MTAT.03.319

Business Data Analytics

Lecture 7

The slides are available under creative common license. The original owner of these slides is the University of Tartu.
Understanding and Mining Text

What can you infer from text?
It appears tweets by Kylie cause Snapchat to lose $1.3 billion today. **TMZ reports** that the social media platform's shares are down 7.2 percent hours after Jenner posted her tweets.
Impact of a text (Tweet)!

- It appears tweets by Kylie cause Snapchat to lose $1.3 billion today. TMZ reports that the social media platform's shares are down 7.2 percent hours after Jenner posted her tweets.

Snapchat CEO's Net Worth Tanks $150 Million After Rihanna's Rant

After Rihanna slammed Snapchat for running an ad that made light of domestic violence, the CEO's net worth fell nearly $150 million over 2 days.
Twitter And Box office (Movies)

A View from Emerging Technology from the arXiv

Twitter Used to Predict Box Office Revenues

The rate of tweeting about a movie accurately predicts its box office revenues, say researchers. But only after a film has been released.

April 1, 2010

Sources:
Twitter And Financial Market

• Two Theories:
  • An “informational” theory:
  • A “sentimental” theory:

Sources:
Twitter And Financial Market

• Two Theories:
  • An “informational” theory:
    An “informational” theory: Information published on Twitter is new, in the sense that it has not yet been incorporated into stock prices. That information can thus be expected to influence the value one could rationally expect for the future cash flows of an asset.
  • A “sentimental” theory:
    A “sentimental” theory: The price of an asset deviates from its fundamental value depending on waves of optimism or pessimism, and Twitter enables the measurement of investor sentiment.

Sources:
Main idea behind the work: what if having people complaining on Twitter about their malfunctioning iPhone will in fact result in Apple experience quarterly losses? Of course, the relationship is neither linear nor simple. It actually depends on the total number of users expressing a certain sentiment as well as on the intensity of their sentiments, and even in perfect circumstances, the correlation between general sentiment and stock oscillations needs to be proved.

*Klout score* (value that indicates the degree of social influence of a certain individual in the social media world) -- varies between 1 and 100 and to a higher value corresponds a higher influence power.

*Interesting* to study this correlation when extraordinary corporate events happen (for example, an acquisition, an IPO, the launch of a new *product*, the company distributing dividends, the CEO been fired, etc.).

Source: http://blogs.lse.ac.uk/usappblog/2017/10/14/can-twitter-sentiment-predict-stock-market-behaviour/
Can Twitter sentiment predict stock market behaviour?

Main idea behind the work: what if having people complaining on Twitter about their malfunctioning iPhone will in fact result in Apple experiencing quarterly losses? Of course, the relationship is neither linear nor simple. It actually depends on the total number of users expressing a certain sentiment as well as on the intensity of their sentiments, and even in perfect circumstances, the correlation between general sentiment and stock oscillations needs to be proved.

Klout score (value that indicates the degree of social influence of a certain individual in the social media world) -- varies between 1 and 100 and to a higher value corresponds a higher influence power.

It is interesting to study this correlation when extraordinary corporate events happen (for example, an acquisition, an IPO, the launch of a new product, the company distributing dividends, the CEO being fired, etc.)

Source: http://blogs.lse.ac.uk/usappblog/2017/10/14/can-twitter-sentiment-predict-stock-market-behaviour/
Source 2: Massive Text in E-commerce

• Consider you are large organization (e-commerce like amazon)
• Millions of customers are interacting every day.
• A automatic process is required to classify and understand customers’
  • Mood (Happy or sad)
  • Sentiment (positive or negative)
  • Message type: Complain or Appreciation
Type of Customers’ messages

• Understanding customer feedback requires understanding text
  • Product reviews
  • Customer feedback forms
  • Email messages
  • Opinion pieces
  • Consumer complaint logs
Why Text Analytics is important?

Some more applications

• Consumer information
  • Product reviews

• Marketing
  • Consumer attitudes
  • Trends

• Politics
  • Politicians want to know voters’ views
  • Voters want to know politicians’ stances and who else supports them

• Social
  • Find like-minded individuals or communities
Some other sources of text (possibly impacting business)

• Social Media Giants
  • Twitter feeds (we already saw it)
  • Blogs
  • Facebook (status updates and comments)
  • Reddit comments

Company may want to track about the brand through tweets or other social media websites
Text Analytics: Enron Email Corpus

• This corpus is valuable to computer scientists and social-network theorists in ways that the e-mails’ authors and recipients never could have intended. Because it is a rich example of how real people in a real organization use e-mail—full of mundane lunch plans, boring meeting notes, embarrassing flirtations that revealed at least one extramarital affair, and the damning missives that spelled out corruption—it has become the foundation of hundreds of research studies in fields as diverse as machine learning and workplace gender studies.

• This research has had widespread applications: computer scientists have used the corpus to train systems that automatically prioritize certain messages in an inbox and alert users that they may have forgotten about an important message. Other researchers use the Enron corpus to develop systems that automatically organize or summarize messages. Much of today's software for fraud detection, counterterrorism operations, and mining workplace behavioral patterns over e-mail has been somehow touched by the data set.

Why Text is difficult?

• Unstructured data
  • No: Table of records with fields (like feature collection)
  • Text fields can have varying number of words
  • Word order matters (and sometimes not)

• Intended for human consumption (and not for computers)

• Dirty text
  • No grammar
  • Misspelled words
  • Synonyms
  • Emojis
  • Sarcasm
Context is important

• Consider the text
  • “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.”

• What you expect the overall sentiment?

• What do you think of the word “incredible”? Positive or Negative?
Challenges

• Harder than topical classification,

• Must consider other features due to...
  • Subtlety of sentiment expression
    • Irony or Sarcasm
    • Expression of sentiment using neutral words
  • Domain/context dependence
    • words/phrases can mean different things in different contexts and domains
  • Effect of syntax on semantics
Clustering and Sentiment Analysis (opinion mining)

• Techniques
  • Lexical based techniques
    • When dictionaries are created in one substantive area and then applied to another problems, serious errors can occur
  • K-NN (bit more sophisticated)
  • Support vector machines (SVM): Another supervised algorithm

Additional reading (if you wish): http://liris.cnrs.fr/Documents/Liris-6508.pdf
Terminologies

• Document: Piece of text (one sentence like tweet or a book)
• Tokens or Terms: A document consists of individual tokens
  • Consider it is as a word for now
• Corpus: Collection of documents.
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.

- Case has been normalized: every term is in lowercase
- Words have been stemmed: their suffixes removed, so that verbs like announces, announced and announcing are all reduced to the term announc
- Stopwords have been removed (common words: the, and, of, and on)

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>Term</th>
<th>Count</th>
<th>Term</th>
<th>Count</th>
<th>Term</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>skype</td>
<td>3</td>
<td>microsoft</td>
<td>3</td>
<td>agreement</td>
<td>2</td>
<td>global</td>
<td>1</td>
</tr>
<tr>
<td>approv</td>
<td>1</td>
<td>announc</td>
<td>1</td>
<td>acquir</td>
<td>1</td>
<td>lead</td>
<td>1</td>
</tr>
<tr>
<td>definit</td>
<td>1</td>
<td>lake</td>
<td>1</td>
<td>communic</td>
<td>1</td>
<td>internet</td>
<td>1</td>
</tr>
<tr>
<td>board</td>
<td>1</td>
<td>led</td>
<td>1</td>
<td>director</td>
<td>1</td>
<td>corp</td>
<td>1</td>
</tr>
<tr>
<td>compani</td>
<td>1</td>
<td>investor</td>
<td>1</td>
<td>silver</td>
<td>1</td>
<td>billion</td>
<td>1</td>
</tr>
</tbody>
</table>
Document Term Frequency

- Each sentence is considered a separate document.
- A simple bag-of-words approach using term frequency would produce a table of term counts

<table>
<thead>
<tr>
<th>d1</th>
<th>d2</th>
<th>d3</th>
</tr>
</thead>
<tbody>
<tr>
<td>jazz music has a swing rhythm</td>
<td>swing is hard to explain</td>
<td>swing rhythm is a natural rhythm</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>explain</th>
<th>hard</th>
<th>has</th>
<th>is</th>
<th>jazz</th>
<th>music</th>
<th>natural</th>
<th>rhythm</th>
<th>swing</th>
<th>to</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>d2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>d3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Starting with Lexical Based

POSITIVE POLARITY (GOOD)
-a mesmerizing cinematic poem from the first frame to the last.
-a well-put-together piece of urban satire.
-one can’t deny its seriousness and quality.
-hard to resist.
-a naturally funny film, home movie makes you crave chris smith’s next movie.
-a true-blue delight.
-a fun ride.
-a surprisingly funny movie.
-the script is smart and dark - hallelujah for small favors.
-a flick about our infantilized culture that isn’t entirely infantile.

NEGATIVE POLARITY (BAD)
-unfortunately the story and the actors are served with a hack script.
-too slow for a younger crowd, too shallow for an older one.
-terminally brain dead production.
-one lousy movie.
-this movie . . . doesn’t deserve the energy it takes to describe how bad it is.
-a cleverly crafted but ultimately hollow mockumentary.
-it’s an 88-minute highlight reel that’s 86 minutes too long.
-the whole affair is as predictable as can be.

source: https://www.youtube.com/watch?v=ytUHvMNnzZk
Comparing text with corpus

Good: 506
Bad: 507
Goodness: $\frac{506}{506+507} = 0.5$
Badness: $\frac{507}{506+507} = 0.5$

Good: 10
Bad: 14
Goodness: $\frac{10}{14+10} = 0.41$
Badness: $\frac{14}{14+10} = 0.59$

“it's rather like a lifetime special -- pleasant, sweet and forgettable.”

Good: 15
Bad: 6
Goodness: $\frac{15}{6+15} = 0.71$
Badness: $\frac{6}{6+15} = 0.29$

Good: 46
Bad: 22
Goodness: $\frac{46}{46+22} = 0.68$
Badness: $\frac{22}{46+22} = 0.32
Calculating the positive and negative score

"it's rather like a lifetime special -- pleasant, sweet and forgettable."

<table>
<thead>
<tr>
<th></th>
<th>#GOOD</th>
<th>#BAD</th>
<th>GOODNESS</th>
<th>BADNESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>it's</td>
<td>506</td>
<td>507</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>rather</td>
<td>42</td>
<td>63</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>like</td>
<td>242</td>
<td>396</td>
<td>0.61</td>
<td>0.39</td>
</tr>
<tr>
<td>a</td>
<td>3446</td>
<td>3112</td>
<td>0.53</td>
<td>0.47</td>
</tr>
<tr>
<td>lifetime</td>
<td>3</td>
<td>5</td>
<td>0.38</td>
<td>0.62</td>
</tr>
<tr>
<td>special</td>
<td>29</td>
<td>40</td>
<td>0.42</td>
<td>0.58</td>
</tr>
<tr>
<td>pleasant</td>
<td>15</td>
<td>6</td>
<td>0.71</td>
<td>0.29</td>
</tr>
<tr>
<td>sweet</td>
<td>46</td>
<td>22</td>
<td>0.68</td>
<td>0.32</td>
</tr>
<tr>
<td>and</td>
<td>3198</td>
<td>2371</td>
<td>0.57</td>
<td>0.43</td>
</tr>
<tr>
<td>forgettable</td>
<td>10</td>
<td>14</td>
<td>0.42</td>
<td>0.58</td>
</tr>
</tbody>
</table>

SUM: 5.22 4.8

So we should classify this as a POSITIVE review!

We are ignoring the order of the words in the sentence.
“it's rather like a lifetime special -- pleasant, sweet and forgettable.”

<table>
<thead>
<tr>
<th>Word</th>
<th>#GOOD</th>
<th>#BAD</th>
<th>GOODNESS</th>
<th>BADNESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>it's</td>
<td>506</td>
<td>507</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>rather</td>
<td>42</td>
<td>63</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>like</td>
<td>242</td>
<td>396</td>
<td>0.61</td>
<td>0.39</td>
</tr>
<tr>
<td>a</td>
<td>3446</td>
<td>3112</td>
<td>0.53</td>
<td>0.47</td>
</tr>
<tr>
<td>lifetime</td>
<td>3</td>
<td>5</td>
<td>0.38</td>
<td>0.62</td>
</tr>
<tr>
<td>special</td>
<td>29</td>
<td>40</td>
<td>0.42</td>
<td>0.58</td>
</tr>
<tr>
<td>pleasant</td>
<td>15</td>
<td>6</td>
<td>0.71</td>
<td>0.29</td>
</tr>
<tr>
<td>sweet</td>
<td>46</td>
<td>22</td>
<td>0.68</td>
<td>0.32</td>
</tr>
<tr>
<td>and</td>
<td>3198</td>
<td>2371</td>
<td>0.57</td>
<td>0.43</td>
</tr>
<tr>
<td>forgettable</td>
<td>10</td>
<td>14</td>
<td>0.42</td>
<td>0.58</td>
</tr>
</tbody>
</table>

SUM: 5.22 4.8
So we should classify this as a POSITIVE review!
We won’t catch everything....

“this movie makes Catwoman look like a great movie.”

“a terrible movie that some people will nevertheless find moving.”

“well-made but mush-hearted.”

“your children will be occupied for 72 minutes.”

... ah well.
Techniques for improvement

- Bi-gram is taking 2 words at a time.

n-grams

“this movie will knock your socks off!”

Stemming

“I really liked this movie!”

“I really like this movie!”

cats, catlike, catty -> cat

watching, watched -> watch

liked, liking -> lik
Techniques for improvement

Stop words

Part of speech tagging

“I did not like this movie”
NOTE: CARELESS STOPWORD ELIMINATION

A word of caution: stopword elimination is not always a good idea. In titles, for example, common words may be very significant. For example, *The Road*, Cormac McCarthy’s story of a father and son surviving in a post-apocalyptic world, is very different from John Kerouac’s famous novel *On the Road*— though careless stopword removal may cause them to be represented identically. Similarly, the recent movie thriller *Stoker* should not be confused with the 1935 film comedy *The Stoker*. [80]
Techniques for improvement

• Use of linguistic knowledge

WordNet

Movie:
Noun

- S: (n) movie, film, picture, moving picture, moving-picture show, motion picture, motion-picture show, picture show, pic, flick (a form of entertainment that enacts a story by sound and a sequence of images giving the illusion of continuous movement) "they went to a movie every Saturday night"; "the film was shot on location"

Wonderful:
Adjective

- S: (adj) fantastic, grand, howling, marvelous, marvellous, rattling, terrific, tremendous, wonderful, wondrous (extraordinarily good or great; used especially as intensifiers) "a fantastic trip to the Orient"; "the film was fantastic!"; "a howling success"; "a marvelous collection of rare books"; "had a rattling conversation about politics"; "a tremendous achievement"
A different approach: k-nearest neighbor (kNN)

“it's rather like a lifetime special -- pleasant, sweet and forgettable.”

IDEA: Find most similar sentence in train set

closest match:

i liked this movie. made for a pleasant evening.
Classification: K-NN Neighbor (not K-Means)

For Categorical

Hamming Distance

\[ D_H = \sum_{i=1}^{k} |x_i - y_i| \]

\[ x = y \Rightarrow D = 0 \]
\[ x \neq y \Rightarrow D = 1 \]

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Male</td>
<td>0</td>
</tr>
<tr>
<td>Male</td>
<td>Female</td>
<td>1</td>
</tr>
</tbody>
</table>

For continuous

Distance functions

- Euclidean: \[ \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \]
- Manhattan: \[ \sum_{i=1}^{k} |x_i - y_i| \]
- Minkowski: \[ \left( \sum_{i=1}^{k} |x_i - y_i|^q \right)^{1/q} \]

N.B

Keep K as odd (in case you have two classes)
K= sqrt (Numbers of training data)

Source: https://www.youtube.com/watch?v=v6278Cjf_qA
K-NN Neighbor (not K-Means)

1. A very simple classification and regression algorithm
   a. In case of classification, new data points get classified in a particular class
   b. In case of regression, new data gets labeled based on the avg. value of k nearest neighbour
2. It is a lazy learner because it doesn’t learn much from the training data
3. It is a supervised learning algorithm

Source: https://www.youtube.com/watch?v=v6278Cjf_qA
K-NN

- We can now use the training set to classify an unknown case (Age=48 and Loan=$142,000) using Euclidean distance.
- If K=1 then the nearest neighbor is the last case in the training set with Default=Y.
- \[ D = \sqrt{(48-33)^2 + (142000-150000)^2} = 8000.01 \] >> Default=Y

<table>
<thead>
<tr>
<th>Age</th>
<th>Loan</th>
<th>Default</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>$40,000</td>
<td>N</td>
<td>102000</td>
</tr>
<tr>
<td>35</td>
<td>$60,000</td>
<td>N</td>
<td>82000</td>
</tr>
<tr>
<td>45</td>
<td>$80,000</td>
<td>N</td>
<td>62000</td>
</tr>
<tr>
<td>20</td>
<td>$20,000</td>
<td>N</td>
<td>122000</td>
</tr>
<tr>
<td>35</td>
<td>$120,000</td>
<td>N</td>
<td>22000</td>
</tr>
<tr>
<td>52</td>
<td>$180,000</td>
<td>N</td>
<td>124000</td>
</tr>
<tr>
<td>23</td>
<td>$95,000</td>
<td>Y</td>
<td>47000</td>
</tr>
<tr>
<td>40</td>
<td>$62,000</td>
<td>Y</td>
<td>80000</td>
</tr>
<tr>
<td>60</td>
<td>$100,000</td>
<td>Y</td>
<td>42000</td>
</tr>
<tr>
<td>48</td>
<td>$220,000</td>
<td>Y</td>
<td>78000</td>
</tr>
<tr>
<td>33</td>
<td>$150,000</td>
<td>Y</td>
<td>80000</td>
</tr>
</tbody>
</table>

With K=3, there are two Default=Y and one Default=N out of three closest neighbors. The prediction for the unknown case is again Default=Y

N.B
Keep K as d
Standardizing (and paying the price)

Using the standardized distance on the same training set, the unknown case returned a different neighbor which is not a good sign of robustness.

<table>
<thead>
<tr>
<th>Age</th>
<th>Loan</th>
<th>Default</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.125</td>
<td>0.11</td>
<td>N</td>
<td>0.7652</td>
</tr>
<tr>
<td>0.375</td>
<td>0.21</td>
<td>N</td>
<td>0.5200</td>
</tr>
<tr>
<td>0.625</td>
<td>0.31</td>
<td>N</td>
<td>0.3160</td>
</tr>
<tr>
<td>0</td>
<td>0.01</td>
<td>N</td>
<td>0.9245</td>
</tr>
<tr>
<td>0.375</td>
<td>0.50</td>
<td>N</td>
<td>0.3428</td>
</tr>
<tr>
<td>0.8</td>
<td>0.00</td>
<td>N</td>
<td>0.6220</td>
</tr>
<tr>
<td>0.075</td>
<td>0.38</td>
<td>Y</td>
<td>0.6669</td>
</tr>
<tr>
<td>0.5</td>
<td>0.22</td>
<td>Y</td>
<td>0.4437</td>
</tr>
<tr>
<td>1</td>
<td>0.41</td>
<td>Y</td>
<td>0.3650</td>
</tr>
<tr>
<td>0.7</td>
<td>1.00</td>
<td>Y</td>
<td>0.3861</td>
</tr>
<tr>
<td>0.325</td>
<td>0.65</td>
<td>Y</td>
<td>0.3771</td>
</tr>
</tbody>
</table>

\[
X_s = \frac{X - \text{Min}}{\text{Max} - \text{Min}}
\]
Processing pipeline

1. Input sentence: “a surprisingly funny movie.”

2. Split it up: [a, surprisingly, funny, movie, . ]

3. Clean up (remove stop words, punctuation, etc): [surprisingly, funny, movie]

4. Stem: [surprisingly, funny, movie]
Later at test time...

“surprisingly funny movie!”

[surprisingly, funny, movie]

[funny, movie, i, recommend, it]

“funny movie, i recommend it.”
Later at test time...

funni occurs in 0.1 of all reviews

movi occurs in 0.7 of all reviews

funni, movi, i, recommend, it

SCORE = 1/0.1 + 1/0.7
       = 10 + 1.43
       = 11.43
Later at test time...

[enigma, surprisingly, funny, movi]

Only occurs in 0.001 of reviews!!!

[enigma, funny, movi, i, recommend, it]

\[
\text{SCORE} = \frac{1}{0.001} + \frac{1}{0.1} + \frac{1}{0.7} \\
= 1000 + 10 + 1.43 \\
= 1011.43
\]
LOG FUNCTION: "SQUASHER" (for values >1)

BASE 10 LOG EXAMPLE:

\[
\begin{align*}
\log(1) &= 0 \\
\log(5) &= 0.7 \\
\log(10) &= 1 \\
\log(100) &= 2 \\
\log(1000) &= 3 \\
\log(3000) &= 3.47 \\
\cdots \\
\log(1000000000000000000) &= 14 \\
\text{Etc.}
\end{align*}
\]
Later at test time...

[enigma, surprisingly, funny, movie]

Only occurs in 0.001 of reviews!!!

[enigma, funny, movie, i, recommend, it]

\[
\text{SCORE} = \log(1/0.001) + \log(1/0.1) + \log(1/0.7) \\
= \log(1000) + \log(10) + \log(1.43) \\
= 3 + 1 + 0.15 \\
= 4.15
\]
K-NN Text Classification

• We just computed the similarity score between the test document with another training document.
• Pick top K similar documents
• Based on majority, assign the polarity or classify the document.
Smart Approach by Facebook!
Demo time!

https://courses.cs.ut.ee/2018/bda/spring/Main/Practice