Data Mining for Business Analytics

Lecture 9: Representing and Mining Text

Stern School of Business
New York University
Spring 2014

Dealing with Text

- Data are represented in ways natural to problems from which they were derived
- Vast amount of text..
- If we want to apply the many data mining tools that we have at our disposal, we must
  - either engineer the data representation to match the tools (representation engineering), or
  - build new tools to match the data

Why Text is Difficult

- Text is “unstructured”
  - Linguistic structure is intended for human communication and not computers
- Word order matters sometimes
- Text can be dirty
  - People write ungrammatically, misspell words, abbreviate unpredictably, and punctuate randomly
  - Synonyms, homograms, abbreviations, etc.
- Context matters

Text Representation

- **Goal**: Take a set of documents –each of which is a relatively free-form sequence of words– and turn it into our familiar feature-vector form
- A collection of documents is called a *corpus*
- *A document is composed of individual tokens or terms*
- *Each document is one instance*
  - *but we don't know in advance what the features will be*

“Bag of Words”

- Treat every document as just a collection of individual words
  - Ignore grammar, word order, sentence structure, and (usually) punctuation
  - Treat every word in a document as a potentially important keyword of the document
- What will be the feature’s value in a given document?
  - Each document is represented by a one (if the token is present in the document) or a zero (the token is not present in the document)
- Straightforward representation
- Inexpensive to generate
- Tends to work well for many tasks

Pre-processing of Text

The following steps should be performed:

- The case should be normalized
  - Every term is in lowercase
- Words should be stemmed
  - Suffixes are removed
  - E.g., noun plurals are transformed to singular forms
- Stop-words should be removed
  - A stop-word is a very common word in English (or whatever language is being parsed)
  - Typical words such as the words *the*, *and*, *of*, and *on* are removed

Term Frequency

- Use the word count (frequency) in the document instead of just a zero or one
  - Differentiates between how many times a word is used

Table 10-1. Three simple documents.

d1	jazz music has a swing rhythm
d2	swing is hard to explain
d3	swing rhythm is a natural rhythm

Table 10-2. Term count representation.

	a	explain	hard	has	is	jazz	music	natural	rhythm	swing	to
d1	1	0	0	1	0	1	1	0	1	1	0
d2	0	1	1	0	1	0	0	0	0	1	1
d3	1	0	0	0	1	0	0	1	2	1	0

Microsoft Corp and Skype Global today announced that they have entered into a definitive agreement under which Microsoft will acquire Skype, the leading Internet communications company, for $8.5 billion in cash from the investor group led by Silver Lake. The agreement has been approved by the boards of directors of both Microsoft and Skype.

Table 10-3. Terms after stemming, ordered by frequency

Term	Count	Term	Count	Term	Count	Term	Count
skype	3	microsoft	3	agreement	2	global	1
approv	1	announc	1	acquir	1	lead	1
definit	1	lake	1	communic	1	internet	1
board	1	led	1	director	1	corp	1
compani	1	investor	1	silver	1	billion	1

Normalized Term Frequency

- Documents of various lengths
- The raw term frequencies are normalized in some way,
  - such as by dividing each by the total number of words in the document
  - or the frequency of the specific term in the corpus
- Words of different frequencies--- the need for inverse document frequency
  - Words should not be *too common* or *too rare*
  - Both upper and lower limit on the number (or fraction) of documents in which a word may occur
  - Feature selection is often employed

TF-IDF

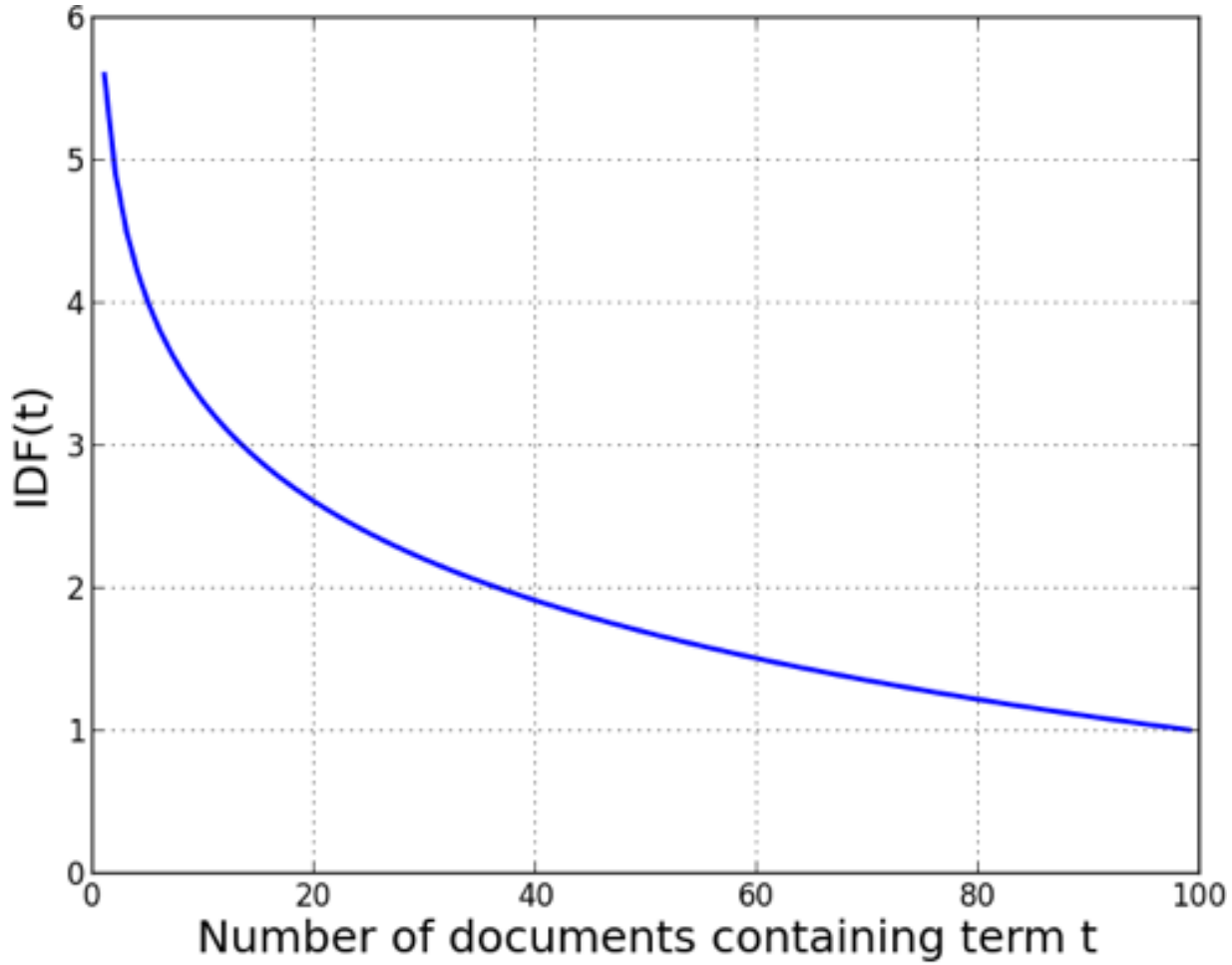
$$\mathrm{TFIDF}(t, d) = \mathrm{TF}(t, d) \times \mathrm{IDF}(t)$$

- Inverse Document Frequency (IDF) of a term

$$\mathrm{IDF}(t) = 1 + \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing } t}\right)$$

The sparseness of a term *t* is measured commonly by an equation called *inverse document frequency* (IDF), which may be thought of as the boost a term gets for being rare.

TFIDF



A graph of IDF(*t*) as a function of the number of documents in which *t* occurs, in a corpus of 100 documents.

Example: Jazz Musicians

- 16 prominent jazz musicians and excerpts of their biographies from Wikipedia

Charlie Parker

Charles "Charlie" Parker, Jr., was an American jazz saxophonist and composer. Miles Davis once said, "You can tell the history of jazz in four words: Louis Armstrong. Charlie Parker." Parker acquired the nickname "Yardbird" early in his career and the shortened form, "Bird," which continued to be used for the rest of his life, inspired the titles of a number of Parker compositions, […]

Duke Ellington

Edward Kennedy "Duke" Ellington was an American composer, pianist, and big-band leader. Ellington wrote over 1,000 compositions. In the opinion of Bob Blumenthal of *The Boston Globe*, "in the century since his birth, there has been no greater composer, American or otherwise, than Edward Kennedy Ellington." A major figure in the history of jazz, Ellington's music stretched into various other genres, including blues, gospel, film scores, popular, and classical.[…]

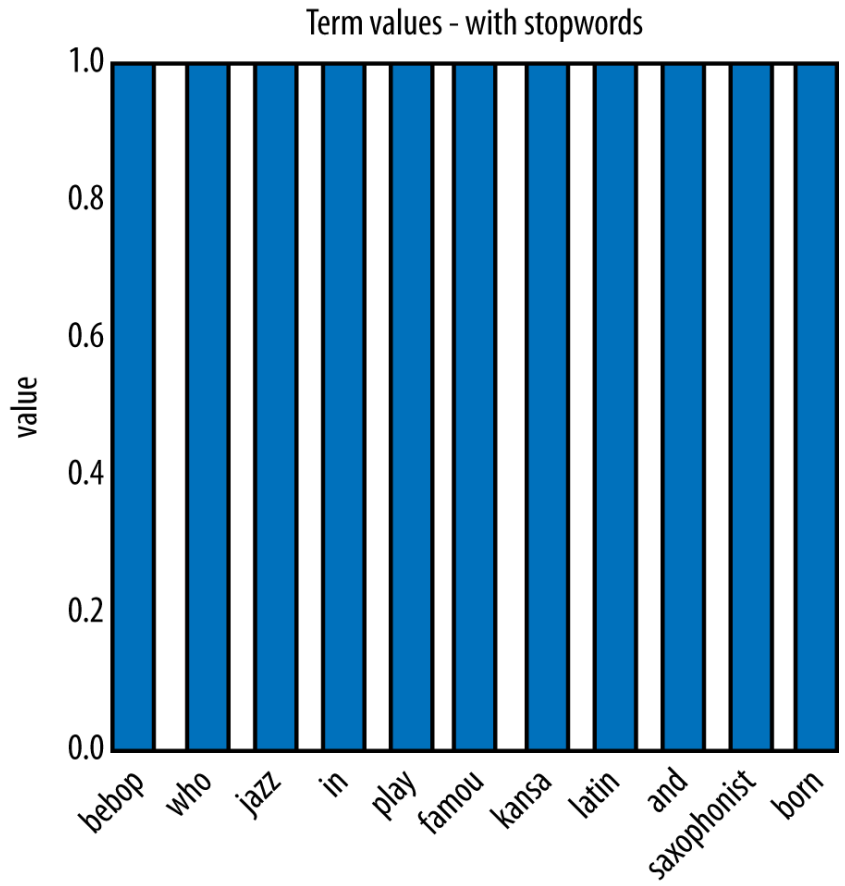
Miles Davis

Miles Dewey Davis III was an American jazz musician, trumpeter, bandleader, and composer. Widely considered one of the most influential musicians of the 20th century, Miles Davis was, with his musical groups, at the forefront of several major developments in jazz music, including bebop, cool jazz, hard bop, modal jazz, and jazz fusion.[…]

Example: Jazz Musicians

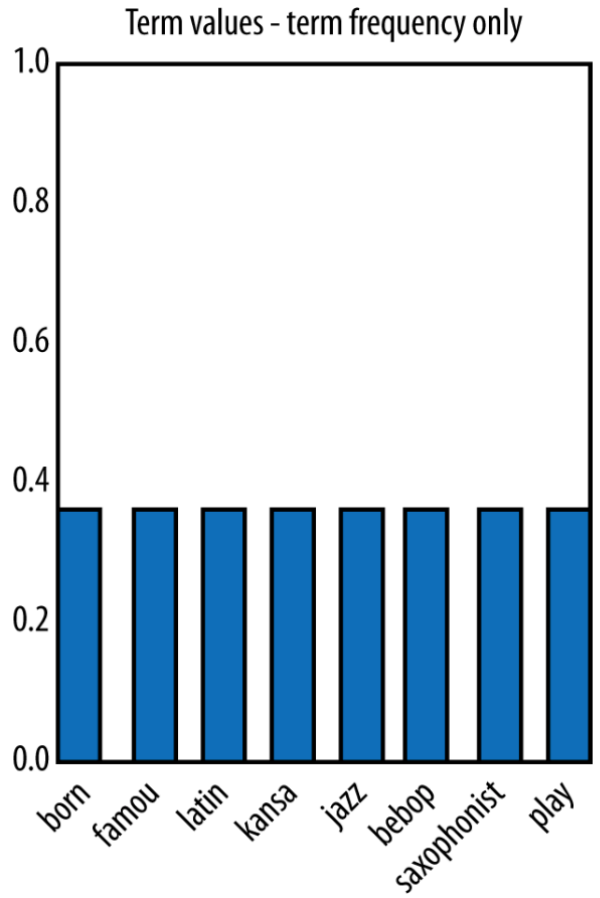
- Nearly 2,000 features after stemming and stop-word removal!
- Consider the sample phrase “Famous jazz saxophonist born in Kansas who played bebop and latin”
  - Our goal is to find the musicians that best matches above phrase
  - Idea:
    - a. develop a feature vector
    - b. compute the distance of above phase’s feature vector and each musician’s feature vector
    - c. pick the musician whose feature vector is most similar to the above phase’s

Example: Jazz Musicians

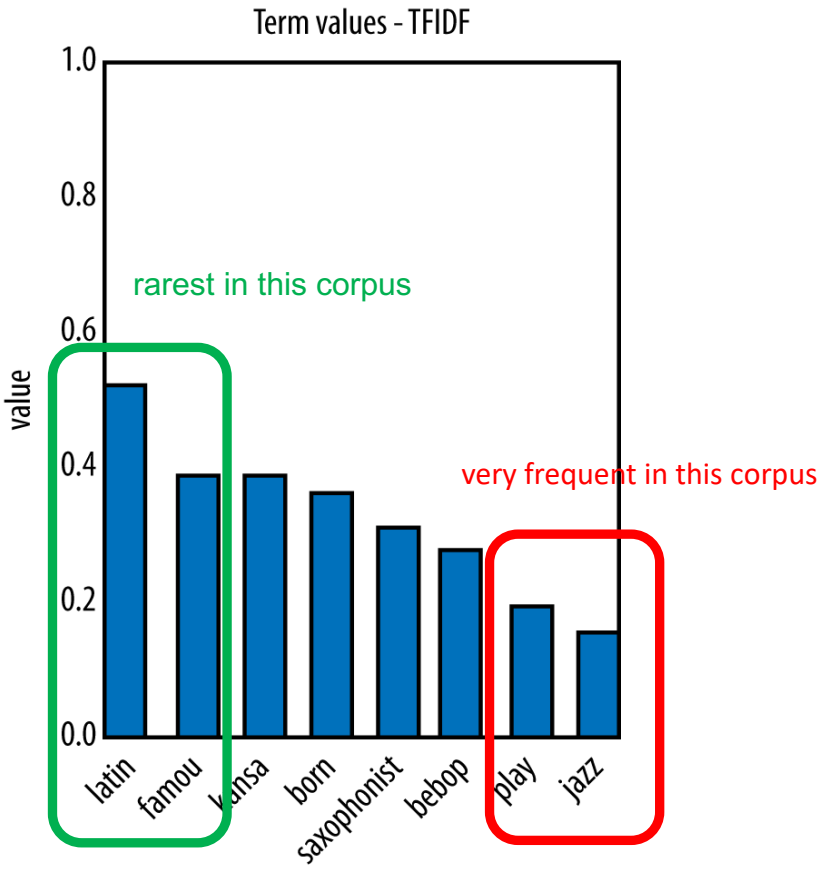


Representation of the query "Famous jazz saxophonist born in Kansas who played bebop and latin" after stemming.

Example: Jazz Musicians



Representation of the query "Famous jazz saxophonist born in Kansas who played bebop and latin" after stopword removal and term frequency normalization.



Final TFIDF representation of the query "Famous jazz saxophonist born in Kansas who played bebop and latin."

Example: Jazz Musicians

Similarity of each musician's text to the query 'Famous jazz saxophonist born in Kansas who played bebop and latin,' ordered by decreasing similarity.

Musician	Similarity	Musician	Similarity
Charlie Parker	0.135	Count Basie	0.119
Dizzie Gillespie	0.086	John Coltrane	0.079
Art Tatum	0.050	Miles Davis	0.050
Clark Terry	0.047	Sun Ra	0.030
Dave Brubeck	0.027	Nina Simone	0.026
Thelonius Monk	0.025	Fats Waller	0.020
Charles Mingus	0.019	Duke Ellington	0.017
Benny Goodman	0.016	Louis Armstrong	0.012

$$d_{cosine}(\mathbf{X}, \mathbf{Y}) = 1 - \frac{\mathbf{X} \cdot \mathbf{Y}}{\| \mathbf{X} \|_2 \cdot \| \mathbf{Y} \|_2}$$

where $||\cdot||_2$ again represents the L2 norm, or Euclidean length, of each feature vector

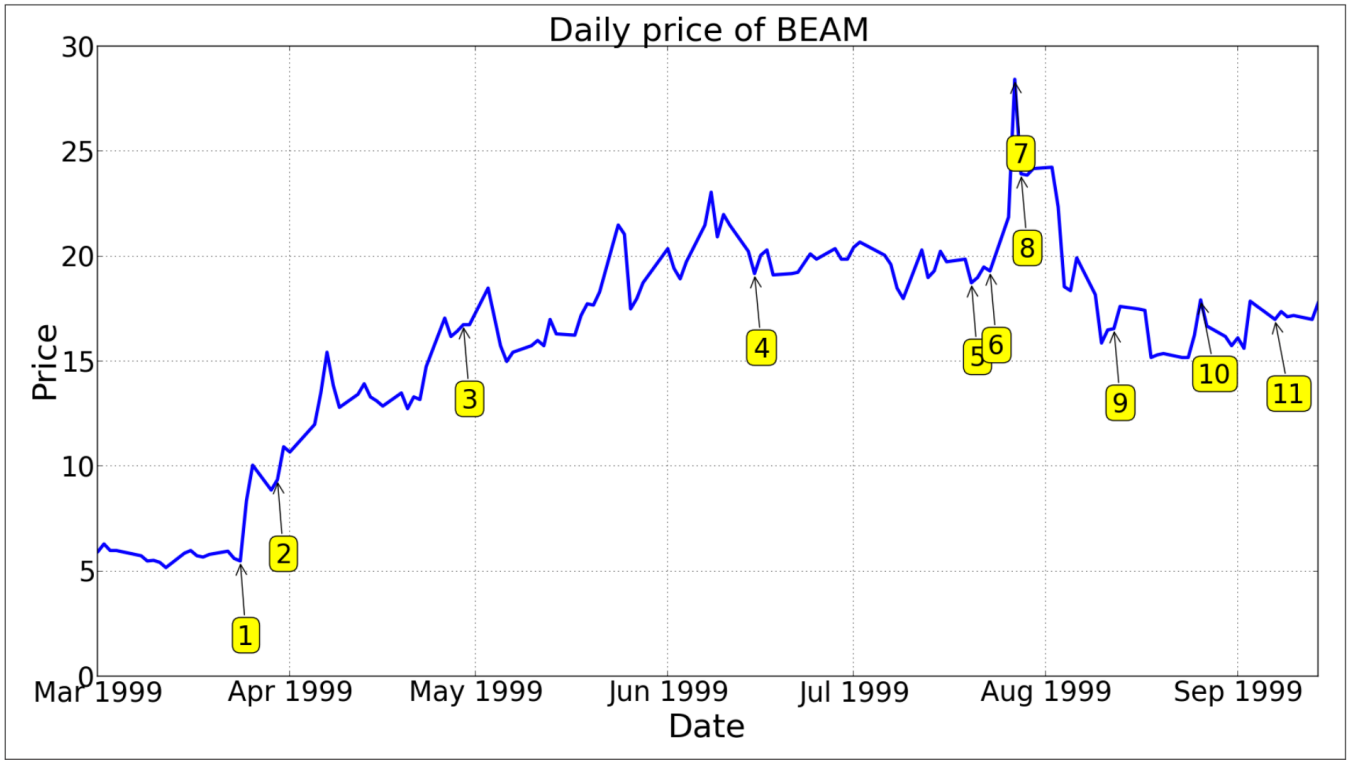
Beyond “Bag of Words”

- N-gram Sequences
- Topic Models

N-gram Sequences

- In some cases, **word order is important** and you want to preserve some information about it in the representation
- A next step up in complexity is to include sequences of adjacent words as terms
- Adjacent pairs are commonly called bi-grams
- Example: “The quick brown fox jumps”
  - It would be transformed into {quick, brown, fox, jumps, quick_brown, brown_fox, fox_jumps}
- N-grams they greatly increase the size of the feature set

Example: Mining News Stories to Predict Stock Price Movement



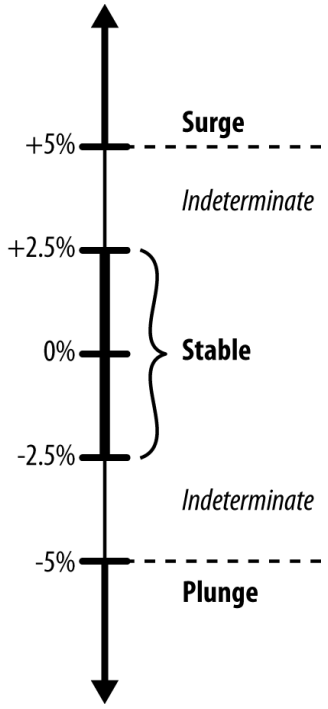


Figure 10-8. Graph of stock price of Summit Technologies, Inc., (NASDAQ:BEAM) annotated with news story summaries.

1 Summit Tech announces revenues for the three months ended Dec 31, 1998 were $22.4 million, an increase of 13%.

2 Summit Tech and Autonomous Technologies Corporation announce that the Joint Proxy/Prospectus for Summit's acquisition of Autonomous has been declared effective by the SEC.

3 Summit Tech said that its procedure volume reached new levels in the first quarter and that it had concluded its acquisition of Autonomous Technologies Corporation.

4 Announcement of annual shareholders meeting.

5 Summit Tech announces it has filed a registration statement with the SEC to sell 4,000,000 shares of its common stock.

6 A US FDA panel backs the use of a Summit Tech laser in LASIK procedures to correct nearsightedness with or without astigmatism.

7 Summit up 1-1/8 at 27-3/8.

A story

1999-03-30 14:45:00
WALTHAM, Mass.--(BUSINESS WIRE)--March 30, 1999--Summit Technology, Inc. (NASDAQ:BEAM) and Autonomous Technologies Corporation (NASDAQ:ATCI) announced today that the Joint Proxy/Prospectus for Summit's acquisition of Autonomous has been declared effective by the Securities and Exchange Commission. Copies of the document have been mailed to stockholders of both companies. "We are pleased that these proxy materials have been declared effective and look forward to the shareholder meetings scheduled for April 29," said Robert Palmisano, Summit's Chief Executive Officer.

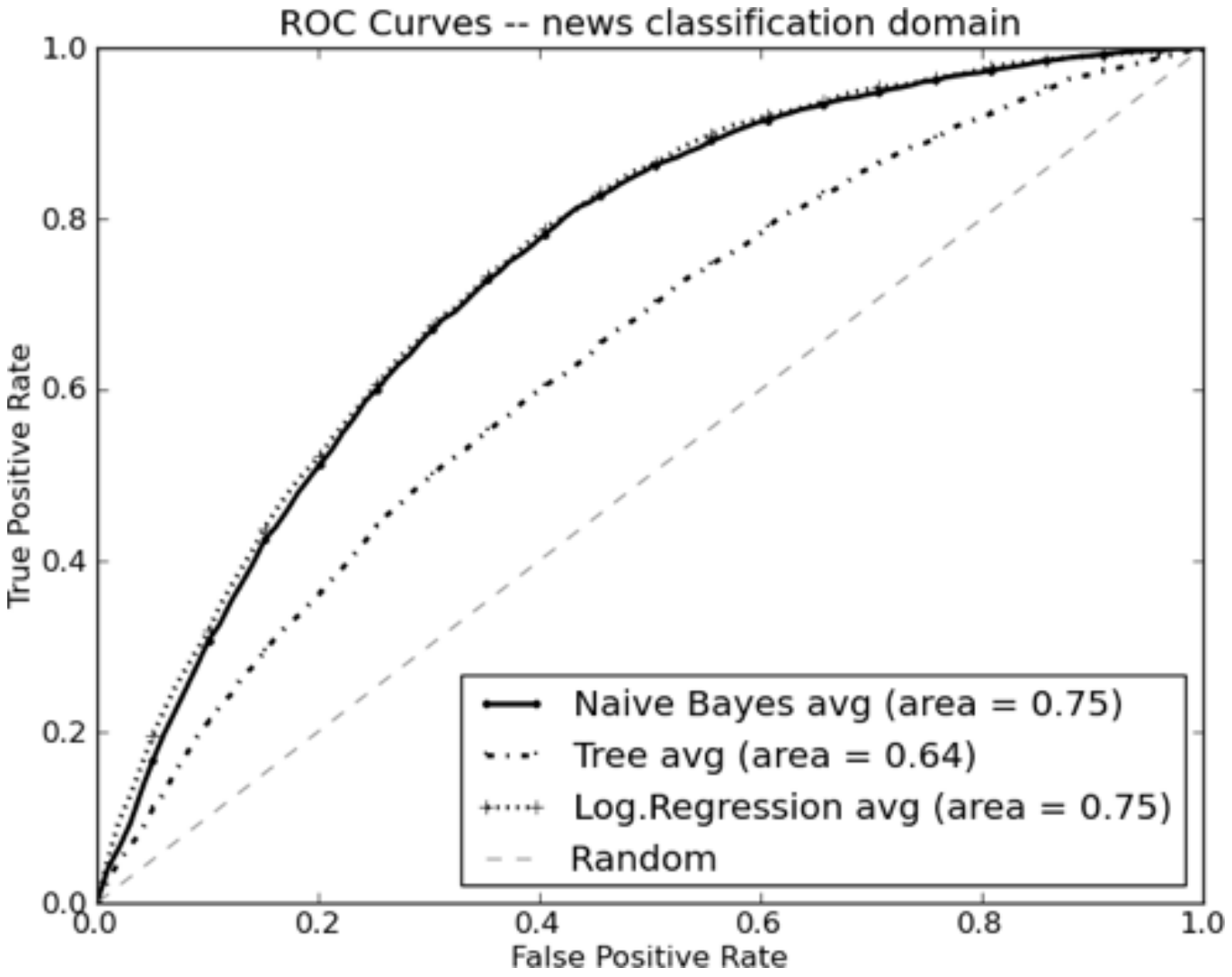
Features and Labels

"Bag of Words" was applied to reduce each story to a TFIDF representation. In particular, each word was case-normalized and stemmed, and stopwords were removed. Finally, we created n-grams up to two, such that every individual term and pair of adjacent terms were used to represent each story.

Subject to this preparation, each story is tagged with a label (**change** or **no change**) based on the associated stock(s) price movement

This results in about 16,000 usable tagged stories. For reference, the breakdown of stories was about 75% no change, 13% surge, and 12% plunge. The surge and plunge stories were merged to form **change**, so 25% of the stories were followed by a significant price change to the stocks involved, and 75% were not.

Mining News Stories to Predict Stock Price Movement



Mining News Stories to Predict Stock Price Movement

terms with high information gain

alert(s,ed), architecture, auction(s,ed,ing,eers), average(s,d), award(s,ed), bond(s), brokerage, climb(ed,s,ing), close(d,s), comment(ator,ed,ing,s), commerce(s), corporate, crack(s,ed,ing), cumulative, deal(s), dealing(s), deflect(ed,ing), delays, depart(s,ed), department(s), design(ers,ing), economy, econtent, edesign, eoperate, esource, event(s), exchange(s), extens(ion,ive), facilit(y,ies), gain(ed,s,ing), higher, hit(s), imbalance(s), index, issue(s,d), late(ly), law(s,ful), lead(s,ing), legal(ity,ly), lose, majority, merg(ing,ed,es), move(s,d), online, outperform(s,ance,ed), partner(s), payments, percent, pharmaceutical(s), price(d), primary, recover(ed,s), redirect(ed,ion), stakeholder(s), stock(s), violat(ing,ion,ors)

Thanks!

Questions?