet vs merlot vs pinot vs zinfandel

Can you predict the price? Rating points?

regression

Each tasting note is short, but we have a lot of them

Does text add value?

Have numerical data, traditional predictive features

Does information in the text add value?

Tasting Notes

Data

21,000 tasting notes from Beverage Tasting Institute

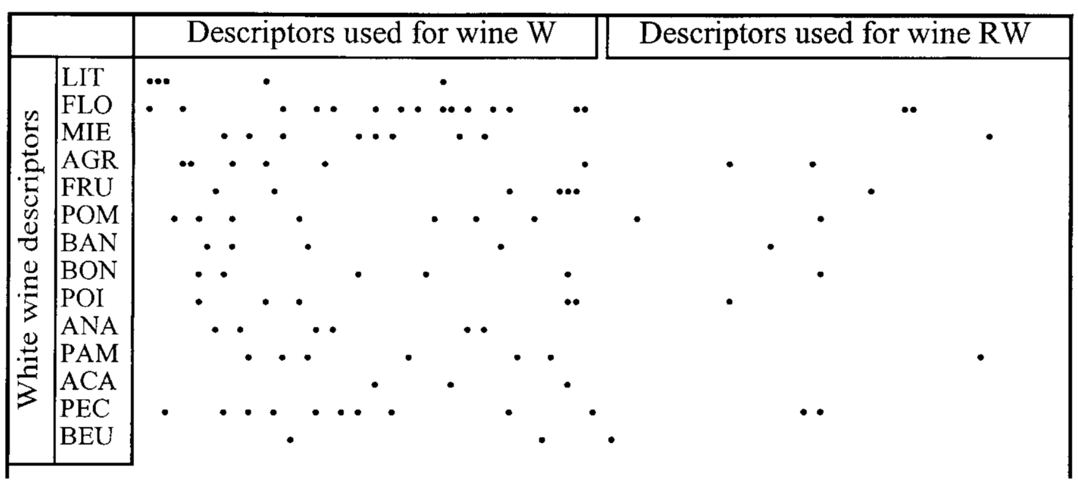
“Earthy, herbal, slightly herbaceous aromas. A medium-bodied palate leads to a short finish that is earthy, tart and has limited fruit.”

“Toasty oak, cherry and thyme aromas. A rich entry leads to a full-bodied palate and a well-structured finish with vibrant acidity, refined tannins, and lovely varietal fruit.”

Lots of tasting notes, but each is relatively short

Mark Liberman   http://languagelog.ldc.upenn.edu/nll/?p=3887/

Do people describe taste, or do they describe color?



“The color of odors”

Typical Steps

90% or more of effort

Prepare data

- Deciding on role for text
- Editing: removing weird characters, such as html markup
- Feature engineering: eg making regression variables

Modeling choices, issues

- Unsupervised (clustering) vs supervised (regression)
- Structural (prob model) vs predictive (conditional mean)

Inference

- What is the inferential context? Do you have a sample?

Browsing the Data

Always good to wander around in your data

Visual, interactive software tools like JMP make this painless

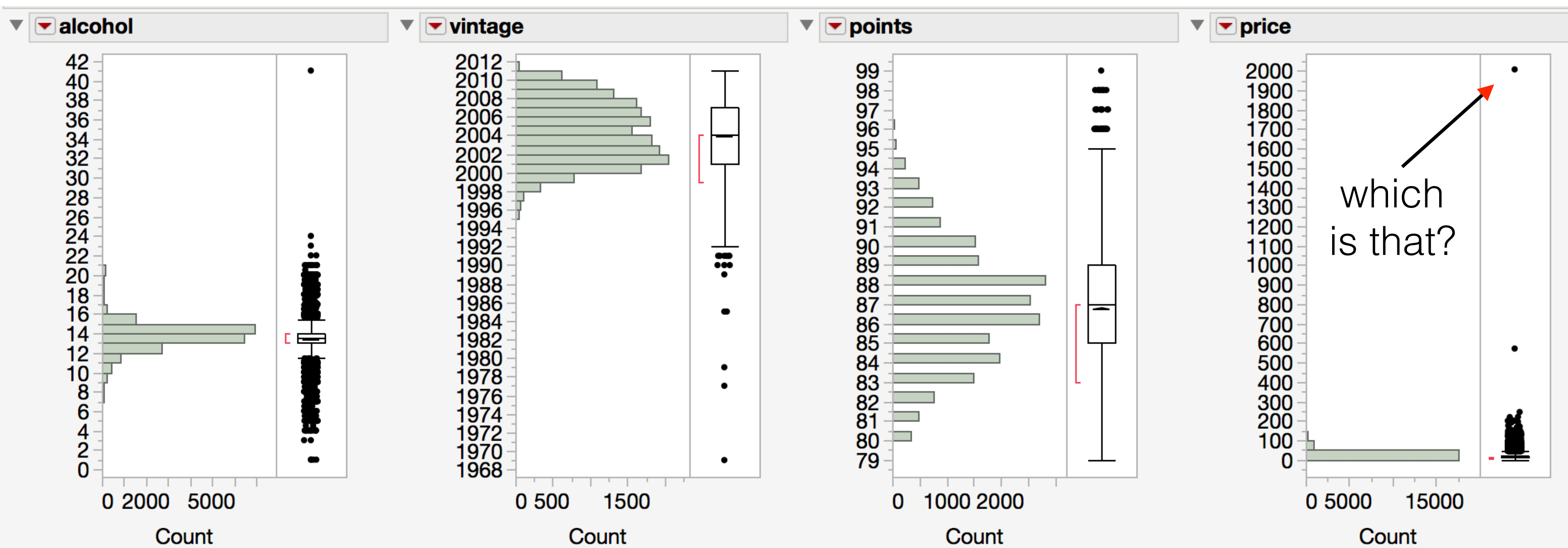
Novelty for stat data: Several columns are long strings

	review	id	label	description
1	1	163522	Corey Creek Vineyard 2000 Gewurztraminer, North Fork Lon…	Lemon oil and grapefruit aromas follow th
2	2	163528	Kunde 1999 Estate Bottled, Syrah, Sonoma Valley $23.	Bacon fat, black cherry, dill, oak aromas.
3	3	163529	Martha Clara Vineyard 1999 Syrah, North Fork of Long Islan…	Earthy, herbal, slightly herbaceous aroma
4	4	163530	Valiano 2000 Chianti Classico $13.99.	Cedar, cherry tomato, and herbal aromas
5	5	163531	Valiano 1998 Chianti Classico Riserva $23.99.	Dusty, cedary, cherry aromas. A rich entry
6	6	163532	Tenuta di Montecucco 1999 Le Coste, Sangiovese, Montecu…	Toasty oak, cherry and thyme aromas. A
7	7	163533	Piccini 2000 Superiore, Chianti $10.99.	Cherry, thyme, cedar aromas follow throu
8	8	163534	Piccini 1998 Patriarca di Piccini, Sangiovese, Tuscany $16.99.	Simple cherry and coffee ground aromas.
9	9	163535	Umberto Fiore 1999 "Torraltra", Dolcetto di Dogliani $15.99.	Dried herb, black cherry and cedar aroma
10	10	163536	Giacomo Brezza 1997 Cannubi, Barolo $63.99.	Cedar, truffle, tar, and sun-dried tomato a
11	11	163537	Giacomo Brezza 1997 "Sarmassa", Barolo $63.99.	Caraway seed, cedar and truffle aromas.
12	12	163538	Santini NV Moscato Bianco, Asti $6.99.	Sweet peach and floral aromas follow thr
13	13	163539	Giacomo Brezza 1997 Classico, Barolo $50.99.	Earthy, cedary, dried cherry aromas. A ric
14	14	163540	Cedar Mountain 1999 Amador County, Zinfandel, Amador C…	Port-like blackberry and cough syrup aro

Browsing the Data

Always good to wander around in your data

- Visual, interactive software tools like JMP make this painless
- Several quantitative variables were extracted from label
  - Regular expressions used to match patterns in data



wine.jmp

Regression Model for Price

Traditional multiple regression

Log(price) as response

Features alcohol, vintage, color, and points

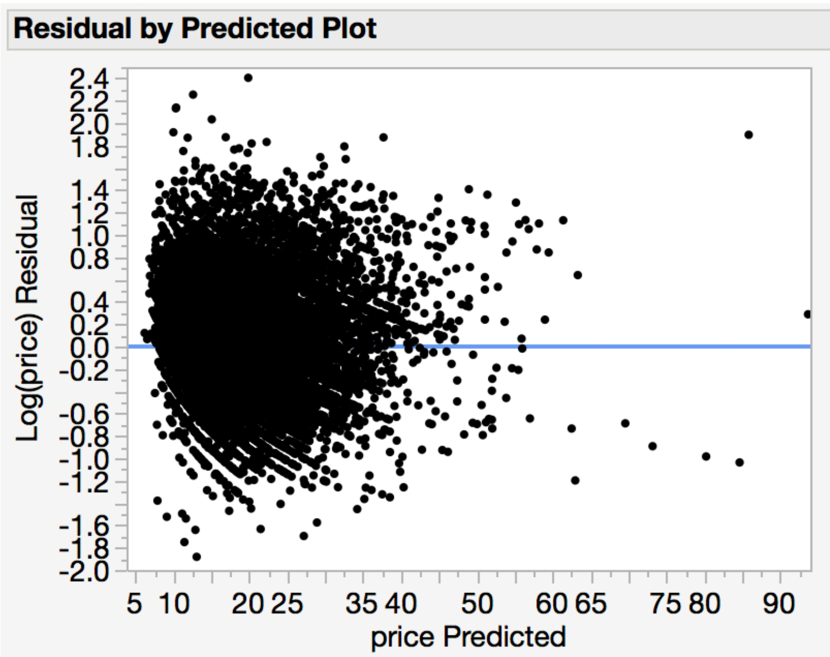
Too many varieties to use this one

With n=16,421, every feature is statistically significant

numerous missing prices

RSquare	0.320011
RSquare Adj	0.319804
Root Mean Square Error	0.476934
Mean of Response	2.893028
Observations (or Sum Wgts)	16421

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	78.879	2.400	32.86	<.0001*
alcohol	0.054	0.003	19.46	<.0001*
vintage	-0.042	0.001	-35.09	<.0001*
color[NA]	0.129	0.008	15.24	<.0001*
color[Red]	-0.044	0.006	-7.61	<.0001*
color[White]	-0.084	0.006	-13.78	<.0001*
points	0.092	0.001	71.81	<.0001*



Wharton
Department of Statistics

Be careful interpreting these…
the response is on a log scale.

What’s the benefit of text?

Does adding information gleaned from the tasting notes improve this regression?

- Is the model more predictive? Does R^2 grow?
- If so, can we interpret the effects of adding text?
- Analogous to using physician notes in diagnostic medicine

How can we find out? Two approaches

- Feature engineering: Hand-craft new variables
- At the moment Black Box: JMPs “Text Explorer" tool
    - We will look inside this tool in the coming lectures

Feature Engineering

Make new variables

Rationale for length of the tasting note: probably write more about a good wine than a crummy wine

Recode other features, particularly variety, to make useful

Indicators for "special" words: "yummy", "delicious", "great"

Sentiment analysis and no peeking at the response!

R^2 grows from 0.32 to 0.35

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	90.4312	2.4048	37.60	<.0001*
alcohol	0.0530	0.0028	19.16	<.0001*
vintage	-0.0477	0.0012	-39.64	<.0001*
color[NA]	0.1354	0.0085	15.89	<.0001*
color[Red]	-0.0788	0.0071	-11.17	<.0001*
color[White]	-0.0566	0.0070	-8.09	<.0001*
points	0.0819	0.0014	60.49	<.0001*
Length of Desc (words)	0.0082	0.0004	19.70	<.0001*
Variety[cabernet]	0.0566	0.0111	5.10	<.0001*
Variety[chardonnay]	-0.0267	0.0133	-2.00	0.0455*
Variety[merlot]	-0.0960	0.0131	-7.33	<.0001*
Variety[other]	-0.0323	0.0077	-4.19	<.0001*
Variety[pinot]	0.2466	0.0137	17.94	<.0001*
Variety[syrah]	-0.0248	0.0176	-1.41	0.1579
Variety[zinfandel]	-0.1233	0.0165	-7.46	<.0001*

Interesting to see effects of varieties

Going Deeper into Text

Explore the description more carefully

What other characteristics can be exploited?

What words, phrases are common enough to be "interesting"

What's a token?

Text Explorer for description

Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-empty Cases	Portion Non-empty per Case
5516	20507	703712	34.3157	20507	1.0000

Term and Phrase Lists

term = word type

Term	Count
aromas	18954
medium	16636
finish	11833
entry	9235
fruit	9135
body	9117
full	7950
bodied	7742
leads	6571
fruity	6406
acidity	6258
cherry	5757
dry	5743
palate	5636

Phrase	Count	N
medium body	5955	2
entry leads	5395	2
medium bodied	5106	2
aromas medium	4720	2
bodied palate	4705	2
fruity medium	3406	2
medium full	3313	2
dry yet	2852	2
yet fruity	2838	2
dry yet fruity	2828	3
full body	2739	2
full bodied	2548	2
yet fruity medium	2462	3
dry yet fruity medium	2454	4

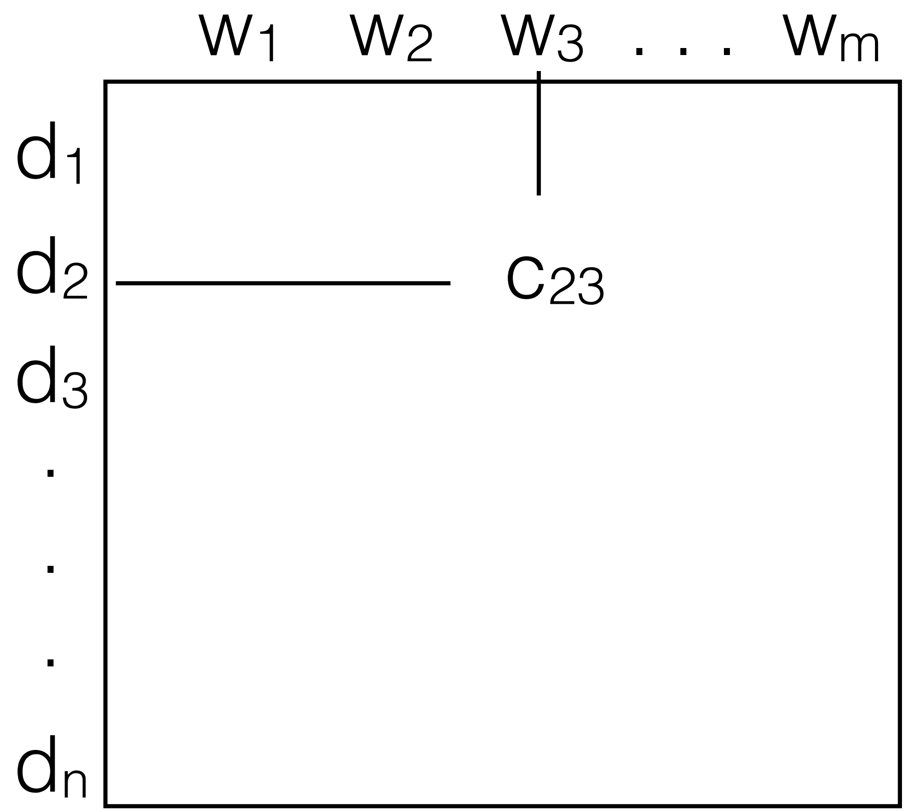
Author likes to use the word "medium" in a phrase.

Document Term Matrix

Count word types that appear in each document

One row for every document (an observation)

One column for every word type (a variable)



number of times word type w_3 appears in document 2

Document Term Matrix

Count word types that appear in each document

- What's a word?
- Where did common words like "a" and "the" go?
- Stemming? Are "herb" and "herbs" different words?
- Accept defaults for now, with explicit choices when using R

DTM is "huge"

- One row for every document, one column for every type
- Sparse: Most tokens are common, most types are rare
- Treat large matrix using idea from stat: Principal Components

Latent Semantic Analysis

LSA

Principal components analysis of the document term matrix

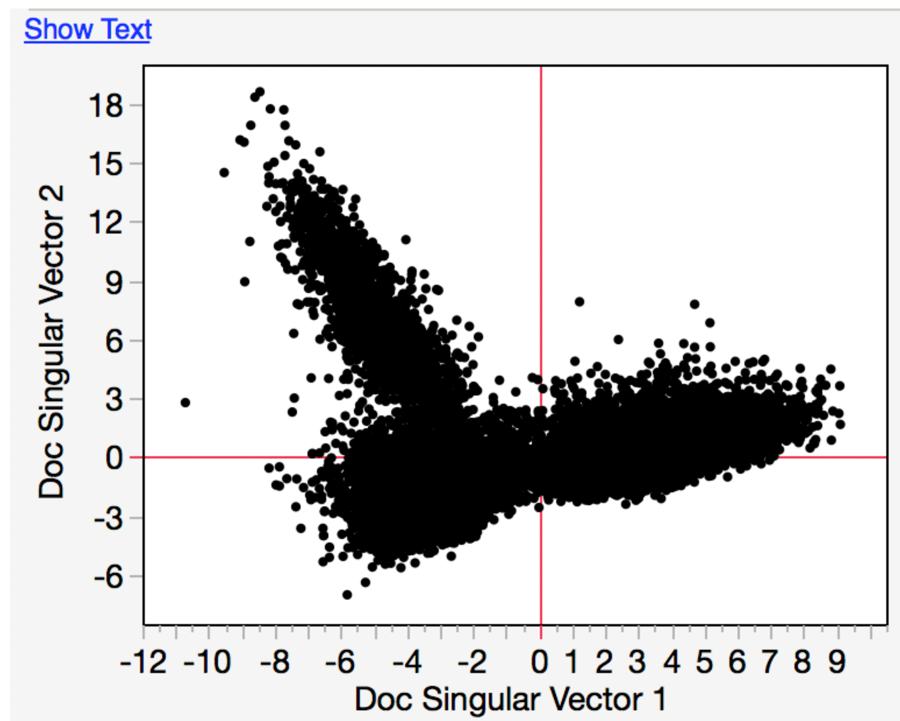
Variations based on how one normalizes the variables

just like standardizing variables in regression analysis

Default results

Specifications for Terms and Weights

Maximum Number of Terms	1703
Minimum Term Frequency	10
Weighting	TF IDF
Number of Singular Vectors	10
Centering and Scaling	Centered and Scaled



Do you see clusters???

Using the Principal Components

Add the principal components to the regression

Come back Tuesday and Wednesday to find out how this magic works and what those components mean.

The model improves again

R^2 grows from 0.32 to 0.35 to 0.40

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	104.51873	3.518718	29.70	<.0001*
alcohol	0.0430385	0.002752	15.64	<.0001*
vintage	-0.053975	0.001752	-30.80	<.0001*
color[NA]	0.1210612	0.008924	13.57	<.0001*
color[Red]	-0.11057	0.009522	-11.61	<.0001*
color[White]	-0.010491	0.007963	-1.32	0.1877
points	0.0688272	0.001864	36.93	<.0001*
Length of Desc (words)	0.0021044	0.000606	3.47	0.0005*
Variety[cabernet]	0.0315045	0.010779	2.92	0.0035*
Variety[chardonnay]	-0.053843	0.013069	-4.12	<.0001*
Variety[merlot]	-0.079787	0.012641	-6.31	<.0001*
Variety[other]	-0.035346	0.007473	-4.73	<.0001*
Variety[pinot]	0.2656546	0.013376	19.86	<.0001*
Variety[syrah]	-0.013322	0.01694	-0.79	0.4316
Variety[zinfandel]	-0.114862	0.015951	-7.20	<.0001*
Singular Vector 1	0.0043076	0.001636	2.63	0.0085*
Singular Vector 2	0.0355563	0.002044	17.40	<.0001*
Singular Vector 3	0.0289218	0.002608	11.09	<.0001*
Singular Vector 4	0.0129098	0.002441	5.29	<.0001*
Singular Vector 5	-0.013955	0.001625	-8.59	<.0001*
Singular Vector 6	0.0289976	0.00217	13.36	<.0001*
Singular Vector 7	0.0525495	0.002374	22.13	<.0001*
Singular Vector 8	0.012453	0.001823	6.83	<.0001*
Singular Vector 9	0.0232932	0.002337	9.97	<.0001*
Singular Vector 10	-0.001965	0.002459	-0.80	0.4242

Should we add more?

Next Steps

What's the science behind the success of using text?

"Description" features alone explain 28% of variation in price

Details, details…

Glossed over several choices

What's a word?
Do we keep all the words? What about phrases?
What's this singular value thing?

The choices might actually not matter, but you need to know what the choices are and why they might matter.

Software

JMP is pretty neat, but it does not implement some methods, such as sentiment analysis and topic models

Plus, its not free (at least not after a 30 day trial)