Introduction to Text Mining
Virtual Data Intensive Summer School
July 10, 2013

Dean Abbott
Abbott Analytics, Inc.
email: dean@abbottanalytics.com
url: http://www.abbottanalytics.com
blog: http://abbottanalytics.blogspot.com
Twitter: @deanabb

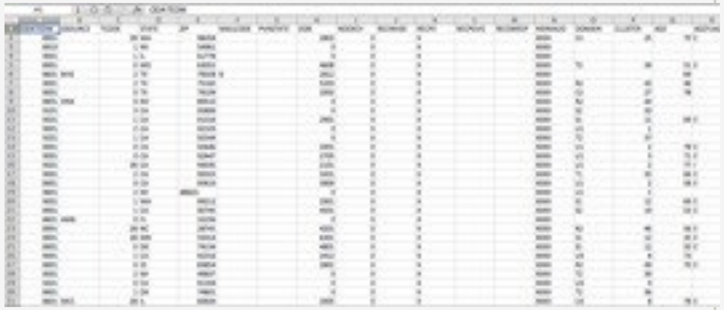
Why Text?

- How much data? 1.8 zettabytes (1.8 trillion GB)
- Most of the World's Data is Unstructured
  - 2009 HP survey: 70%
  - Gartner: 80%
  - Jerry Hill (Teradata), Anant Jhingran (IBM): 85%
- Structured (stored) data often misses elements critical to predictive modeling
  - Un-transcribed fields, notes, comments
  - Ex: examiner/adjuster notes, surveys with free-text fields, medical charts

Why Text Mining?

- Leveraging text ***should*** improve decisions and predictions
- Text mining is gaining momentum
  - Sentiment Analysis (twitter, facebook)
    - Predicting stock market
    - Predicting churn
    - Customer influence
  - Customer Service and Help Desk
- Not to mention Watson!

Structured vs. Unstructured Data

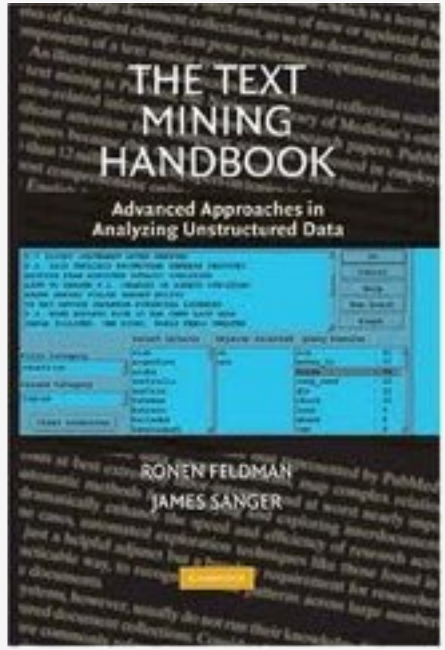


- Structured data
  - “Loadable into a spreadsheet”
    - Rows and columns
    - Each cell filled, or could be filled
    - Data is consistent, uniform
  - Data mining friendly

- Unstructured data
  - Microsoft Word, HTML, Adobe PDF documents, ...
    - This PPT document is unstructured text
    - Unstructured data often converted to XML -> semi-structured
  - Not structured into “cells”
    - Variable record length; notes, free-form survey answers
    - Text is relatively sparse, inconsistent, and not uniform
    - Also...images, video, music, etc.

How Unstructured is "Unstructured"?

- Feldman and Sanger
  - "Weakly Structured" data: few structural cues to text based on layout or markups
    - Research papers
    - Legal memoranda
    - News Stories
  - "Semistructured" data: extensive format elements, metadata, field labels
    - Email
    - HTML web pages
    - PDF files



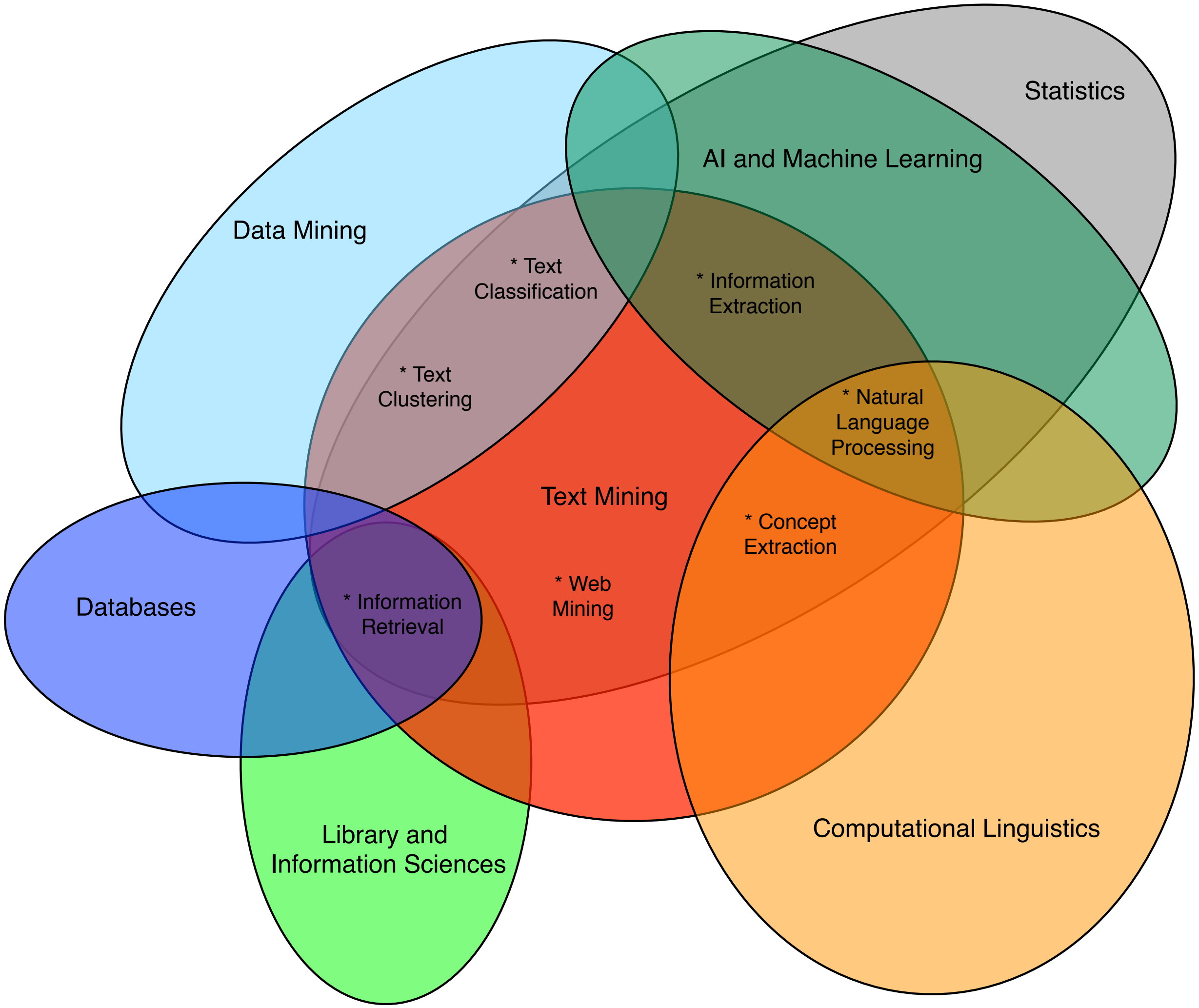
Why is Text Mining Hard

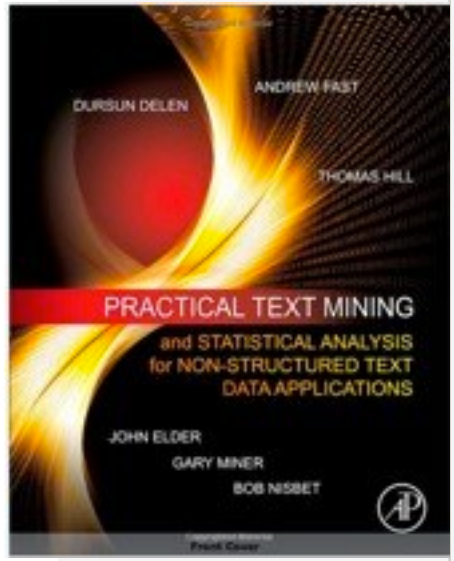
- Language is ambiguous
  - Context is needed to clarify
  - The same words can mean different things (homographs)
    - Bear (verb) - to support or carry
    - Bear (noun) - a large animal
  - Different words can mean the same thing (synonyms)
- Language is subtle
- Concept / Word extraction usually results in huge number of “dimensions”
  - Thousands of new fields
  - Each field typically has low information content (sparse)
- Mispellings, abbreviations, spelling variants
  - Renders search engines, SQL queries, Regex, ... ineffective

Four Text Mining Ambiguities

- **Homonomy**: same word, different meaning by accident of history
  - Bank
    - a. Mary walked along the bank of the river.
    - b. HarborBank is the richest bank in the city.
- **Synonymy**: synonyms, different words, similar or same meaning; can substitute one word for the other without changing the meaning of the sentence substantively.
  - Synonyms can have differing connotations...
    - a. Miss Nelson became a kind of big sister to Benjamin.
    - b. Miss Nelson became a kind of large sister to Benjamin.
- **Polysemy**: same word or form, but different, albeit related meaning
  - Bank
    - a. The bank raised its interest rates yesterday.
    - b. The store is next to the newly constructed bank.
    - c. The bank appeared first in Italy in the Renaissance.
- **Hyponymy**: concept hierarchy or subclass (subordinates)
  - Animal (noun)
    - a. dog
    - b. cat
  - Injury
    - a. Broken leg, contusion...

From Practical Text Mining (Delen, Fast, Hill, Miner, Elder, Nisbet)





Seven Types of Text Mining
(from Miner, Elder, et al)

Seven Types of Text Mining
(from Miner, Elder, et al)

1. **Search and Information Retrieval (IR)**: Storage and retrieval of text documents, including search engines and keyword search

Seven Types of Text Mining
(from Miner, Elder, et al)

1. **Search and Information Retrieval (IR)**: Storage and retrieval of text documents, including search engines and keyword search

2. **Document Clustering**: Grouping and categorizing terms, snippets, paragraphs or documents using data mining clustering methods

Seven Types of Text Mining
(from Miner, Elder, et al)

1. **Search and Information Retrieval (IR)**: Storage and retrieval of text documents, including search engines and keyword search

2. **Document Clustering**: Grouping and categorizing terms, snippets, paragraphs or documents using data mining clustering methods

3. **Document Classification**: Grouping and categorizing snippets, paragraphs, or document using data mining classification methods, based on models trained on labeled examples.

Seven Types of Text Mining
(from Miner, Elder, et al)

1. **Search and Information Retrieval (IR)**: Storage and retrieval of text documents, including search engines and keyword search

2. **Document Clustering**: Grouping and categorizing terms, snippets, paragraphs or documents using data mining clustering methods

3. **Document Classification**: Grouping and categorizing snippets, paragraphs, or document using data mining classification methods, based on models trained on labeled examples.

4. **Web Mining**: Data and Text Mining on the Internet with a specific focus on the scale and interconnectedness of the web.

Seven Types of Text Mining
(from Miner, Elder, et al)

1. **Search and Information Retrieval (IR)**: Storage and retrieval of text documents, including search engines and keyword search

2. **Document Clustering**: Grouping and categorizing terms, snippets, paragraphs or documents using data mining clustering methods

3. **Document Classification**: Grouping and categorizing snippets, paragraphs, or document using data mining classification methods, based on models trained on labeled examples.

4. **Web Mining**: Data and Text Mining on the Internet with a specific focus on the scale and interconnectedness of the web.

5. **Information Extraction (IE)**: Identification and extraction of relevant facts and relationships from unstructured text; the process of making structured data from unstructured and semi-structured text

Seven Types of Text Mining
(from Miner, Elder, et al)

1. **Search and Information Retrieval (IR)**: Storage and retrieval of text documents, including search engines and keyword search
2. **Document Clustering**: Grouping and categorizing terms, snippets, paragraphs or documents using data mining clustering methods
3. **Document Classification**: Grouping and categorizing snippets, paragraphs, or document using data mining classification methods, based on models trained on labeled examples.
4. **Web Mining**: Data and Text Mining on the Internet with a specific focus on the scale and interconnectedness of the web.
5. **Information Extraction (IE)**: Identification and extraction of relevant facts and relationships from unstructured text; the process of making structured data from unstructured and semi-structured text
6. **Natural Language Processing (NLP)**: Low-level language processing and understanding tasks (e.g., tagging part of speech); often used synonymously with computational linguistics

Seven Types of Text Mining

(from Miner, Elder, et al)

1. **Search and Information Retrieval (IR)**: Storage and retrieval of text documents, including search engines and keyword search

2. **Document Clustering**: Grouping and categorizing terms, snippets, paragraphs or documents using data mining clustering methods

3. **Document Classification**: Grouping and categorizing snippets, paragraphs, or document using data mining classification methods, based on models trained on labeled examples.

4. **Web Mining**: Data and Text Mining on the Internet with a specific focus on the scale and interconnectedness of the web.

5. **Information Extraction (IE)**: Identification and extraction of relevant facts and relationships from unstructured text; the process of making structured data from unstructured and semi-structured text

6. **Natural Language Processing (NLP)**: Low-level language processing and understanding tasks (e.g., tagging part of speech); often used synonymously with computational linguistics

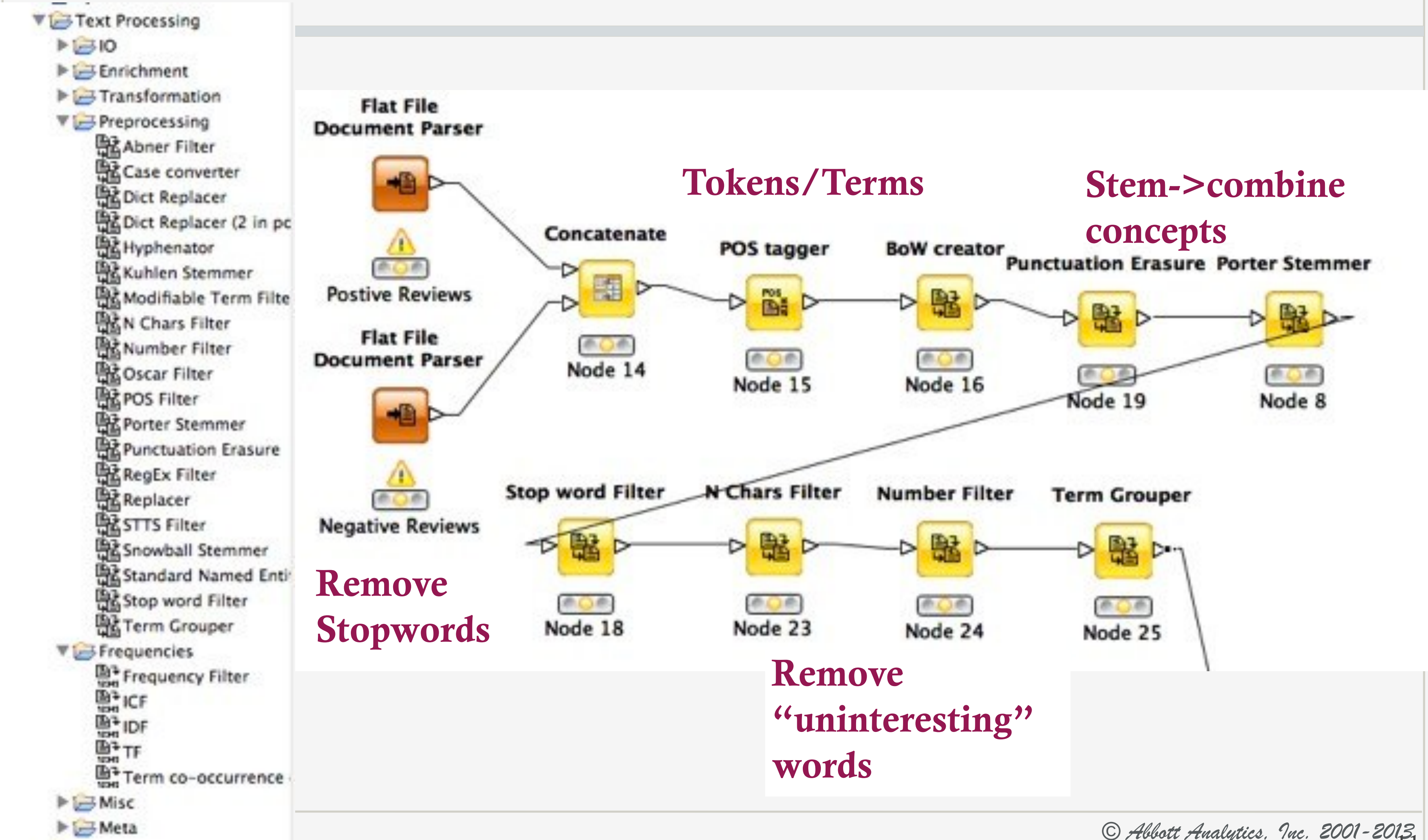
7. **Concept Extraction**: Grouping of words and phrases into semantically similar groups

Text Mining Process Flow (from Miner, Elder, *et al*)

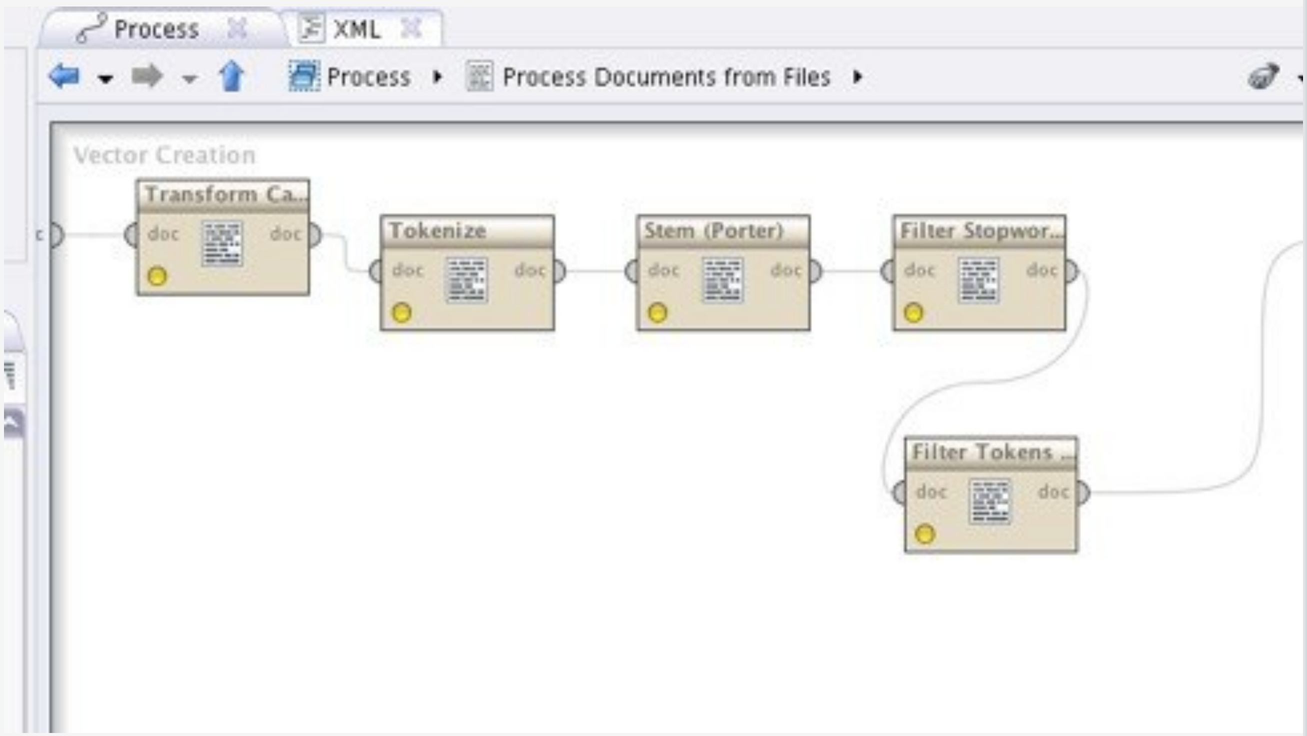
1. Phase 1. Determine the purpose of the study
2. Phase 2. Explore the availability and the nature of the data
3. Phase 3. Prepare the Data.
   (details in next slide)
4. Phase 4. Develop and Assess the Models
5. Phase 5. Evaluate the results
6. Phase 6. Deploy the results

Note the similarity between this process flow and CRISP-DM

Text Mining Pre-Processing Steps in KNIME



Text Mining Pre-Processing Steps in RapidMiner



Select Data Source

- Most tools support
  - Single Text File
    - Each record contains a different "comment"
  - Folders containing multiple text files of a particular type->load multiple folders
  - RSS Feeds and URLs (spider a site)
  - Documents (.doc, .pdf, .xml)

Text Normalization

- Case
  - Make all lower case (if don't care about proper nouns, titles, etc.)
- Clean up transcription and typing errors
  - do n't, movei, ...
- Correct mispelled words
  - Phonetically
    - Use fuzzy matching algorithms such Soundex, Metaphone, or string-edit distance
  - Dictionaries
    - Use POS and context to make good guess

Part of Speech (POS) Tagging

- Useful for recognizing names of people, places, organizations, titles
- English language
  - Minimum set includes noun, verb, adjective, adverb, prepositions, conjunctions
  - Penn Tree Bank contains 36 POS

POS Tags from Penn Tree Bank

Number	Tag	Description
1	CC	Coordinating conjunction
2	CD	Cardinal number
3	DT	Determiner
4	EX	Existential *there*
5	FW	Foreign word
6	IN	Preposition or subordinating conjunction
7	JJ	Adjective
8	JJR	Adjective, comparative
9	JJS	Adjective, superlative
10	LS	List item marker
11	MD	Modal
12	NN	Noun, singular or mass
13	NNS	Noun, plural
14	NNP	Proper noun, singular
15	NNPS	Proper noun, plural
16	PDT	Predeterminer
17	POS	Possessive ending
18	PRP	Personal pronoun

Number	Tag	Description
19	PRP$	Possessive pronoun
20	RB	Adverb
21	RBR	Adverb, comparative
22	RBS	Adverb, superlative
23	RP	Particle
24	SYM	Symbol
25	TO	*to*
26	UH	Interjection
27	VB	Verb, base form
28	VBD	Verb, past tense
29	VBG	Verb, gerund or present participle
30	VBN	Verb, past participle
31	VBP	Verb, non-3rd person singular present
32	VBZ	Verb, 3rd person singular present
33	WDT	Wh-determiner
34	WP	Wh-pronoun
35	WP$	Possessive wh-pronoun
36	WRB	Wh-adverb

Example of Brill Tagging

Example of Brill Tagging

- In this talk, Mr. Pole discussed how Target was using Predictive Analytics including descriptions of using potential value models, coupon models, and...yes...predicting when a woman is due.

Example of Brill Tagging

- In this talk, Mr. Pole discussed how Target was using Predictive Analytics including descriptions of using potential value models, coupon models, and...yes...predicting when a woman is due.

- In/IN this/DT talk/NN ,/, Mr./NNP Pole/NNP discussed/VBD how/WRB Target/NNP was/VBD using/VBG Predictive/NNP Analytics/NNP including/VBG descriptions/NNS of/IN using/VBG potential/JJ value/NN models/NNS ,/, coupon/NN models/NNS ,/, and...yes...predicting/VBG when/WRB a/DT woman/NN is/VBZ due/JJ./.

Example of Brill Tagging

/DT
determiner
/IN
preposition
/JJ
adjective
/NN
noun
/NNP
proper noun
/NNS
plural noun
/PRP
possessive
pronoun
/VBD
verb,
past tense
/VBZ
verb,
3rd prsn
/WRB
Wh adverb

- In this talk, Mr. Pole discussed how Target was using Predictive Analytics including descriptions of using potential value models, coupon models, and…yes…predicting when a woman is due.

- In/IN this/DT talk/NN ,/, Mr./NNP Pole/NNP discussed/VBD how/WRB Target/NNP was/VBD using/VBG Predictive/NNP Analytics/NNP including/VBG descriptions/NNS of/IN using/VBG potential/JJ value/NN models/NNS ,/, coupon/NN models/NNS ,/, and…yes…predicting/VBG when/WRB a/DT woman/NN is/VBZ due/JJ./.

POS Tagging: How Hard is it?

- ~89% of English words have only one part of speech (unambiguous)
  - However, many common words in English are ambiguous
  - But even these can largely be disambiguated by rules or probabilistically
- Taggers can be rule-based, stochastic (training on a labelled set of words using HMMs), or a combination (most popular combination is the "Brill" tagger)
- Example of stochastic tagging
  - NNP VBZ VBN TO VB NR
    Secretariat is expected to race tomorrow
  - NNP VBZ VBN TO NN NR
    Secretariat is expected to race tomorrow
  - P(NN|TO) = 0.00047
  - P(VB|TO) = 0.83 -> "race" is most likely a verb

POS Tagging: How Hard is it?

- ~89% of English words have only one part of speech (unambiguous)
  - However, many common words in English are ambiguous
  - But even these can largely be disambiguated by rules or probabilistically

- Taggers can be rule-based, stochastic (training on a labelled set of words using HMMs), or a combination (most popular combination is the “Brill” tagger)

/NNP
proper noun
/VB
Base verb
/VBN
verb,
past participle
/VBZ
verb,
3rd prsn
/TO
to

- Example of stochastic tagging
  - NNP VBZ VBN TO VB NR
    Secretariat is expected to race tomorrow
  - NNP VBZ VBN TO NN NR
    Secretariat is expected to race tomorrow

  - P(NN|TO) = 0.00047
  - P(VB|TO) = 0.83 -> “race” is most likely a verb

Tokenization

- Convert streams of characters into "words"
- Main clues (in English): white space.
- Words can contain special characters, such as these: . , ' – etc.
  - Examples: Dr.   O'Malley 555-1212
- No single algorithm "works" always
- Some languages do not have white space
  - Chinese, Japanese, Korean; German compound nouns

Tokenization Example (from KNIME)

i find zero effect to be an immensely funny and witty film .
kasdan's humor is of the best kind -- soft spoken , and mostly dialogue-driven (though there are som
it's the kind of humor that's funny even after you've seen it five or six times .
there's one scene in which zero talks about how detached he is , an how that makes him such a great d
what we see during this narration are various shots of him sitting on a bed , or standing motionlessl
unshaven face , his eyes pointing to something off camera , but obviously to nothing in particular .
i can't convey to you how funny this is , but what makes it great film making is that it has a point
characterization of zero .
as a side note , i don't consider myself an average viewer when it comes to comedy (not to sound eli

narration [RB (POS)]	"zero effect gets its title from the main character , daryl zero (bill pullman) , although we do n't understand what it truly means until the very last line of dialogue in the film ."

- Narration is split out as a token
- Part of Speech is listed as "adverb"
  - The (POS) means it was part of the corpus of documents I labelled as "positive reviews"
- Others: word--it, --quite, son+s

Stemming

- Normalizes / unifies variations of the same idea
  - "walking", "walks", "walked", "walker" => "walk".
- Inflectional Stemming
  - Remove plurals
  - Normalize verb tenses
  - Remove other affixes
- Stemming to root
  - Reduce word to most basic element
  - More aggressive than inflectional
  - Examples
    - "denormalization" -> "norm"
    - "apply", "applications", "reapplied" -> "apply"

Stemming Example (from KNIME)

i find zero effect to be an immensely funny and witty film .
kasdan's humor is of the best kind -- soft spoken , and mostly dialogue-driven (though there are som
it's the kind of humor that's funny even after you've seen it five or six times .
there's one scene in which zero talks about how detached he is , an how that makes him such a great d
what we see during this narration are various shots of him sitting on a bed , or standing motionlessl
unshaven face , his eyes pointing to something off camera , but obviously to nothing in particular .
i can't convey to you how funny this is , but what makes it great film making is that it has a point
characterization of zero .
as a side note , i don't consider myself an average viewer when it comes to comedy (not to sound eli

narrat[RB (POS)]	"zero effect get it titl from the main charact daryl zero bill pullman although we do nt understand what it truli mean until the veri last line of dialogu in the film"	"zero effect gets its title from the main character daryl zero bill pullman although we do nt understand what it truly means until the very last line of dialogue in the film"

- narration becomes narrat
- title becomes titl
- character becomes charact

Common English Stop Words

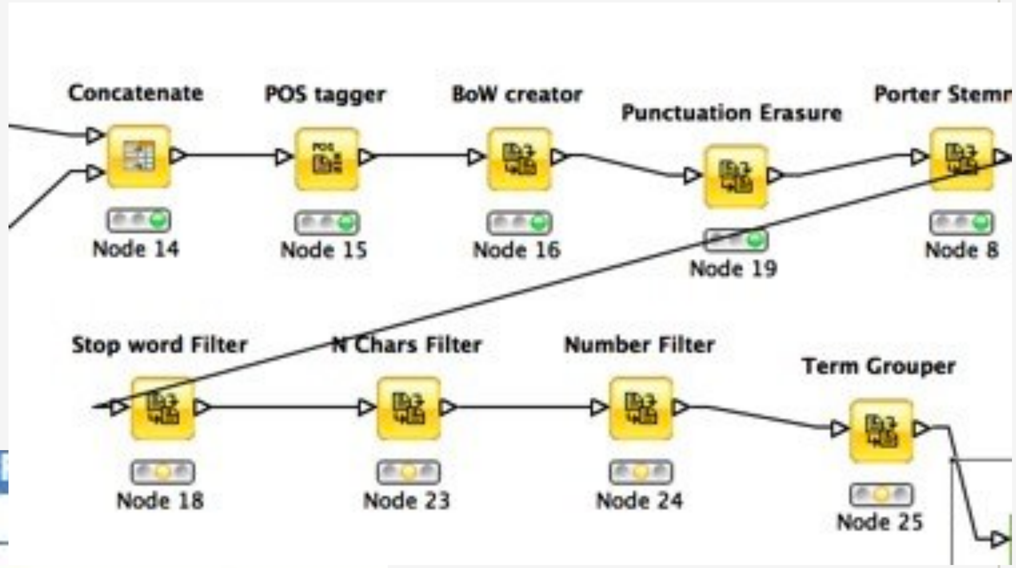
- a, an, and, are, as, at, be, but, by, for, if, in, into, is, it, no, not, of, on, or, such, that, the, their, then, there, these, they, this, to, was, will, with
- Stop words are very common and rarely provide useful information for information extraction or concept extraction.
- Removing stop words also reduces dimensionality

Dictionaries and Lexicons

- Highly recommended!! Can be **very** time consuming...
- Reduces set of keywords to focus on
    - Words of interest
    - Dictionary words
- Increase set of keywords to focus on
    - Proper nouns, special names/phrases
    - Acronyms
    - Titles
    - Numbers
- Key ways to use dictionary
    - Local dictionary (specialized words)
    - Stopwords and "too frequent" words
    - Stemming: reduce stems to dictionary words
    - Synonyms: replace synonyms with root word in list
    - Resolve Abbreviations and Acronyms

What Counts Can Look Like: # Records == # Tokens (“Words”)

In KNIME:
200+ documents turn into more than 70K Term-Document pairs
(average of >350 terms per review



Step	Num Records	
Load Files	203	
Bag of Words	72349	
Remove Punctuation	70801	2.1%
Porter Stemming	70652	0.2%
Stop Word Filtering	49212	30.3%
N char filter	47350	3.8%
Number Filter	47168	0.4%
Term Grouper	43123	8.6%
Document Vector	203	99.5%

We've Got the Terms… Now What?

Tag Cloud of Movie Review Terms

01-2013

Text Mining Features

- Keywords
  - Keyword Flags: Bag of Words flags
  - TF: Counts of keywords in field
  - Binned counts (doesn't exist, exists once, exists more than once)
  - IDF
    - IDF: Inverse Document Frequency = log(1 + NumDocs / NumDocs with Term)
  - TF*IDF
- Multiple-word phrases: n-grams
- Reduced dimensionality features: PCA

Keyword ("Bag of Words") Representation as Binary Flag in KNIME

File

Table "default" – Rows: 203 | Spec – Columns: 10488 | Properties

Row ID	Docum...	appar[]	highlig...	art[]	action[]	perfect[]	filmma...	site[]	fight[]	mere[]
1	""	1	1	1	1	1	1	1	1	1
2	""	0	0	0	0	0	0	0	0	0
3	"000-foot...	0	0	0	1	0	0	1	0	1
4	"accept os...	0	0	0	0	0	0	0	0	0
5	"ado noth"	1	0	0	0	0	0	0	0	0
6	"african a...	1	0	0	1	1	0	0	0	0
7	"ago japa...	0	0	0	1	0	1	0	1	1
8	"ago john...	0	0	0	1	0	0	0	0	0
9	"ahh teen...	0	1	0	0	0	0	0	0	0
10	"america l...	1	0	0	1	0	0	0	0	0
11	"american...	0	0	1	1	0	0	0	1	0
12	"american...	0	0	0	0	1	0	0	1	0
13	"anniversa...	0	0	0	0	0	0	0	0	0
14	"anoth for...	0	0	0	1	0	0	0	0	1
15	"anoth thi...	1	0	0	1	0	0	0	0	0
16	"anticip sa...	0	0	0	1	0	0	0	1	0
17	"appar dir...	1	0	0	0	1	0	0	1	0
18	"averag te...	0	0	0	0	0	0	0	0	0
19	"beatl nob...	0	0	0	0	0	1	0	0	0
20	"befor re...	0	0	0	0	1	0	0	0	0
21	"believ re...	0	0	0	0	0	0	0	0	0
22	"blade mo...	0	0	0	0	0	0	0	0	0

Term Frequency: TF

- # Times the term occurs in a document, d
- Assumptions:
  - If term occurs more often, it measures something important
  - 2x as many occurrences is 2x as important
    - This can be mitigated if need be
      - Common “fix”: log transform, log10(1+TF)
  - Each occurrence is an independent event (not a replicate)
  - Is it true?
    - Information retrieval: probably “yes”
    - Fraud detection, notes, log files: maybe “no”

Document Frequency: DF

- # Documents the term occurs in
- Assumptions:
  - Terms that occur in fewer documents are more specified to a document and more descriptive of the content: rarity matters
  - Terms that occur in most documents are common words, not as descriptive
  - Is it true?
    - Sometimes "yes"
    - Sometimes just reflect textual variants (synonyms), regional differences, personal style

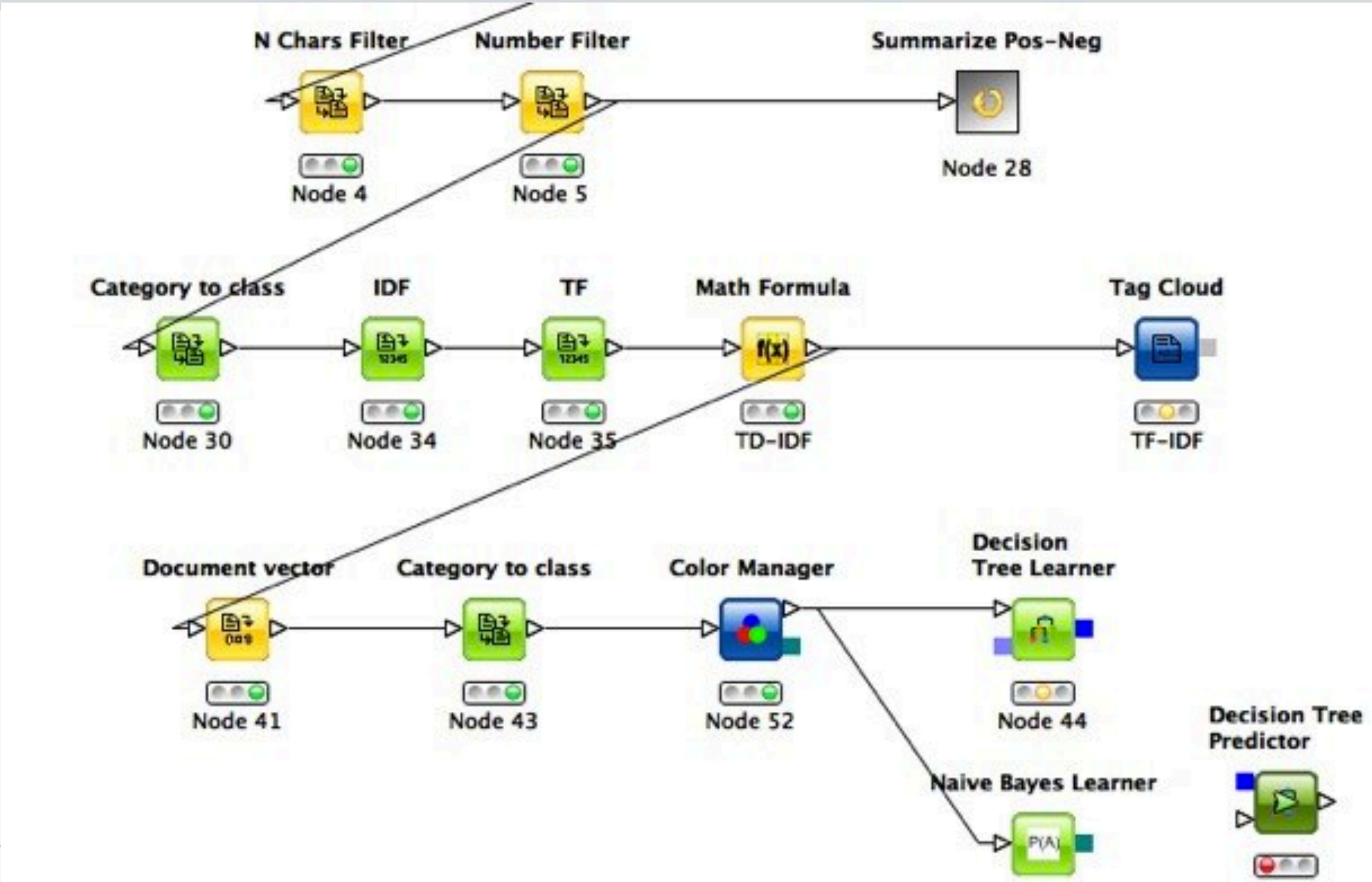
Inverse Document Frequency: IDF

- DF: smaller is better
  - We often want a larger number to be “better”
- Solution: IDF
  - Too severe if we define
    IDF = # docs / # docs with term
  - Typical definition:
    IDF = log10(1 + # docs / # docs with term)
  - Why? Because it seems to work better

Tying Them Together: TF-IDF

- Separately, DF and IDF can be good features
- Together, they represent a good idea
  - TF-IDF = TF * IDF
  - Higher frequency of terms that are rare may indicate a very important concept
  - Why multiply? Are these “independent”?
    - No, but multiplying seems to work just fine

Building Document Classification Models in



Wednesday, July 10, 13

What TF-IDF Looks Like:
>> By Term

Filtered table – 2:128 – Column Filter

File

Table "default" – Rows: 6842 | Spec – Columns: 8 | Properties | Flow Variables

...	Term	TotalHits	PctPosi...	PctNeg...	Document class	IDF	TF a...	TFIDF
...	john[NN(POS)]	35	57.143	42.857	POSITIVE	0.831	15	12.46
...	fight[NN(POS)]	21	42.857	57.143	NEGATIVE	1.026	12	12.313
...	films[NNS(POS)]	87	54.023	45.977	POSITIVE	0.521	11	5.735
...	star[NN(POS)]	36	61.111	38.889	POSITIVE	0.82	10	8.203
...	little[RB(POS)]	44	47.727	52.273	NEGATIVE	0.747	10	7.475
...	little[JJ(POS)]	59	59.322	40.678	NEGATIVE	0.646	10	6.458
...	characters[NNS(POS)]	94	42.553	57.447	POSITIVE	0.498	10	4.982
...	city[NN(POS)]	27	55.556	44.444	POSITIVE	0.928	9	8.356
...	john[NN(POS)]	35	57.143	42.857	NEGATIVE	0.831	9	7.476
...	home[NN(POS)]	38	55.263	44.737	POSITIVE	0.8	9	7.204
...	characters[NNS(POS)]	94	42.553	57.447	POSITIVE	0.498	9	4.483
...	hard[RB(POS)]	22	59.091	40.909	NEGATIVE	1.008	8	8.063
...	sex[NN(POS)]	23	30.435	69.565	NEGATIVE	0.99	8	7.924
...	city[NN(POS)]	27	55.556	44.444	NEGATIVE	0.928	8	7.428
...	guy[NN(POS)]	45	46.667	53.333	POSITIVE	0.739	8	5.916
...	comedy[NN(POS)]	48	43.75	56.25	POSITIVE	0.717	8	5.734
...	films[NNS(POS)]	87	54.023	45.977	POSITIVE	0.521	8	4.171
...	fight[NN(POS)]	21	42.857	57.143	NEGATIVE	1.026	7	7.183
...	family[NN(POS)]	23	78.261	21.739	NEGATIVE	0.99	7	6.933

What TF-IDF Looks Like:
>> By Document

Documents output table - 2:41 - Document vector

File

Table "default" - Rows: 202 | Spec - Columns: 648 | Properties | Flow Variables

Row ID	D movie(...	D watchi...	D charac...	D worst(J...	D acting(...	D superb...	D guy(N...	D reason...	D variou...	D fight(N...	D nor(C...	D story(...
318	10.467	0	1.307	0.928	0.756	0	0	0	0	6.157	0	0.881
319	2.617	0	1.742	0.928	0	0	0	0	0	0	0	0.44
320	1.495	0	0.436	0	0	0	0	0	0	2.052	0	0
321	1.121	1.188	0	0	0.756	0	0	0	0	0	0	0
322	1.121	0	1.307	0	0	0	0	0.888	1.087	0	0	0.44
323	1.869	0	0.436	0	0	0	0.739	0	0	3.078	0	0
324	0.748	0	0.436	0	0.756	0	0	0.888	0	0	0	0
325	1.121	0	0	0	0	0	1.479	0	0	0	0	0.44
326	1.495	0	0	0.928	0	0	0	0	0	0	0	0.44
327	1.869	0	0	0	0	0	0	0	0	0	0	0
328	1.495	0	0.871	0	0	0	0	0	0	1.026	0	0
329	1.121	0	0	0.928	0	0	0	0	0	0	0	0
330	0	0	0	0	0	0	0	0	1.087	0	0	0
331	0.374	0	0.436	0	0	0	0	0	0	0	0	0
332	5.981	0	0.871	0	0	0	0	0	0	0	0	0
333	0.374	0	0.436	0	0	0	0.739	0	0	0	0	0
334	1.121	0	0	0	0	1.218	0	0	0	0	0	0
335	0.374	0	0.436	0	0	0	0	0	1.087	3.078	0	1.321
336	1.495	1.188	0.871	0	0	0	0	0.888	0	0	0	0
337	2.617	0	0	0	1.511	0	2.958	1.777	0	0	0	0.44
338	1.121	0	0.436	0	0	0	0.739	0	0	0	0	0
339	2.243	0	0.871	0	0	0	0	0	0	0	0	0
340	0.374	0	1.742	0	1.511	1.218	0.739	0	0	0	0	0
341	2.243	0	0	0	0.756	0	0	0	0	0	1.087	0.44
342	1.121	0	0.436	0	0	0	0	0	0	0	0	0
343	0	1.188	0	0	0	0	0	0.888	0	0	0	0.44
344	4.112	0	0.871	0	0.756	0	0	0	0	0	1.087	0
345	0.748	0	1.307	0	0	0	0	0	0	0	0	0

Multi-Word Features: N-Grams

Multi-Word Features: N-Grams

- N-Grams
  - Combinations of characters or words
  - "N" means how many character or word groups you identify and extract
    - 2-grams (bigrams, digrams): "vice president"
    - 3-grams (trigrams): "central intelligence agency"
    - 4-grams: "united states of america"

Multi-Word Features: N-Grams

- N-Grams
  - Combinations of characters or words
  - "N" means how many character or word groups you identify and extract
    - 2-grams (bigrams, digrams): "vice president"
    - 3-grams (trigrams): "central intelligence agency"
    - 4-grams: "united states of america"
- Constructing bigrams: a simple model, P(A|B)
  - Example Corpus (from Jurafsky and Martin):
    - <s> I am Sam </s>
    - <s> Sam I am </s>
    - <s> I do not like green eggs and ham </s>
    - P(I|<s>) = 2/3; P(Sam|<s>) = 1/3; P(Sam|am) = 1/2; P(am|Sam)= 0/2, etc.

Multi-Word Features: N-Grams

- N-Grams
  - Combinations of characters or words
  - "N" means how many character or word groups you identify and extract
    - 2-grams (bigrams, digrams): "vice president"
    - 3-grams (trigrams): "central intelligence agency"
    - 4-grams: "united states of america"
- Constructing bigrams: a simple model, P(A|B)
  - Example Corpus (from Jurafsky and Martin):
    - <s> I am Sam </s>
    - <s> Sam I am </s>
    - <s> I do not like green eggs and ham </s>
    - P(I|<s>) = 2/3; P(Sam|<s>) = 1/3; P(Sam|am) = 1/2; P(am|Sam)= 0/2, etc.
- More sophisticated: allow gaps between words

1-Gram (Term) Relationship to Sentiment

For 100 Reviews, 25/27 with "worst" are negative

Filtered – 6:100:30 – Row Filter

File

Table "default" – Rows: 647 | Spec – Columns: 4 | Properties | Flow Variables

Text

Row ID	Term	Tota...	NumN...	PctPositive
Row15412	supposed[VBN(POS)]	13	13	0
Row9829	mess[NN(POS)]	14	13	7.143
Row1750	worst[JJS(POS)]	27	25	7.407
Row4647	dull[JJ(POS)]	13	12	7.692
Row1703	boring[JJ(POS)]	11	10	9.091
Row6413	girls[NNS(POS)]	11	10	9.091
Row11778	poor[JJ(POS)]	16	14	12.5
Row1122	bad[RB(POS)]	14	12	14.286
Row1804	break[NN(POS)]	14	12	14.286
Row13095	ridiculous[JJ(POS)]	13	11	15.385
Row11743	pointless[JJ(POS)]	12	10	16.667
Row15244	stupid[JJ(POS)]	22	18	18.182
Row11327	pathetic[JJ(POS)]	11	9	18.182
Row12349	quick[JJ(POS)]	11	9	18.182
Row17121	waste[NN(POS)]	11	9	18.182
Row1701	boring[VBG(POS)]	15	12	20

Sample Bigrams in Movie Review Data

Table View - 2:15 - Interactive Table (44 x 6)

File Hilite Navigation View Output

Row ID	S Word1	S Word2	I Sum(C...	I Sum(P...	D Sum(P...	D Pos...
Row1583	some	reason	52	1	15	28.846
Row2600	worse	than	52	1	15	28.846
Row877	have	some	63	1	18	28.571
Row650	figure	out	71	1	20	28.169
Row31	action	scenes	80	1	22	27.5
Row302	bad	guy	93	1	25	26.882
Row600	even	worse	58	1	15	25.862
Row1758	the	bad	138	1	35	25.362
Row1554	should	have	191	1	46	24.084
Row2402	van	damme	66	1	15	22.727
Row1354	only	thing	81	1	16	19.753
Row2210	the	worst	207	1	38	18.357
Row304	bad	movie	62	0	0	0
Row2360	too	bad	60	0	0	0

- Negative Reviews: 169 documents with “the worst”
- Positive Reviews: 38 documents with “the worst”
  - Negative Rate for “worst”: 199/262 = 76.0%
  - Negative Rate for “the worst” = 169/207 = 81.6%

The Obvious

The Obvious

casper van dien has the chiseled facial features and tan skin which will make mo
st girls swoon .
to most guys , he will come across as artificial .
fortunately , van dien is the worst of the cast , despite how hard he tries to c
ome off as a real character .

The Obvious

casper van dien has the chiseled facial features and tan skin which will make mo
st girls swoon .
to most guys , he will come across as artificial .
fortunately , van dien is the worst of the cast , despite how hard he tries to c
ome off as a real character .

despite all them flashy effects and big explosions , deep rising is still , at h
eart , a good 'ol b movie .
luckily , it's a very good b movie .
the worst cliches in movie history are a b movie's bread and butter .
therefore , things that would destroy a serious movie actually help us have a go
od time while watching a movie of lower calibre .

The Obvious

casper van dien has the chiseled facial features and tan skin which will make mo
st girls swoon .
to most guys , he will come across as artificial .
fortunately , van dien is the worst of the cast , despite how hard he tries to c
ome off as a real character .

despite all them flashy effects and big explosions , deep rising is still , at h
eart , a good 'ol b movie .
luckily , it's a very good b movie .
the worst cliches in movie history are a b movie's bread and butter .
therefore , things that would destroy a serious movie actually help us have a go
od time while watching a movie of lower calibre .

it is a real shame to see kelly , definitely the worse actor than saxon , to ste
al the scenes from him only because his lines , being the worst possible blaxplo
itation cliches , sound so damn over the top .
other actors , not including shih kien who turns han into typical , although not
very convincing bondian villain , are nothing more than fist fodder for bruce l
ee (among them is young jackie chan) .

Ambiguities

Ambiguities

the film has gotten some negative reviews (a friend of mine actually thinks it's the worst in the series) , but i'm not really sure why .
it's an exciting , often hilarious movie that engaged me and left me ready for the next star trek film .
some say it's a bit too light , and more of a long episode than a film .
others say the special effects are cheesy and that it's boring .
i simply enjoyed the film .

Ambiguities

the film has gotten some negative reviews (a friend of mine actually thinks it'
s the worst in the series) , but i'm not really sure why .
it's an exciting , often hilarious movie that engaged me and left me ready for t
he next star trek film .
some say it's a bit too light , and more of a long episode than a film .
others say the special effects are cheesy and that it's boring .
i simply enjoyed the film .

in a wonderfully executed performance by ian michael smith , simon birch is conv
inced that he's god's instrument , and there must be a reason he's so small .
joe doesn't quite buy it , but he sticks by simon ; the two tough it out through
the worst of times and are each other's only true friends .

Ambiguities

the film has gotten some negative reviews (a friend of mine actually thinks it'
s the worst in the series) , but i'm not really sure why .
it's an exciting , often hilarious movie that engaged me and left me ready for t
he next star trek film .
some say it's a bit too light , and more of a long episode than a film .
others say the special effects are cheesy and that it's boring .
i simply enjoyed the film .

in a wonderfully executed performance by ian michael smith , simon birch is conv
inced that he's god's instrument , and there must be a reason he's so small .
joe doesn't quite buy it , but he sticks by simon ; the two tough it out through
the worst of times and are each other's only true friends .

all in all , " scream 3 " does not fall victim to the most lamented principle of
a trilogy : it is not the worst of the films .
as a trilogy , " scream " was a lot of fun ; refreshing , humorous , offbeat .
almost sad to see it go .

Ambiguities

the film has gotten some negative reviews (a friend of mine actually thinks it's the worst in the series) , but i'm not really sure why .
it's an exciting , often hilarious movie that engaged me and left me ready for the next star trek film .
some say it's a bit too light , and more of a long episode than a film .
others say the special effects are cheesy and that it's boring .
i simply enjoyed the film .

in a wonderfully executed performance by ian michael smith , simon birch is convinced that he's god's instrument , and there must be a reason he's so small .
joe doesn't quite buy it , but he sticks by simon ; the two tough it out through the worst of times and are each other's only true friends .

all in all , " scream 3 " does not fall victim to the most lamented principle of a trilogy : it is not the worst of the films .
as a trilogy , " scream " was a lot of fun ; refreshing , humorous , offbeat .
almost sad to see it go .

he found the film to be a `sea of sugary bromides' and condemned mr . voight's character as `hopelessly wooden ? adopts an accent even more indeterminate than the one he came up with for anaconda . '
in addition , `entertainment weekly' slam-dunked the film , condemning it as `the worst family film of the year . '
there have been so many other bad reviews like this , too .
my suggestion : disregard the critics .

What Regular Expressions Do

- Match strings: REs are pattern matchers, not “language interpreters”
- Are **more flexible** than string operations in Excel, C, or other languages
- Key: wildcards
  - Allows matching varying length patterns
  - Allows matching existence or non-existence of patterns
  - Matches combinations very efficiently
- have rules (assumptions) that describe how they match
  - Typically, they start at the “left” end of a chunk of text, and match the ***leftmost*, longest** string

Adding Flexibility with Regular Expressions

Adding Flexibility with Regular Expressions

- REs match string patterns; they are not “language interpreters”

Adding Flexibility with Regular Expressions

- REs match string patterns; they are not “language interpreters”
- Are more flexible than string operations in Excel, C, or other languages

Adding Flexibility with Regular Expressions

- REs match string patterns; they are not “language interpreters”
- Are more flexible than string operations in Excel, C, or other languages
- Key to their success: wildcards
  - Allows matching varying length patterns
  - Allows matching existence or non-existence of patterns
  - Matches combinations very efficiently, including nested search

Adding Flexibility with Regular Expressions

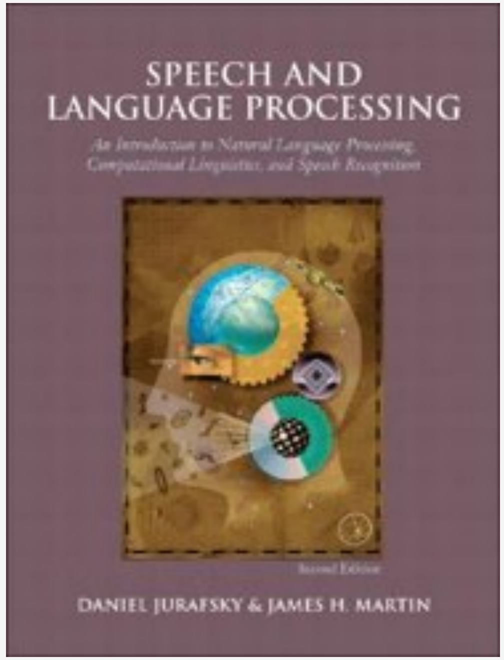
- REs match string patterns; they are not "language interpreters"
- Are more flexible than string operations in Excel, C, or other languages
- Key to their success: wildcards
  - Allows matching varying length patterns
  - Allows matching existence or non-existence of patterns
  - Matches combinations very efficiently, including nested search
- Typically, they start at the "left" end of a chunk of text and match the ***leftmost*****, longest** string defined by the RE

How Widespread is the Use of Regular Expressions?

- First chapter in the book “Speech and Language Processing: An Introduction to Natural Language Processing, Linguistics, and Speech Recognition” (Jurafsky and Martin, 2010, ISBN 978-7-115-23892-4) is on Regular Expressions

- “One of the unsung successes in standardization in computer science has been the regular expression (RE), a language for specifying text search strings.”

- “the regular expression is an important theoretical tool [used] throughout computer science and linguistics.”



What to do with Features?

- Create new columns for each document (row)
  - Each column represents a measure for a keyword or concept (phrase)
  - Can include multiple representations for each concept (keyword flags *and* TF-IDF, for example)
- But…
  - Could be many (way too many) columns!

Reducing Keyword Features

- Remove "useless" words: stopwords, articles, etc.
  - Is "The" in "The Who" useless?
- Assess features one-at-a-time
  - Keep features with predictive power (via chi-square or other test)
- Reduce features through Principal Component Analysis (PCA), Singular Value Decomposition (SVD), or clustering
  - Determines which features "load" together (i.e. are correlated)
  - One approach: keep factors that explain enough variance in data
- Cosine transform
  - Monotonic (for angles between 0 deg to 180 deg)
  - Product of TF-IDF for document and TF-IDF for corpus
    - Normalize by TF-IDF for all documents (sqrt of sum of squares)

Text Mining Resources

Dean Abbott
Abbott Analytics, Inc.
email: dean@abbottanalytics.com
url: http://www.abbottanalytics.com
blog: http://abbottanalytics.blogspot.com
Twitter: @deanabb

Miner, Elder, Hill, Nisbet, Delen, and Fast Text Mining Book

Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications [Hardcover]

Gary Miner (Author), John Elder IV (Author), Thomas Hill (Author), Robert Nisbet (Author), Dursun Delen (Author), Andrew Fast (Author)

Be the first to review this item | Like (10)

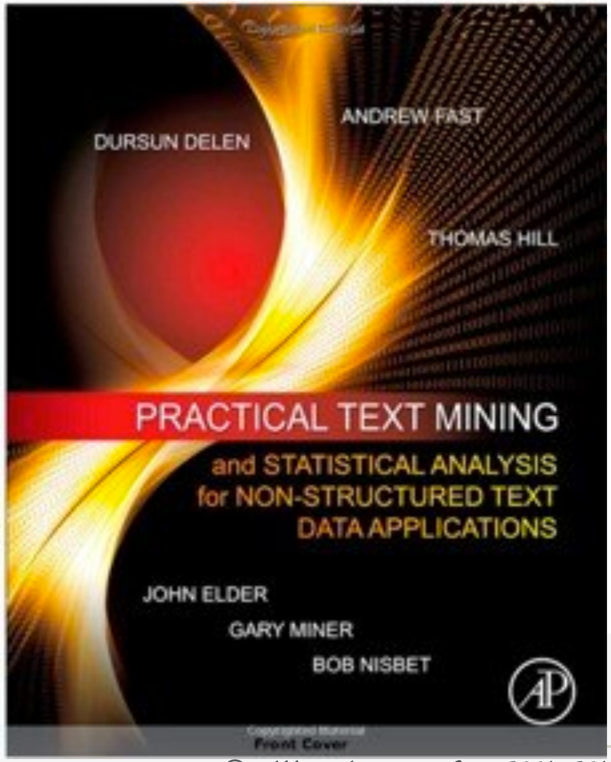
List Price: ~~$79.95~~
Price: $71.66 Prime
You Save: $8.29 (10%)
Special Offers Available

In Stock.
Ships from and sold by **Amazon.com**. Gift-wrap available.

Want it delivered Saturday, February 4? Order it in the next 22 hours and 45 minutes, and

2 new from $67.66   1 used from $68.03

Formats	Amazon Price	New from	Used from
Kindle Edition	$43.97	--	--
Hardcover	$71.66 Prime	$67.66	$68.03



Wednesday, July 10, 13

abbott analytics

Weiss, Indurkhya, Zhang, Damerau Text Mining Book

Text Mining: Predictive Methods for Analyzing Unstructured Information [Paperback]

Sholom M. Weiss (Author), Nitin Indurkhya (Author), Tong Zhang (Author), Fred Damerau (Author)

(5 customer reviews) | Liked (2)

List Price: ~~$109.00~~

Price: $81.04 Prime

You Save: $27.96 (26%)

Special Offers Available

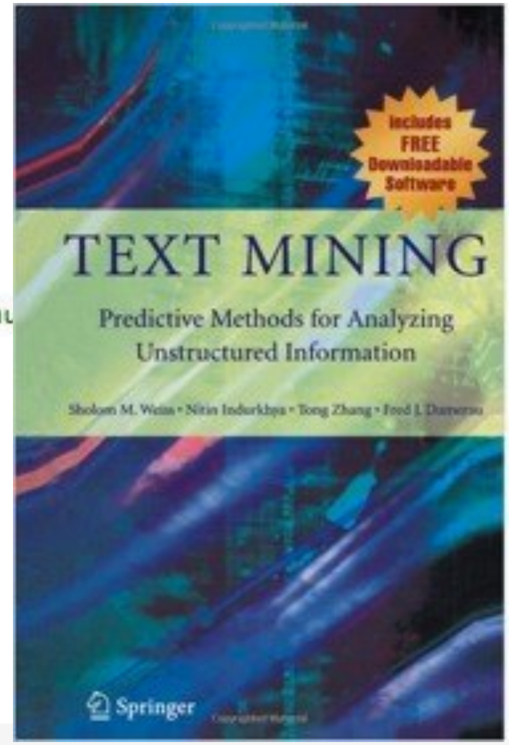
In Stock.

Ships from and sold by **Amazon.com**. Gift-wrap available.

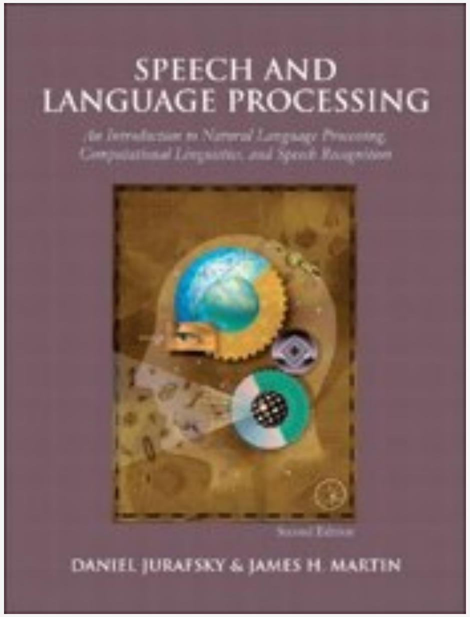
Want it delivered Saturday, February 4? Order it in the next 17 hours and 44 minu at check

12 new from $81.04 8 used from $109.38

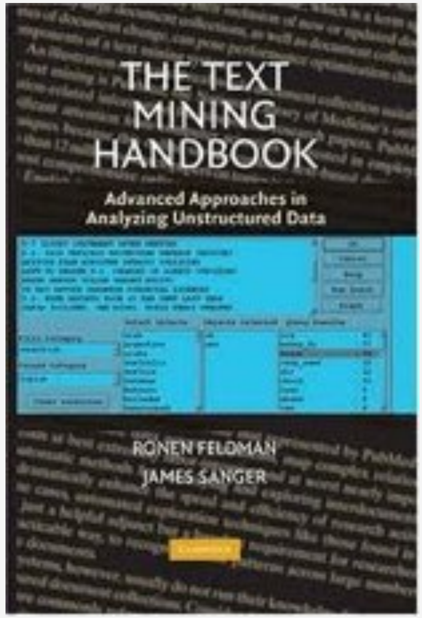
Formats	Amazon Price	New from	Used from
Kindle Edition	$66.88	--	--
Hardcover	$74.31 Prime	$60.00	$64.99
Paperback	$81.04 Prime	$81.04	$109.38



Text Mining Books on the Computational Linguistics Spectrum



"Speech and Language Processing: An Introduction to Natural Language Processing, Linguistics, and Speech Recognition" (Jurafsky and Martin, 2010, ISBN 978-7-115-23892-4)



"The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data", Ronen Feldman and James Sanger, ISBN-13: 978-0521836579

Some Interesting URLs

- KNIME software: http://knime.org/downloads/overview
- CST's POS Tagger (Brill): http://cst.dk/online/pos_tagger/uk/
- CRISP-DM: ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/
- Ngram software: http://homepages.inf.ed.ac.uk/lzhang10/ngram.html
- Statistical Analysis of Corpus Data with R: http://cogsci.uni-osnabrueck.de/~severt/SIGIL/sigil_R/

Regular Expression References

- Good references:
  - http://www.regular-expressions.info/tutorial.html
  - Regular Expressions Cookbook (O'Reilly) ISBN: 978-0596520687
  - Mastering Regular Expressions (O'Reilly) ISBM: 978-0596528126
  - Stanford Free Lectures
    - http://www.youtube.com/watch?v=hwDhO1GLb_4&feature=relmfu
    - Full course description: https://class.coursera.org/nlp/auth/welcome