Chapter 10. Representing and Mining Text

Fundamental concepts: preparation and representation of text data for mining

Exemplary techniques: Bag of words, TFIDF scores, n-grams, stemming, named entity extraction, topic models

Text data are extremely common nowadays, largely due to Internet which has become a ubiquitous channel of communication.

One important challenge is to represent each text data point (i.e. a document) as a numerical vector such that the data mining tools are become directly applicable.

The basic idea is also helpful in dealing with other types of non-numerical data.

Further Reading:

Provost and Fawcett (2013): Chapter 10

Why Text is Important? – It is everywhere!

Medical records, consumer complaint logs, product inquiries, and repair records are all in the form of text, for communication between people.

Internet is the new media: most of it still in the form of text – personal web pages, Twitter feeds, email, Facebook status updates, product descriptions, blog postings etc

Google and Bing are based on massive amounts of text-oriented data science.

Exploiting this vast amount of data requires converting text to the format which is meaningful to computers, i.e. a vector consisting of numerical attributes.

Why Text is Difficult?

- *Unstructured*: no uniform structure across different texts. Each text has its own free-form sequence of words, length, number of paragraphs, symbols, tables and figures.

- *Dirty*: some documents may be written ungrammatically, with misspell words, or words together, abbreviate unpredictably, and punctuate randomly.

- *Ambiguity*: different words share the same meaning, or the same words mean differently in different contexts.

  Texts are intended for human consumption, *context* is important. The same words or statements may mean different things in different context. It can be difficult to evaluate any particular word or phrase here without taking into account the entire context.

> “The first part of this movie is far better than the second. The acting is poor and it gets out-of-control by the end, with the violence overdone and an incredible ending, but it’s still fun to watch.”

In this movie review excerpt, it is not clear if the overall sentiment is positive or negative, or if the word *incredible* is used positively or negatively?

Text must undergo serious preprocessing before it can be used for data mining

Document: one piece of text (regardless its length or contents)

Corpus: a collection of documents concerned.

Term or *Token*: a word, a phrase, or several connected words.

Bag of Words – a basic tools for text data representation

Treat every document as just a collection of individual words, ignoring grammar, word order, sentence structure, and punctuation.

This is a very simple approach, inexpensive to generate, and tends to work well for many tasks.

However some preprocessing is necessary:

- *Case-normalization*: make every word in lower-case

  `iPhone`, `iphone` and `IPHONE` are treated as one word

- *Stemming*: remove suffixes

  verbs like `announces`, `announced` and `announcing` are all reduced to `announc` change noun plurals to singular, e.g. `directors` is recorded as `director`

- *Stopwords*: such as `and, a, an, of, on, at`. Those words are very common and tend to occur in all documents.

  For some applications (but not the information retrieval!), one may also exclude words which occur too rare, say, under 3% of the documents in the corpus

  On the other hand, the words occurring in most documents are not useful either for, e.g. classification and clustering, should be removed for those applications.

After the above preprocessing, every remaining word is a possible feature. There are several ways to present the value for each feature.

1. *Binary*: each word is a token with value 1 if token occurs in the document, and 0 otherwise.

Each document is represented by the set of words contained in it, represented by a long vector consisting of 1 and 0. The length of the vector is the total number of words contained in all the documents in the corpus.

2. *Term Frequency*: using the word count (frequency) in the document instead of just 1 or 0.

   An obvious drawback: longer documents tend to produce larger TF scores.

   The TF may be divided by the total number of words in the document

3. *TFIDF*: The TFIDF value of a term t in a given document d is defined as

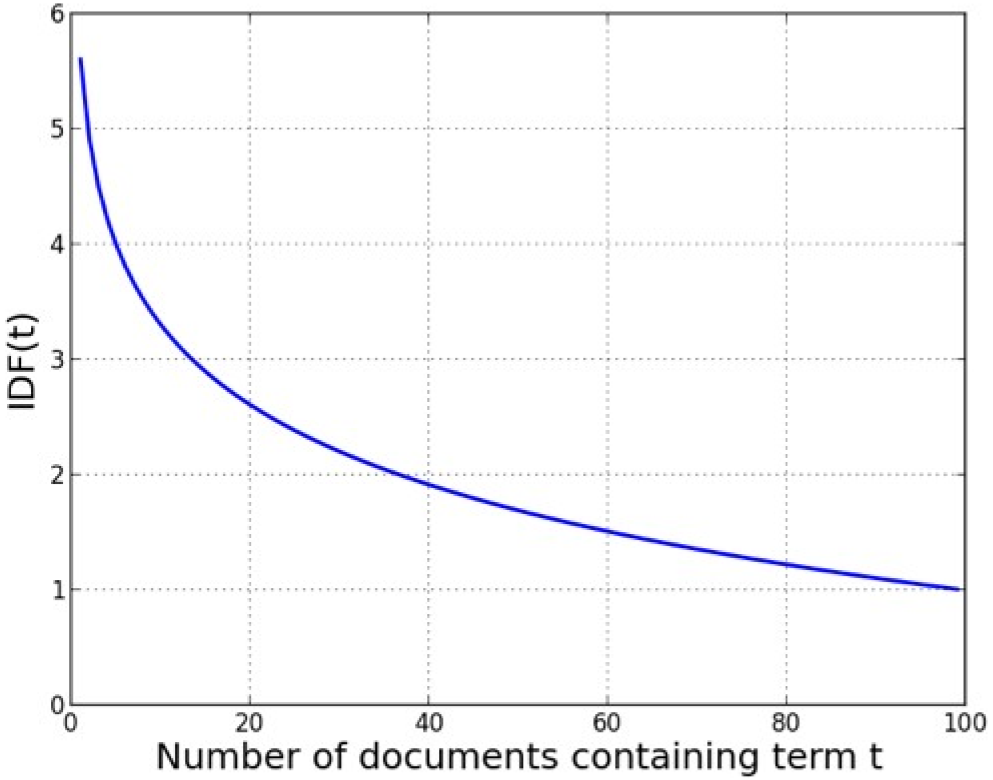
$$\text{TFIDF} = \text{TF (term frequency)} \times \text{IDF (inverse document frequency)}$$

$$\text{TF(t, d)} = \text{No. of times of word t occurring in document d}$$

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

Term counts within the documents form the TF values for each term, and the document counts across the corpus form the IDF values.

IDF boosts rare terms.



IDF of a term t within a corpus of 100 documents

Since the number of features is often excessively large. Feature selection is often necessary, which can be carried out by imposing minimum and maximum thresholds of term counts, and/or using a measure such as information gain to rank the terms by importance so that low-gain terms can be culled.

The bag-of-words text representation approach treats words in a document as independent terms of the document by assigning values to each term. TFIDF is a very commonly used, based on frequency and rarity. But it could be binary, term frequency, with normalization or without.

Experiment with different representations to see which produces the best results.

Example: Jazz Musicians

Data: Excerpts of the biographies from Wikipedia for 16 jazz musicians.

- 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, $\cdots$

- Duke Ellington

  Edward Kennedy ‘‘Duke’’ Ellington was an American composer, pianist, and bigband 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, bandlea 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. ···

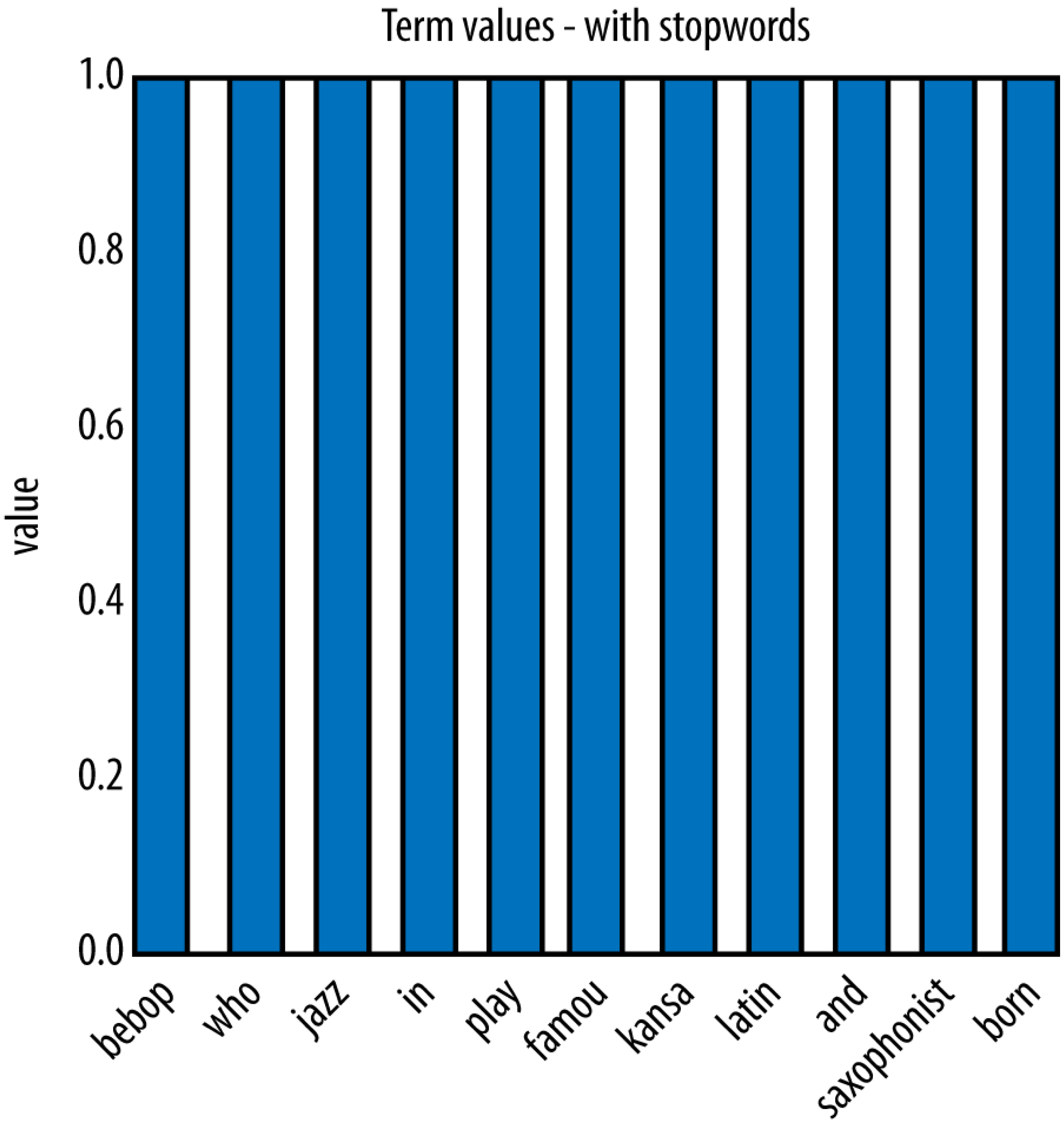
  ......

For this small corpus with $n = 16$, its vocabulary are large with $p \approx 2000$ after stemming and stopword removal. Applying the Bags of Words technical above with TFIDF scores, we translate each biography into a p-vector.

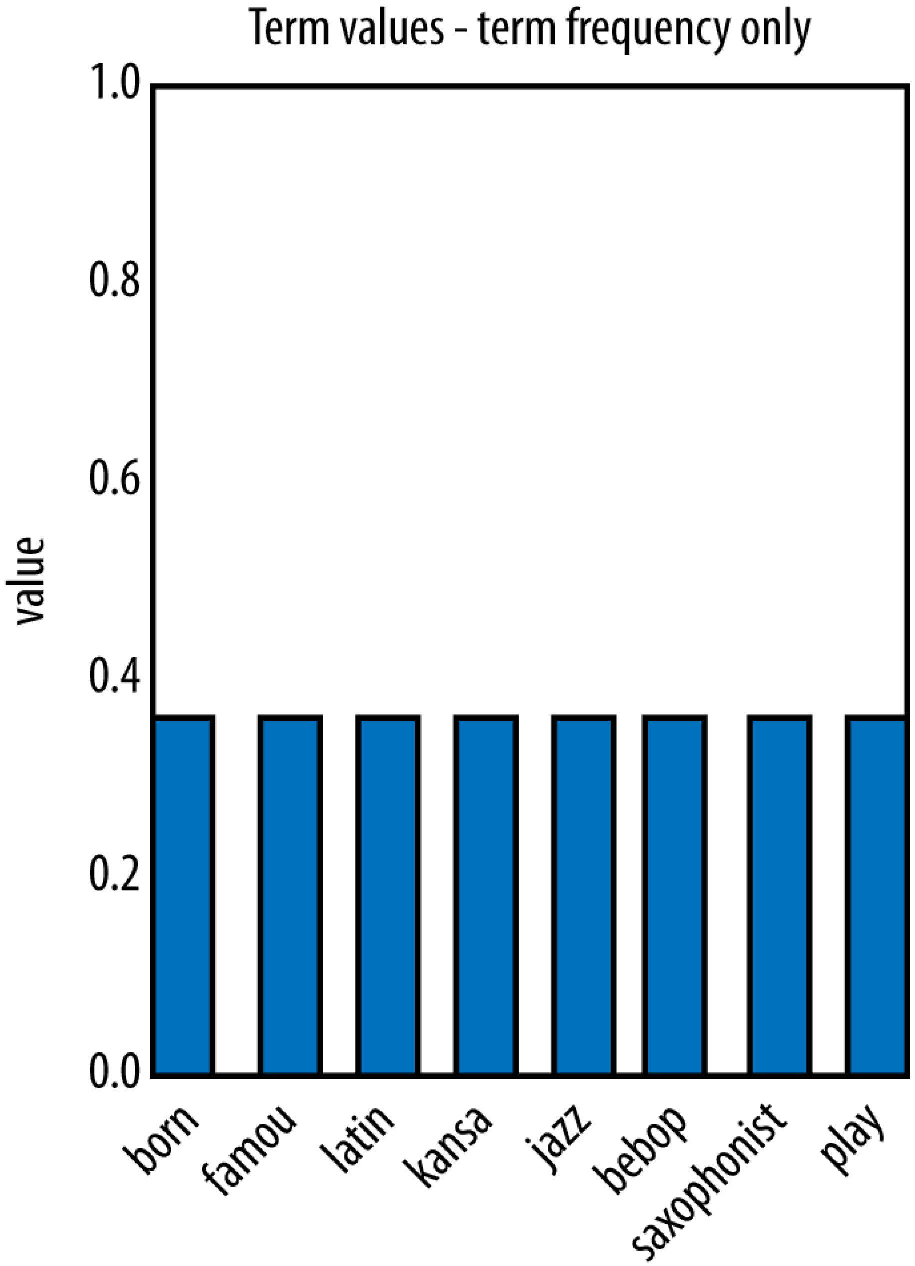
To illustrate its usefulness, suppose a search engine received a query:

```
Famous jazz saxophonist born in Kansas who played bebop and latin.
```

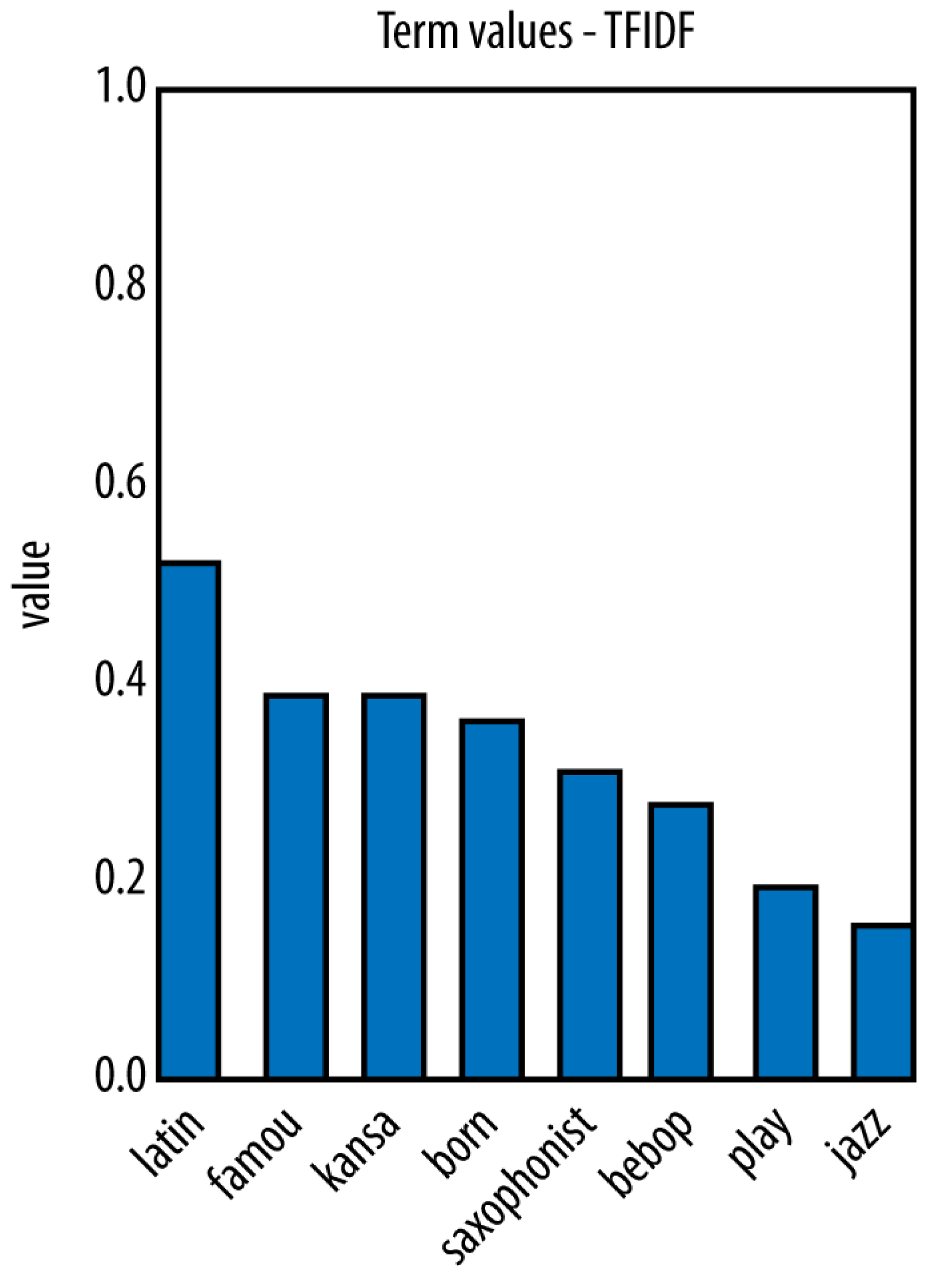
It treats the query exactly as a document to process it using the Bag-of-Words techniques.



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



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.’

The IDF scores were calculated based on 16 biographies in the corpus.

Using the correlation-based measure,

$$\rho(\mathbf{x},\mathbf{y}) = \sum_i x_i y_i \Big/ \sqrt{\sum_i x_i^2 \sum_j y_j^2}$$

the similarity between the query and each of the 16 Jazz musicians' biography was calculated.

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

Charlie Parker is the closest match. He in fact is a saxophonist born in Kansas and who played the bebop style of jazz. He sometimes combined other genres, including Latin, a fact that is mentioned in his biography.

Beyond Bag of Words

The basic bag of words approach is relatively simple, requires no linguistic analysis. It performs surprisingly well on a variety of tasks.

But some further improvements are required for many applications.

k-gram Sequences

To take into account of the order of words, take k adjacent words as a term.

For example, the sentence '`The quick brown fox jumps`' will generate terms {`quick, brown, fox, jump, quick-brown, brown-fox, fox-jump`} in a bag-of-words with 2-gram sequences

The advantage of k-gram is obvious when particular phrases are significant but their component words may not be. For example, the tri-gram `exceed-analyst-expectation` is more meaningful than the 3 individual words.

However it increases the number of attributes substantially, demanding more storage and computing/searching power.

Named Entity Extraction

Many text-processing toolkits include a named entity extractor, to extract phrases annotated with terms like *person* or *organization*.

This knowledge has to be learned from a large corpus, or coded by hand.

Some extractors may have particular areas of expertise, such as industry, government, or popular culture.

Example: Mining News Stories to Predict Stock Price Movement

Background: Companies make and announce decisions of mergers, new products, earnings projections, and so forth. Investors read these news stories, possibly change their beliefs about the prospects of the companies involved, and trade stock accordingly. Then stock prices change.

Ideally we would like predict in advance and with precision the change in a company's stock price based on the stream of news. In reality there are many complex factors involved in stock price changes, many of which are not conveyed in news stories.

Goal: to mine the news to recommend interesting news stories.

A piece of news is interesting if it will likely result in a significant change in a stock's price.

Assumptions:

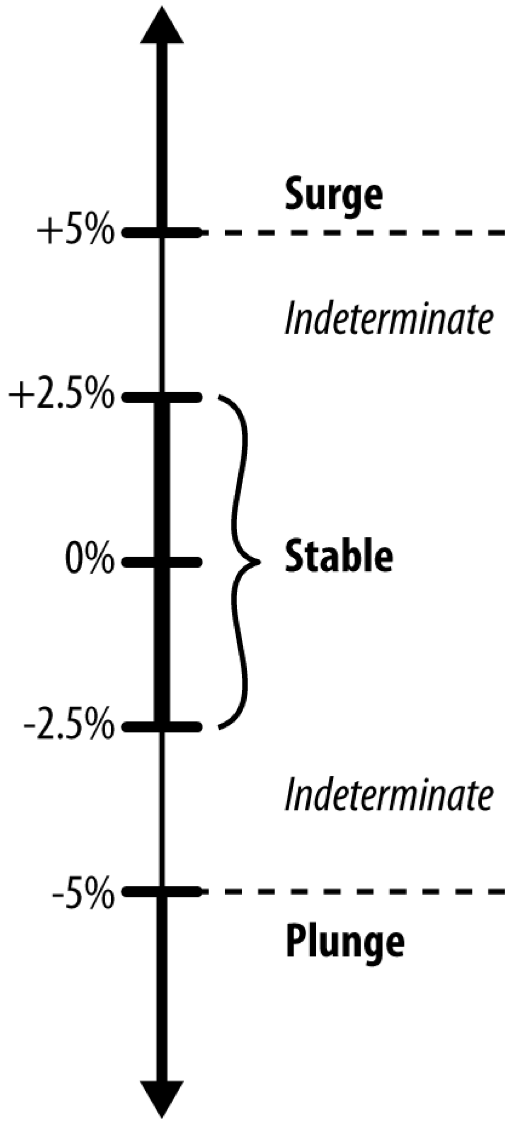
1. Only the changes in price on the same day are considered, as too difficult to predict the impact in the future.

2. Simplify stock price movements into two categories: `change` and `no change`, as predicting the exact changes are too difficult. (Here the direction of change is ignored.)

3. Only count for relatively large changes, ignoring the subtlety of small fluctuations.

4. Only news stories mentioning a specific stock will affect that stock's price.

This is inaccurate of course, and is a simplification to make the analysis easier.



BPercentage change in price, and corresponding label.

No change = stable

change = {surge, plunge}

Data. Two separate time series: the stream of news stories (text documents), and a corresponding stream of daily stock prices in 1999, for the stocks listed on the New York Stock Exchange and NASDAQ.

About 36,000 news stories.

For example, to see what news stories are available about Apple Computer, Inc., see the corresponding Yahoo!Finance page. Yahoo! aggregates news stories from a variety of sources such as Reuters, PR Web, and Forbes.

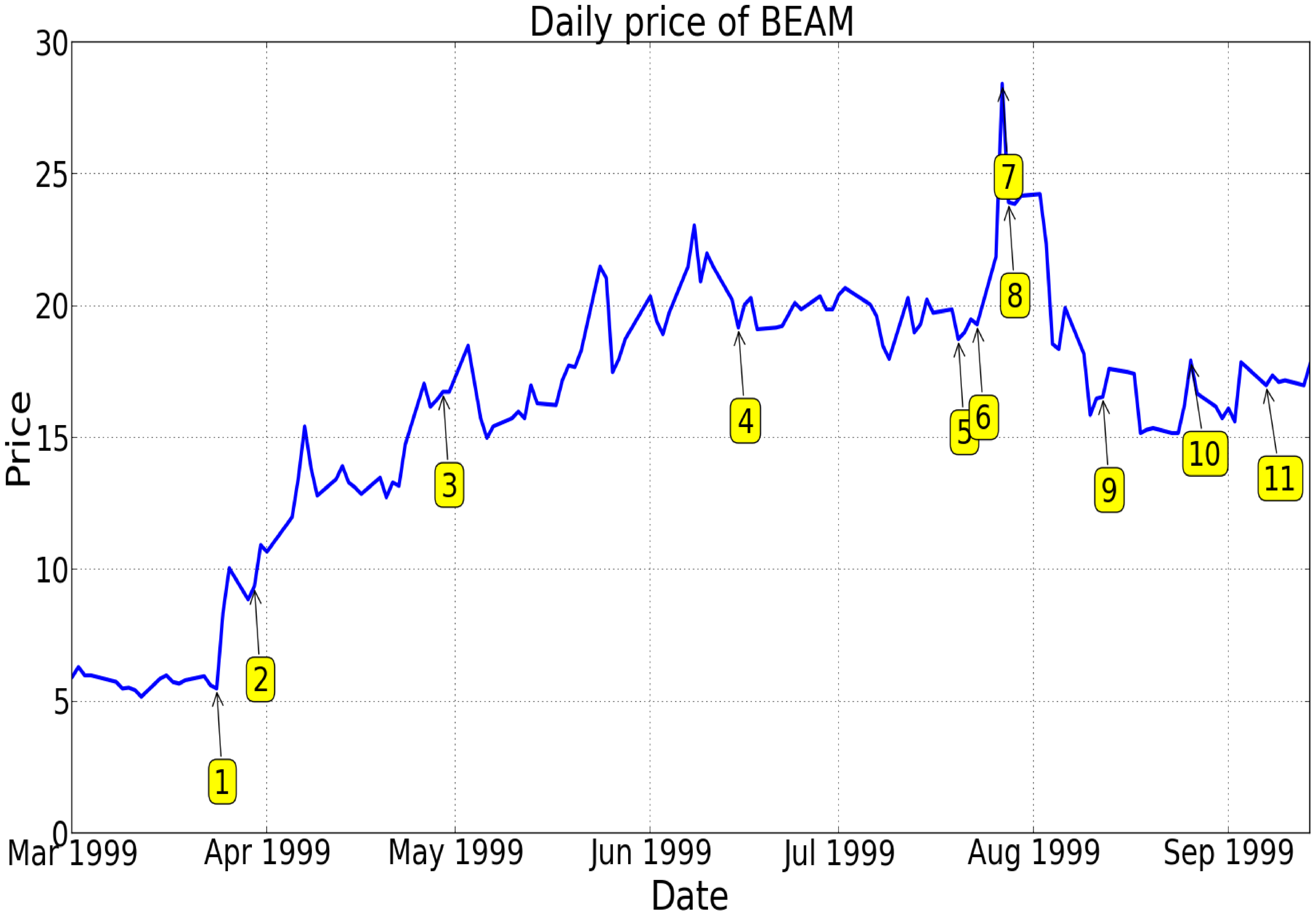
The new stories contain many miscellaneous materials: date and time, the news source, stock symbols, links to other sites, as well as background material not strictly germane to the news.

Here is one news 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.
```

Each such story is tagged with the stock mentioned.

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


Daily price of BEAM
Price
Date
0
5
10
15
20
25
30
Mar 1999
Apr 1999
May 1999
Jun 1999
Jul 1999
Aug 1999
Sep 1999
1
2
3
4
5
6
7
8
9
10
11

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.

8 Summit Tech said today that its revenues for the three months ended June 30, 1999 increased 14% ...

9 Summit Tech announces the public offering of 3,500,000 shares of its common stock priced at $16/share.

10 Summit announces an agreement with Sterling Vision, Inc. for the purchase of up to six of Summit‘s state of the art, Apex Plus Laser Systems.

11 Preferred Capital Markets, Inc. initiates coverage of Summit Technology Inc. with a Strong Buy rating and a 12-16 month price target of $22.50.

News is Messy

- News comprises a wide variety of stories, including earnings announcements, analysts' assessments, market commentary, SEC filings, financial balance sheets, and so on. Companies are mentioned for many different reasons and a single document may actually comprise multiple unrelated news blurbs of the day.

- Stories come in different formats, some with tabular data, some in multiparagraphs. Much of the meaning is imparted by context.

- Stock tagging is not perfect, tends to be overly permissive, such that stories are included in the news feed of stocks that were not actually referenced in the story.

Data Preprocessing

Each stock has an opening (at 9:30am EST) and closing (at 4pm EST) price for each day.

To classify each day into '`change`', '`no change`' or not classified, let

$$\text{PercenC} = 100 \times \frac{\text{(Price at 4pm) - (Price at 10am)}}{\text{(Price at 10am)}}$$

If $|\text{PercenC}| \geq 5$, '`change`'

If $|\text{PercenC}| < 2.5$, '`no change`'

Why use prices at 10am?

News also occurs off trading hours, and fluctuations near the opening hours can be erratic. Therefore in addition, define the change between days

$$\text{PercenC} = 100 \times \frac{\text{(Price at 10am) - (Price at 4pm yesterday)}}{\text{(Price at 4am yesterday)}}$$

The news stories require more care!

Stories without timestamps are discarded.

Stories mentioning two stocks or more are discarded.

Each story is aligned with the correct stock at correct trading day/window.

Stories corresponding to unclassified trading days/windows are discarded.

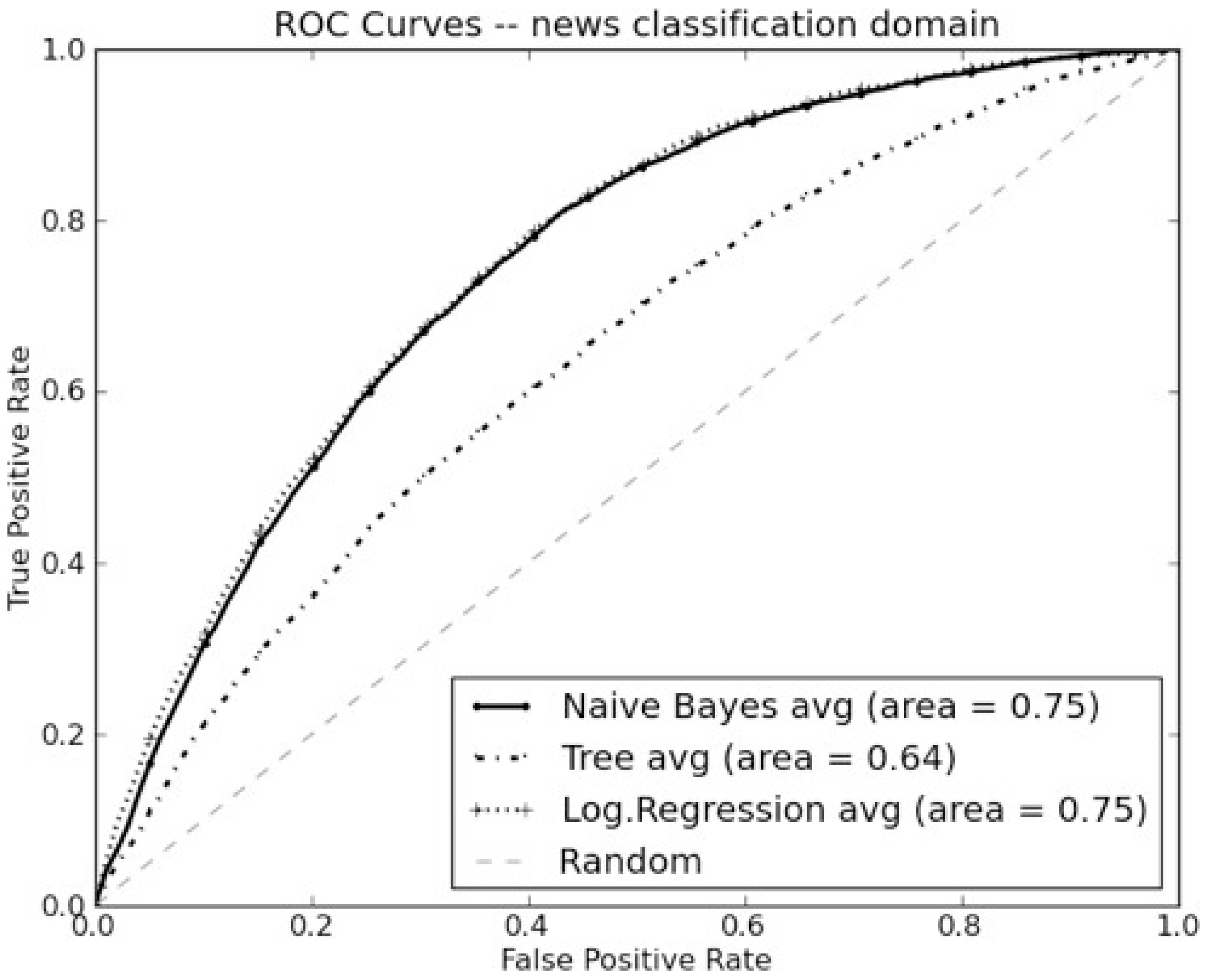
Remaining stories are gained a label `‘change’` or `‘no change’`.

Features are extracted from each of those stories using *Bag-of-Words* with 2-grams, after case-normalization, stemming, and stopword-removing.

Finally, there are about 16,000 usable tagged stories, with about 75% with label `no change` and 25% with label `change` (actually 13% for `surge` and 12% for `plunge`).

Results

As the goal is modest, i.e. to identify the news stories which are associated to substantial stock price changes, no cost and benefit analysis here and no expected value calculation either.



Average from ten-fold cross-validation, using change as the positive class and no change as the negative class.

1. Predictive signal of news stories is indicated by the ‘bowing out’ of the curves above the diagonal line (random classifiers). The AUCs are greater 0.5 substantially.

2. Logistic regression and Naive Bayes perform similarly, whereas the classification tree (Tree) is considerably worse.

3. No classifiers (with any threshold values) are close to the perfection point (0, 1).

Below are the words (or stems) in the 'Bag of Words' which are most informative (i.e. with smallest conditional entropies):

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)

Many are suggestive of of good or bad news for a company or its stock price.

Some of them (econtent, edesign, eoperate) are also suggestive of the ‘Dotcom Boom’ of the late 1990s

This is perhaps the most complex example encountered so far. However it still represents an excessively simplistic approach to a real and complex project.

- No particular effort on the extraction of the names of companies and people involved. Furthermore it is not clear from individual words who are the subjects and objects of the events.
- Important modifiers like not, despite, and expect may not be adjacent to the phrases they modify.
- Markets react to news quickly. Hourly or instantaneous price changes should be used in order to trade on the information.
- Consider 3-class classification: `no change, surge, plunge`
- Time series nature of the data is almost completely ignored.

In addition to Chapter 10 of the textbook by Provost and Fawcett, here are a few references on this ‘News-Stock Price’ example:

Mittermayer, M., and Knolmayer, G. (2006). Text mining systems for market response to news: A survey. Working Paper No.184, Institute of Information Systems, University of Bern.

Zhang, J., Haerdle, W.K., Cheng, C.Y. and Bommes, E. (2015). Distillation of news flow into analysis of stock reactions.
`http://edoc.hu-berlin.de/series/sfb-649-papers/2015-5/PDF/5.pdf`