Chapter 10

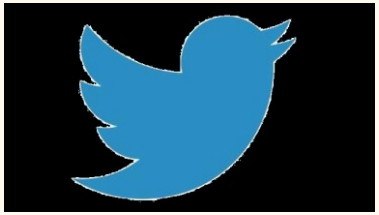
Text Mining

Dr. Steffen Herbold
herbold@cs.uni-goettingen.de

Outline

- Overview
- Challenges for Text Mining
- Summary

Example for Textual Data



Oct 4, 2018 08:03:25 PM Beautiful evening in Rochester, Minnesota. VOTE, VOTE, VOTE! https://t.co/SyxrxvTpZE [Twitter for iPhone]
Oct 4, 2018 07:52:20 PM Thank you Minnesota - I love you! https://t.co/eQC2NqdIil [Twitter for iPhone]
Oct 4, 2018 05:58:21 PM Just made my second stop in Minnesota for a MAKE AMERICA GREAT AGA
@KarinHousley to the U.S. Senate, and we need the strong leadership of @TomEmmer, @Jason2
@PeteStauber in the U.S. House! [Twitter for iPhone]
Oct 4, 2018 05:17:48 PM Congressman Bishop is doing a GREAT job! He helped pass tax reform w
EVERYONE! Nancy Pelosi is spending hundreds of thousands of dollars on his opponent because they
agenda of higher taxes and wasteful spending! [Twitter for iPhone]
Oct 4, 2018 02:29:27 PM “U.S. Stocks Widen Global Lead” https://t.co/Snhv08ulcO [Twitter for iPhone]
Oct 4, 2018 02:17:28 PM Statement on National Strategy for Counterterrorism: https://t.co ycjMAV
[Twitter for iPhone]
Oct 4, 2018 12:38:08 PM Working hard, thank you! https://t.co/6HQVaEXH0I [Twitter for i
Oct 4, 2018 09:17:01 AM This is now the 7th. time the FBI has investigated Judge Kavan uld still not
be good enough for the Obstructionist Democrats. [Twitter for iPhone]
Oct 4, 2018 09:01:13 AM RT @ChatByCC: While armed with the power of our Vote, we p Never
doubt this was an American revolution #MAGA… [Twitter for iPhone]
Oct 4, 2018 08:54:58 AM This is a very important time in our country. Due Process, Fairn now on
trial! [Twitter for iPhone]
Oct 4, 2018 08:34:16 AM Our country’s great First Lady, Melania, is doing really well in A d she
loves them! It is a beautiful thing to see. [Twitter for iPhone]
Oct 4, 2018 07:16:41 AM The harsh and unfair treatment of Judge Brett Kavanaugh is ha pact on
voters. The PEOPLE get it far better than the politicians. Most importantly, this great life c espicable
Democrats and totally uncorroborated allegations! [Twitter for iPhone]

How do you analyze this?

Text Mining in General

- Inferring information from textual data
- Techniques for text mining as diverse as the texts themselves!
- The following demonstrates a relatively standard workflow for processing text data
  - Create corpus
  - Pre-process contents
  - Define bag-of-words

Documents and Corpus

- Create documents to create a corpus

Each tweet a document

Oct 4, 2018 08:03:25 PM Beautiful evening in Rochester, Minnesota. VOTE, VOTE, VOTE! https://t.co/SyxrxvTpZE [Twitter for iPhone]

Oct 4, 2018 07:52:20 PM Thank you Minnesota - I love you! https://t.co/eQC2NqdIil [Twitter for iPhone]

Oct 4, 2018 05:58:21 PM Just made my second stop in Minnesota for a MAKE AMERICA GREAT AGAIN rally. We need to elect @KarinHousley to the U.S. Senate, and we need the strong leadership of @TomEmmer, @Jason2CD, @JimHagedornMN and @PeteStauber in the U.S. House! [Twitter for iPhone]

Oct 4, 2018 05:17:48 PM Congressman Bishop is doing a GREAT job! He helped pass tax reform which lowered taxes for EVERYONE! Nancy Pelosi is spending hundreds of thousands of dollars on his opponent because they both support a liberal agenda of higher taxes and wasteful spending! [Twitter for iPhone]

Oct 4, 2018 02:29:27 PM “U.S. Stocks Widen Global Lead” https://t.co/Snhv08ulcO [Twitter for iPhone]

Oct 4, 2018 02:17:28 PM Statement on National Strategy for Counterterrorism: https://t.co/ajFBg9Elsj https://t.co/Qr56ycjMAV [Twitter for iPhone]

Oct 4, 2018 12:38:08 PM Working hard, thank you! https://t.co/6HQVaEXH0I [Twitter for iPhone]

Oct 4, 2018 09:17:01 AM This is now the 7th. time the FBI has investigated Judge Kavanaugh. If we made it 100, it would still not be good enough for the Obstructionist Democrats. [Twitter for iPhone]

All tweets together the corpus

Unstructured Data → Structured Data

- Identify relevant content

→ Remove irrelevant content

Drop shortened links

Drop device

Oct 4, 2018 08:03:25 PM Beautiful evening in Rochester, Minnesota. VOTE, VOTE, VOTE! https://t.co/SyxrxvTpZE [Twitter for iPhone]

Oct 4, 2018 07:52:20 PM Thank you Minnesota - I love you! https://t.co/eQC2NqdIil [Twitter for iPhone]

Oct 4, 2018 05:58:21 PM Just made my second stop in Minnesota for a MAKE AMERICA GREAT AGAIN rally. We need to elect @KarinHousley to the U.S. Senate, and we need the strong leadership of @TomEmmer, @Jason2CD, @JimHagedornMN and @PeteStauber in the U.S. House! [Twitter for iPhone]

Oct 4, 2018 05:17:48 PM Congressman Bishop is doing a GREAT job! He helped pass tax reform which lowered taxes for EVERYONE! Nancy Pelosi is spending hundreds of thousands of dollars on his opponent because they both support a liberal agenda of higher taxes and wasteful spending! [Twitter for iPhone]

Oct 4, 2018 02:29:27 PM “U.S. Stocks Widen Global Lead” https://t.co/Snhv08ulcO [Twitter for iPhone]

Oct 4, 2018 02:17:28 PM Statement on National Strategy for Counterterrorism: https://t.co/ajFBg9Elsj https://t.co/Qr56ycjMAV [Twitter for iPhone]

Oct 4, 2018 12:38:08 PM Working hard, thank you! https://t.co/6HQVaEXH0I [Twitter for iPhone]

Oct 4, 2018 09:17:01 AM This is now the 7th. time the FBI has investigated Judge Kavanaugh. If we made it 100, it would still not be good enough for the Obstructionist Democrats. [Twitter for iPhone]

Drop timestamp

Punctuation and cases

- Usually do not carry information

→ Can be removed

beautiful evening in rochester minnesota vote vote vote

thank you minnesota i love you

just made my second stop in minnesota for a make america great again rally we need to elect karinhousley to the us senate and we need the strong leadership of tomemmer jason2cd jimhagedornmn and petestauber in the us house

congressman bishop is doing a great job he helped pass tax reform which lowered taxes for everyone nancy pelosi is spending hundreds of thousands of dollars on his opponent because they both support a liberal agenda of higher taxes and wasteful spending

us stocks widen global lead

statement on national strategy for counterterrorism

working hard thank you

this is now the 7th time the fbi has investigated judge kavanaugh if we made it 100 it would still not be good enough for the obstructionist democrats

Stop words

- Most common words in a language (a, the, I, we, to, too, …)

→ Can be removed

beautiful evening rochester minnesota vote vote vote

thank minnesota love

just made second stop minnesota make america great again rally need elect karinhousley senate need strong leadership tomemmer jason2cd jimhagedornmn petestauber house

congressman bishop doing great job helped pass tax reform lowered taxes everyone nancy pelosi spending hundreds thousands dollars opponent both support liberal agenda higher taxes wasteful spending

stocks widen global lead

statement national strategy counterterrorism

working hard thank

now 7th time fbi investigated judge kavanaugh made 100 would still good enough obstructionist democrats

Stemming and Lemmatization

- Stemming: reduce terms to common stem (spending → spend)
- Lemmatization: use common term for synonyms (better → good)

beautiful evening rochester minnesota vote vote vote

thank minnesota love

just make second stop minnesota make america great again rally need elect karinhousley senate need strong leadership tomemmer jason2cd jimhagedornmn petestauber house

congressman bishop do good job help pass tax reform lower tax everyone nancy pelosi spend hundred thousand dollar opponent both support liberal agenda high tax waste spend

stock wide global lead

statement national strategy counterterrorism

work hard thank

now 7th time fbi investigate judge kavanaugh make 100 would still good enough obstruct democrat

Bag-of-Words

Every term one dimension

beauti-ful	evenin g	roches -ter	minne-sota	vote	thank	love	just	make	second	stop	amer-ica	great	…
1	1	1	1	3	0	0	0	0	0	0	0	0	…
0	0	0	1	0	1	1	0	0	0	0	0	0	…
0	0	0	1	0	0	0	0	2	1	1	1	1	…
0	0	0	0	0	0	0	0	0	0	0	0	0	…
0	0	0	0	0	0	0	0	0	0	0	0	0	…
0	0	0	0	0	1	0	0	0	0	0	0	0	…
0	0	0	0	0	0	0	0	1	0	0	0	0	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…

Number of occurences in document

Inverse Document Frequency

- Absolute frequency of words problematic for discrimination
  - Favors words that occur often and in many documents
  - Similar to stop words
- Inverse document frequency for the uniqueness of terms
  - $idf_i = \log \frac{N}{tf_i}$
- Inverse document frequency to weight terms by uniqueness
  - $tf - idf = tf_i \cdot idf_i$

Downstream Analysis

- Bag of words base structure

- Allows many approaches for analysis
  - Classification
    - Usually requires manual labeling of data
  - Clustering
  - Sentiment analysis
    - Based on word counts
  - Information retrieval
    - Identification of related documents
  - Visualizations

Outline

- Overview
- **Challenges for Text Mining**
- Summary

High Dimensional Data

- Bag of words gets high dimensional extremely fast
  - Still over sixty different words after stemming/lemmatization in the tweet example
- Can also have a huge amount of documents
  - >39000 tweets by donald trump in total
- High dimension+many documents → very high runtime
- Requires
  - Very efficient algorithms
  - Often only possible through massive parallelization
  - Naive Bayes is a popular choice

Ambiguities

- Synonyms
  - break (take a break, break something)
  - Will all be grouped together by a bag of words
- Semantic changes in interpretation of the same sentences
  - I hit the man with a stick. (Used a stick to hit the man)
  - I hit the man with a stick (I hit the man who was holding a stick)
- Can often only be inferred from a greater context

→ Often impossible to resolve and leads to noise in the analysis

Capturing Syntax

- Bag of words ignores syntax
  - Nine sentences with different meanings and the same words
    1. Only he told his mistress that he loved her. (Nobody else did)
    2. He only told his mistress that he loved her. (He didn't show her)
    3. He told only his mistress that he loved her. (Kept it a secret from everyone else)
    4. He told his only mistress that he loved her. (Stresses that he had only ONE!)
    5. He told his mistress only that he loved her. (Didn't tell her anything else)
    6. He told his mistress that only he loved her. ("I'm all you got, nobody else wants you.")
    7. He told his mistress that he only loved her. (Not that he wanted to marry her.)
    8. He told his mistress that he loved only her. (Yeah, don't they all...).
    9. He told his mistress that he loved her only. (Similar to above one).

- Capturing word orders and other relationships greatly increases the dimension
  - E.g., n-grams

(bag of words over n-tuples)

And the list goes on

- Bad spelling
- Slang
- Homonyms not captured by lemmatization
- Encodings and special characters
- …

Outline

- Overview
- Challenges for Text Mining
- **Summary**

Summary

- Text mining is the analysis of textual data

- Requires imposing a structure upon the unstructured texts
  - Bag of words
- There are some standard techniques
  - Removing punctuation, cases, stopwords
  - Stemming/Lemmatization
  - Often also removing numbers and special characters

  → Depends on context!

- Many problems due to noise in the textual data