Texts Words Meaning?

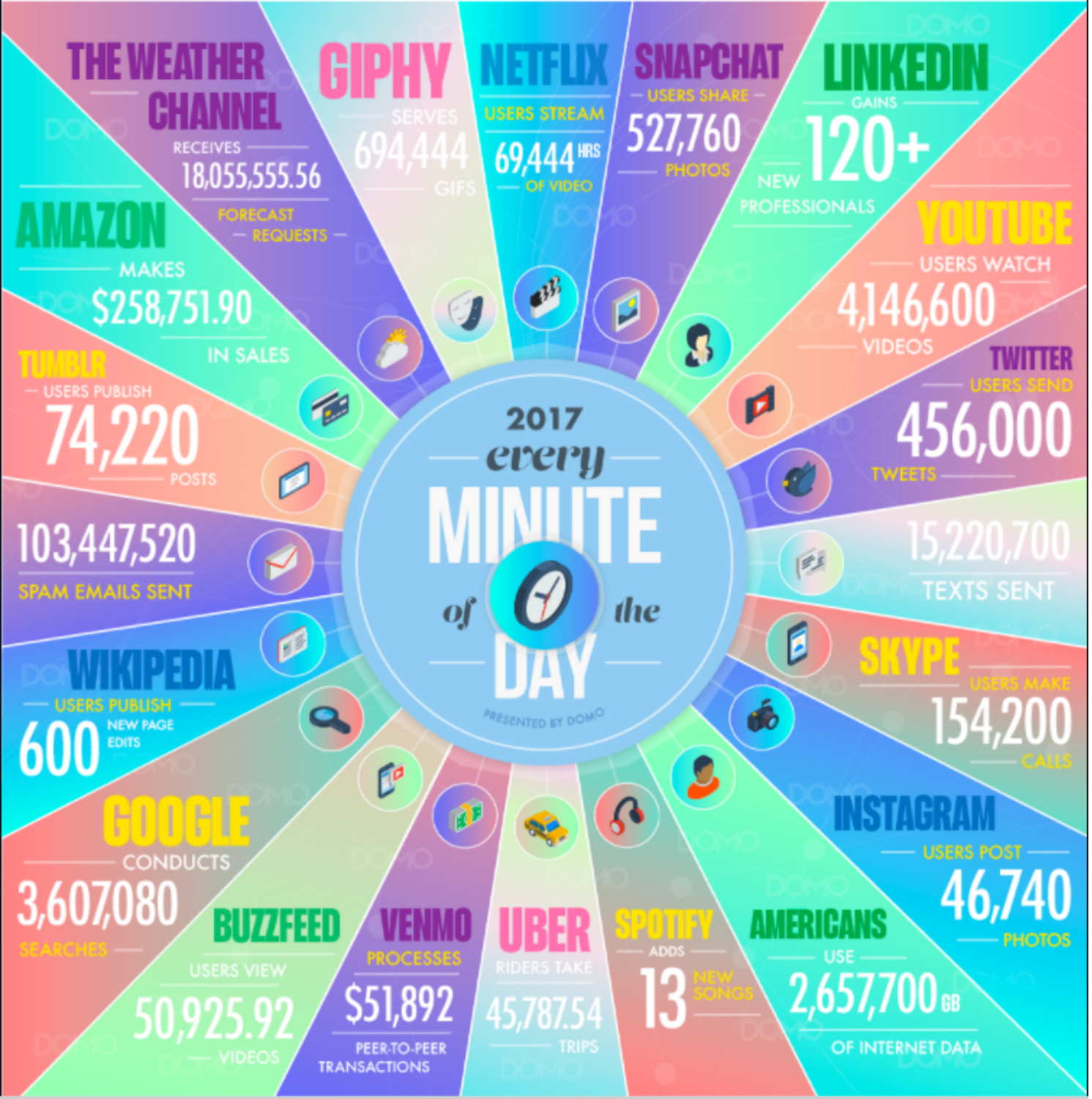
Elisabeth Bommes

mbr targeting / Ströer Labs
and
Ladislaus von Bortkiewicz Chair of Statistics
Humboldt-Universität zu Berlin
http://lvb.wiwi.hu-berlin.de

Rise of Unstructured Data

 instead of 

- Text, speech, images, video
- Merrill Lynch (1998): 80 - 90% of data is unstructured
- Gantz and Reinsel (2012)
  - $<$ 1% of data is analyzed
  - 40 zettabytes in 2020 (1 ZB = 1 billion TB)
- Implicit structure
  - Text: punctuation, part of speech
  - Images: coordinates, colors



Source
domo.com

One fits All?

How to

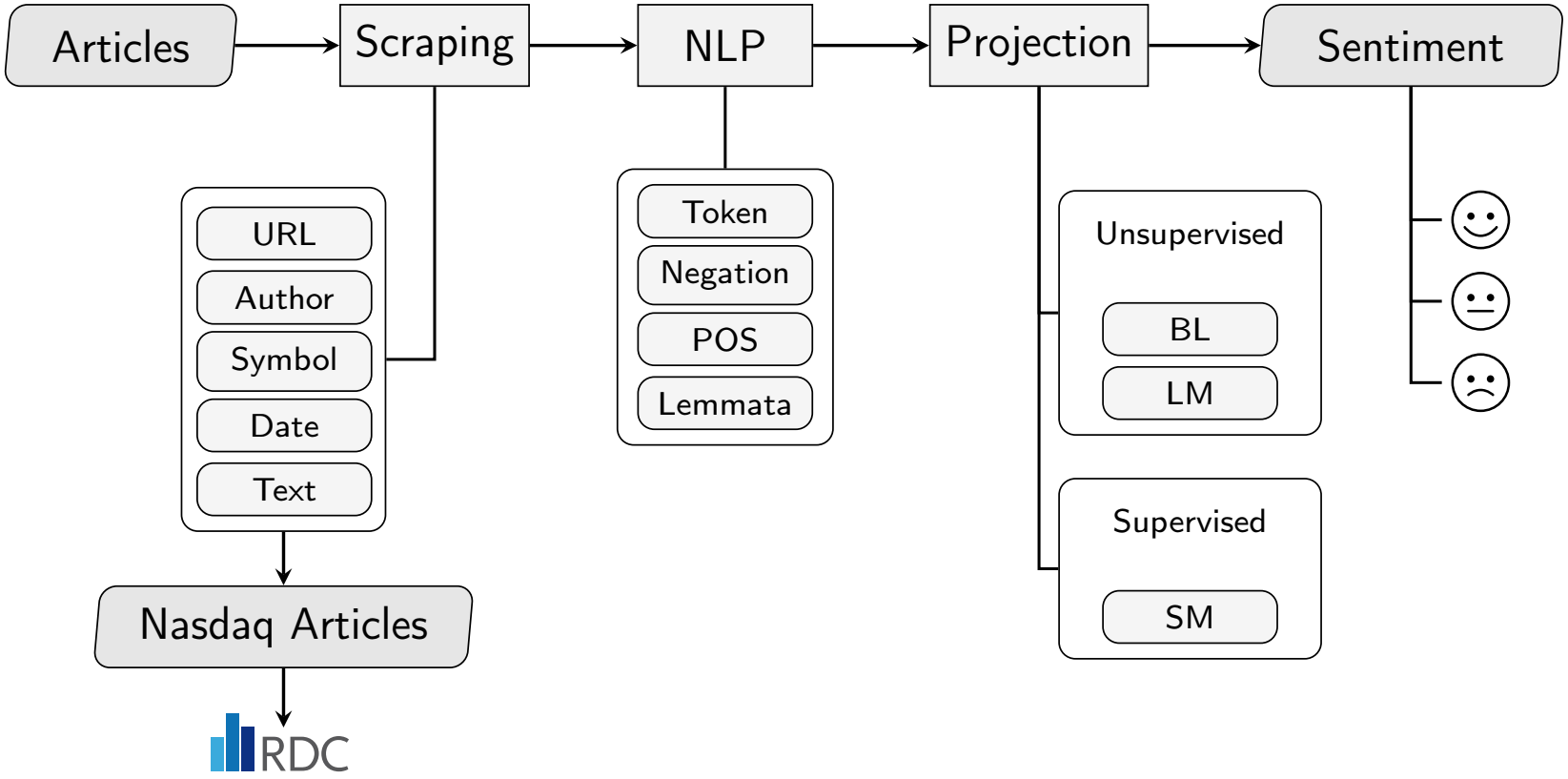
obtain
process
model
interpret

texts

?

Depends on domain and question.

Sentiment Pipeline



Outline

1. Motivation ✓
2. Obtaining Data
3. Processing Data
4. Topic Modeling
5. Sentiment Classification
6. Conclusion

Obtaining

- Ready to use
  - Gold standard corpora (GSC)
  - Classification: 20 Newsgroups REUTERS YouTube
  - Sentiment: amazon.com tripadvisor
- APIs
  - The New York Times theguardian
  - Sometimes no archive, request limitations
- Crawling / Scraping
  - Data for any domain of interest
  - Legality?

Gold Standard Corpora

- For training and evaluation of algorithms
- Manual annotations
  - Syntax, semantics lexical knowledge
  - E.g. entities, grammatical structures
- Multiple experts
  - Inter-annotator agreement
  - Time-consuming and expensive
- Example: Penn Treebank
  - 4.5 million English words
  - GSC for syntactical tagging

Wissler et al. (2014)

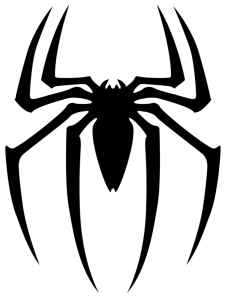
New York Times APIs

- Articles
  - 1851 until today
  - Headlines, abstracts, first paragraph, meta information
- Semantics
  - People, places, organizations
- Book and movie reviews
- Geo

Example: NYT API

Crawling and Scraping

- Crawling
  - Any information
  - Follows links
  - General information extraction



- Scraper
  - Specific information
  - Specific web pages
  - Easy to obtain high quality data

Legality of Web Scraping

- It is public / Google does it
  - Search engines add value
  - Log in systems, paywalls, ...?
- Highly context specific
  - Commerical v non-commercial
  - Internal v third party use
- Technicalities
  - Bandwidth usage
  - Denial-of-service (DoS) attack

European Union

- Ryanair Ltd v PR Aviation BV (2015)
  - PR Aviation: price comparison of flights
  - Copyright and database right infringement?
  - ToS prohibited data extraction for commercial purposes
- Decision by Court of Justice of the European Union
  - No infringement of intellectual property, no creative input
  - ToS still apply, liability in terms of breach of contract
- In contrast NLA v Meltwater (2013)
  - Scraping of news headlines and links to articles
  - Intellectual property is infringed because of creative input

United States

Pro

- Web data is public, should be accessible
- Unfair market power of Facebook, Google, LinkedIn, ...
- First Amendment protects information gathering

Contra

- Copyright infringement
- Breach of contract
- Violation of the Computer Fraud and Abuse Act (CFAA), 1986
- Trespass to chattels

LinkedIn v hiQ and vice versa

If you exclude someone from sites like LinkedIn, Facebook and Twitter, you are excluding them from the modern version of the town square.

Laurence Tribe, Harvard law professor

- hiQ predicts who is when quitting their job
- LinkedIn: CFAA violation, hiQ: blocked
- LinkedIn ordered to give access to public profiles

Academia is save, right?

Aaron Swartz

- Harvard research fellow
- Automatic download of JSTOR articles
- Laptop in restricted closet at MIT
- No civil law suit by MIT and JSTOR
- Federal charges: wire fraud, CFAA violations
- Possible penalty of $1 million and 35 years in prison

Unclear outcome, suicide on January 11, 2013

Bright Side

Cap Verde is beautiful and does not extradite

Ethical Scraping for Academia

- Technical
  - Use API if provided
  - Appear as a bot, not as a human
  - Provide user agent string with contact data
  - Decreased rate of requests
  - Check robots.txt Google's robots.txt
- Usage
  - Strictly non-commercial
  - Restrict further access to academia
- Ask for permission, not for forgiveness!

Scraping How To

- Complete framework: Scrapy
- Fast and easy: Beautiful Soup
- Low level: lxml

- Complete framework: RCrawler
- Fast and easy: rvest
- Low level: XML

Beeradvocate

- Largest beer rating community
- Founded in 1996, print magazine since 2006
- Mostly "exotic" and craft beer

- Still independent
  - Anheuser-Busch InBev acquires beer related websites
  - Examples: RateBeer, The Beer Necessities, ...

- Breweries in the US
  - Less than 100 in 1980s, more than 5,000 in 2016
  - Craft beer boom

Individual Review

Sort by: Recent | High | Low | **Top Raters** | Alström Bros

first ← prev | 1-25 | 26-50 | 51-75 | next → last

✔ **Ratings: 5,928** | **Reviews: 1,915**

4.13/5 rDev -2.1%

look: 4.5 | smell: 4 | taste: 4 | feel: 4 | overall: 4.5

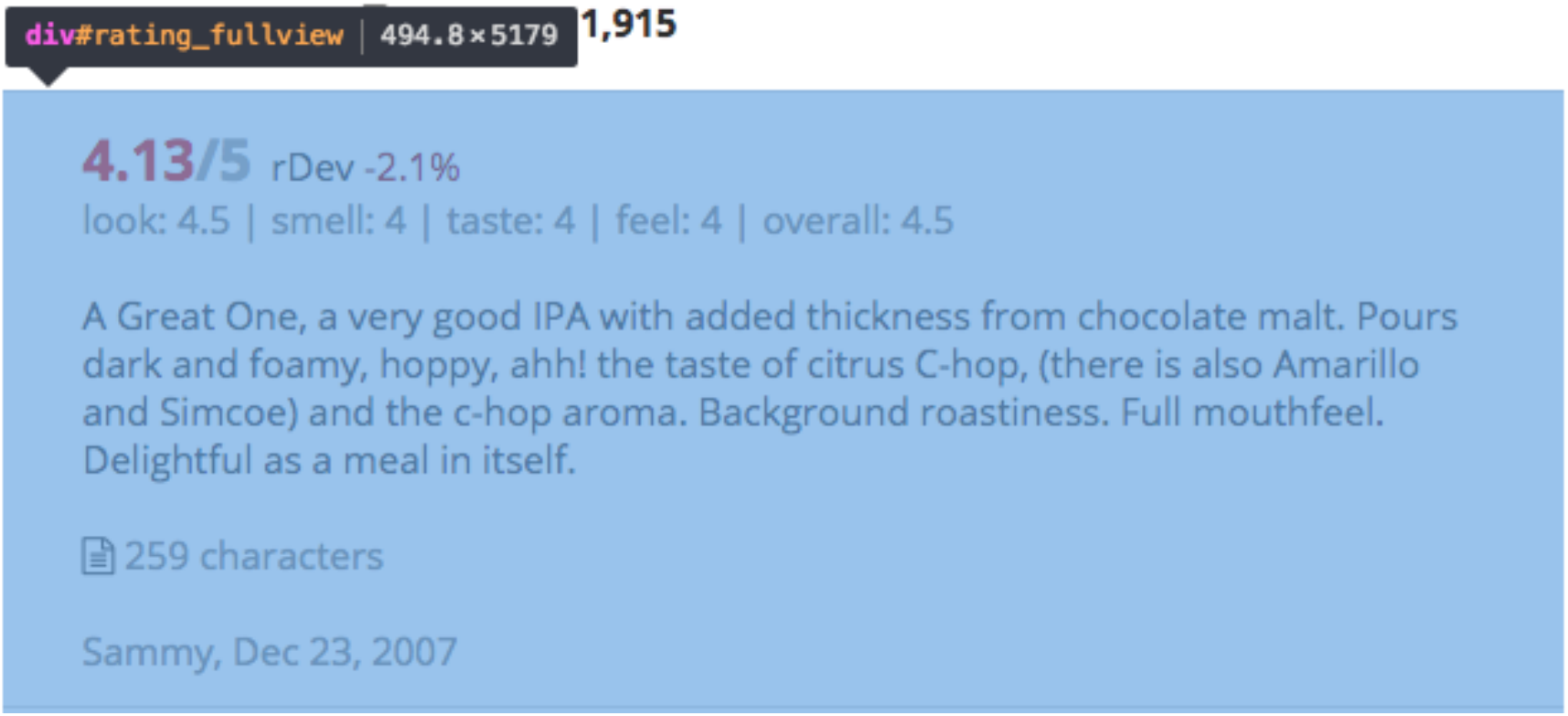
A Great One, a very good IPA with added thickness from chocolate malt. Pours dark and foamy, hoppy, ahh! the taste of citrus C-hop, (there is also Amarillo and Simcoe) and the c-hop aroma. Background roastiness. Full mouthfeel. Delightful as a meal in itself.

259 characters

Sammy, Dec 23, 2007

Source

xPath Inspector



Example: Scraper

Processing

- Always needed in text mining
- Vast improvement of model
- Often depends on theoretical model
  - Time series models for stock returns
- Just a side note in scientific articles
  - Well established "gold standard" processing for many domains
  - No scientific news value

Example: Beer Review

An OK lager. Light, crisp, but nothing special. Stacked against the great pilsners of the world or similar offerings, this beer is mediocre. But still solid enough to put it head and shoulders above any macro.

Stats for all reviews

Text Normalization: Basics

[(an, ok lager), (light, crisp, but, nothing, special), (stacked, against, the, great, pilsners, of, the, world, or, similar, offerings, this, beer, is, mediocre), (but, still, solid, enough, to, put, it, head, and, shoulders, above, any, macro)]

- Lowercase
- Removal of non alphabetic characters
- Word and sentence tokenization

Part of Speech Tagging

- More than nouns, verbs, adjectives
  - Penn Treebank: 45 tags
  - Brown Corpus: 87 tags
- Useful for
  - Information retrieval
  - Word-sense disambiguation
  - Shallow parsing of names and other named entities
- Math Tagging, Kristianto et al (2012)
  - Extract mathematical definitions
  - Pattern matching and machine learning

Nouns

An OK **lager**. **Light**, crisp, but **nothing** special. Stacked against the great **pilsners** of the **world** or similar **offerings**, this **beer** is mediocre. But still solid enough to put it **head** and **shoulders** above any **macro**.

NOUN

- Which topic?
- Which beer category? Lager, pilsner
- Stats for all Nouns

POS Tagging is not perfect

An OK **lager**. **Light**, crisp, but **nothing** special. Stacked against the great **pilsners** of the **world** or similar **offerings**, this **beer** is mediocre. But still solid enough to put it **head** and **shoulders** above any **macro**.

NOUN

- Which topic? Beer
- Which beer category? Lager, pilsner
- Stats for all Nouns

Verbs

An OK **lager**. **Light**, crisp, but **nothing** special. **Stacked** against the great **pilsners** of the **world** or similar **offerings**, this **beer** **is** mediocre. But still solid enough to **put** it **head** and **shoulders** above any **macro**.

NOUN, **VERB**

- Relationships between nouns Dependency Tree
- Here: not very informative

Adjectives

An **OK** **lager**. **Light**, **crisp**, but **nothing** **special**. **Stacked** against the **great** **pilsners** of the **world** or **similar** **offerings**, this **beer** **is** **mediocre**. But still **solid** enough to **put** it **head** and **shoulders** above any **macro**.

NOUN, **VERB**, **ADJ**

- What is the opinion? OK, solid, (nothing) special, mediocre
- How is something? Crisp
- Stats for all Adjectives

Text Filtering: Noun Chunks

(An, OK, lager), (Light), (the, great pilsners)

(the, world), (similar, offering), (this, beer)

(it), (any, macro)

- Parse dependencies of nouns
- Additional filtering for noun, adjective, verb combinations

Text Normalization: Lemmatization

(stacked, stack), (pilsners, pilsner)

(offerings, offering), (is, be)

(it, -PRON-), (shoulders, shoulder)

- Map a word to its canonical form
- Alternative to stemming

Text Normalization: Special Cases

- Dates
  - Canonical form: 22.11.2017 → 11/22/2017
  - Relative to absolute: Yesterday → 11/22/2017
- Abbreviations
  - Common: United States of America → USA
  - Specific: ordinary least squares (OLS)
- Numbers and units
  - Hundred → 100
  - $100 → 100_dollar → 10000_cent

What is the most important aspect of the domain?

Word Sense Disambiguation

 or 

- Dictionary approach by Lesk (1986)
  - Assumption: common topic for words in same neighborhood
  - Choose sense with largest number of counts
- WordNet (Princeton)
  - Groups words into sets of synonyms
  - Demo
- Supervised methods
  - Rely on manually labeled training data
  - New words are problematic

Named Entity Recognition (NER)

- Persons, locations, organizations
- Stanford NER
  - Linear chain conditional random field sequence model
- DBpedia Spotlight
  - Entity recognition with DBpedia
  - Demo
- DBpedia
  - Structured content from Wikipedia
  - Started at FU Berlin and U Leipzig

NLP Toolkits

- NLTK, originated at U of Pennsylvania
- TextBlob, simplified text processing
- SpaCy, industrial-strength NLP

- OpenNLP, Apache
- CoreNLP, Stanford

For Moneybags

- Data
  - RavenPack
  - Sifter
- NLP, categorization, sentiment
  - Google Cloud Natural Language
  - Watson Discovery
- Paid crowd work platforms
  - Amazon's Mechanical Turk
  - CrowdFlower

Topic Modeling

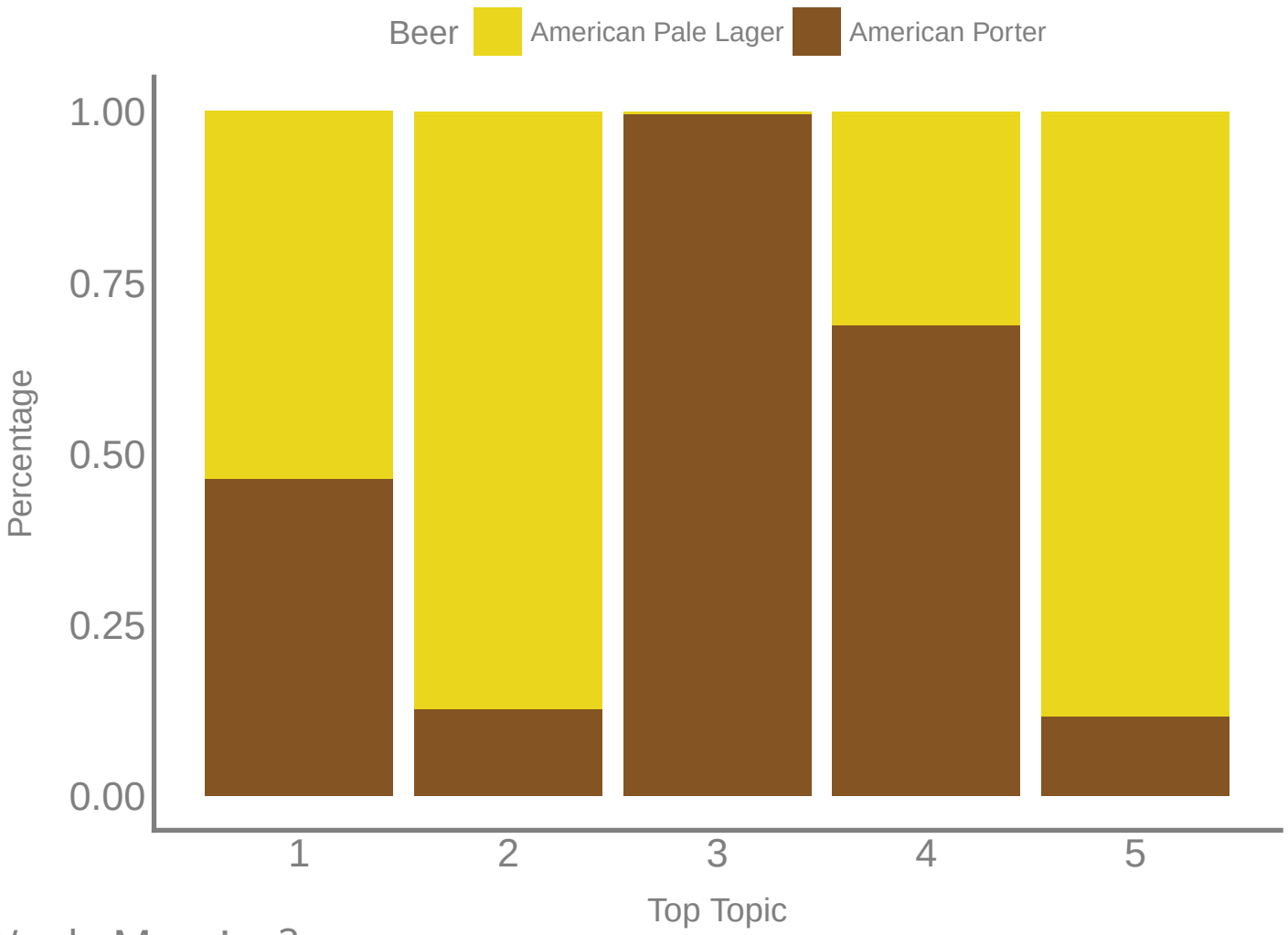
- Group similar documents
- Dimension reduction for documents
- Latent Semantic Analysis (LSA)
  - Singular value decomposition (SVD)
- Blei et al (2003): Latent Dirichlet Allocation (LDA)
  - Generative statistical model
  - Document as mixture of topics

Top Terms in Topics

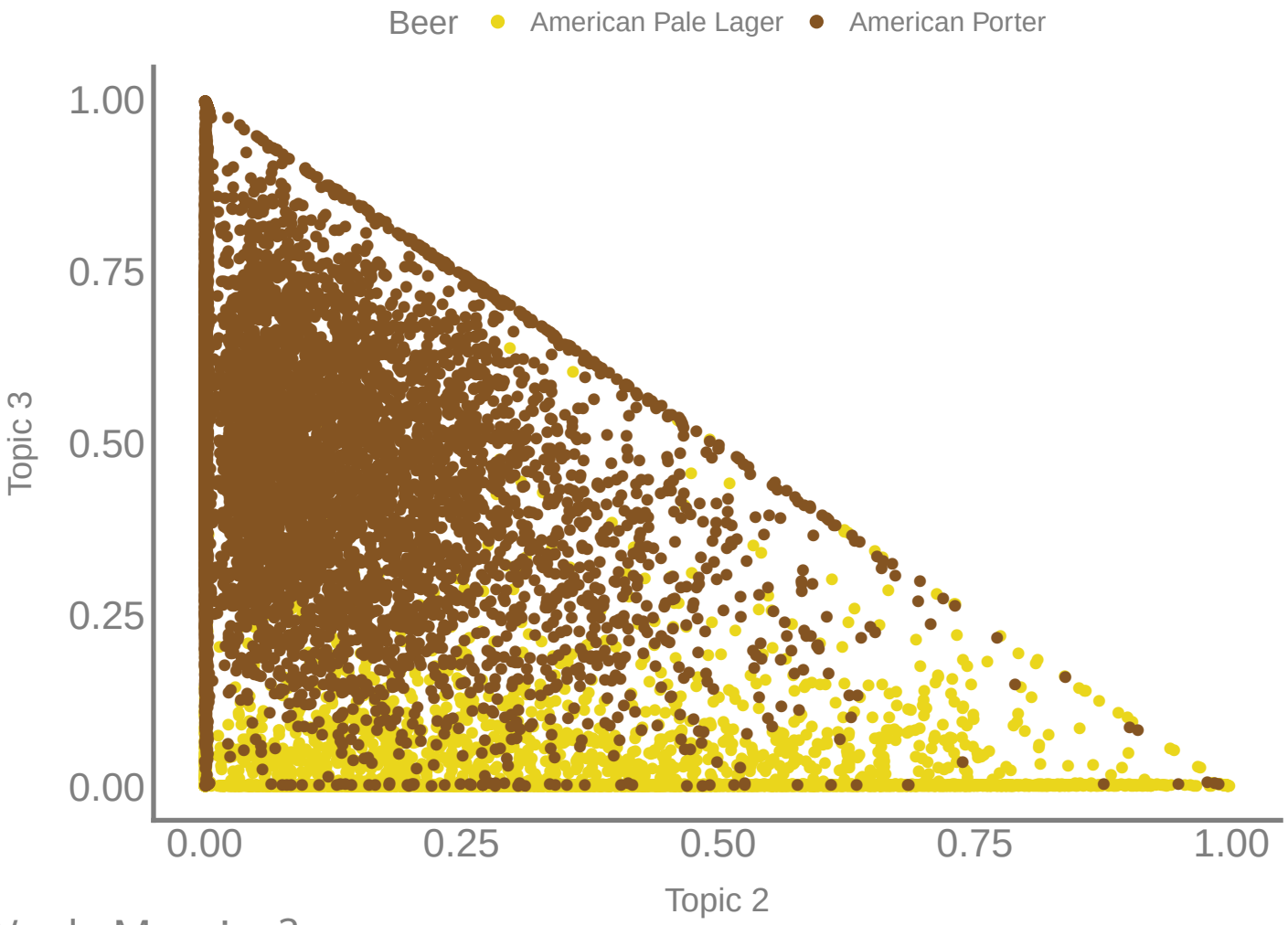
Topic	t1	t2	t3	t4
1	light	malt	nice	s
2	lager	butter	beer	taste
3	chocolate	coffee	dark	porter
4	-pron-	beer	pumpkin	brew
5	hop	light	malt	citrus

- Topic 2: light beers
- Topic 3: dark beers

Top Topic


Beer
American Pale Lager
American Porter
1.00
0.75
0.50
0.25
0.00
Percentage
1
2
3
4
5
Top Topic

Topic Values



Discrimination Possible?

- Classify style by topic value
- Discrimination rule

Lager if Topic 2 > Topic 3

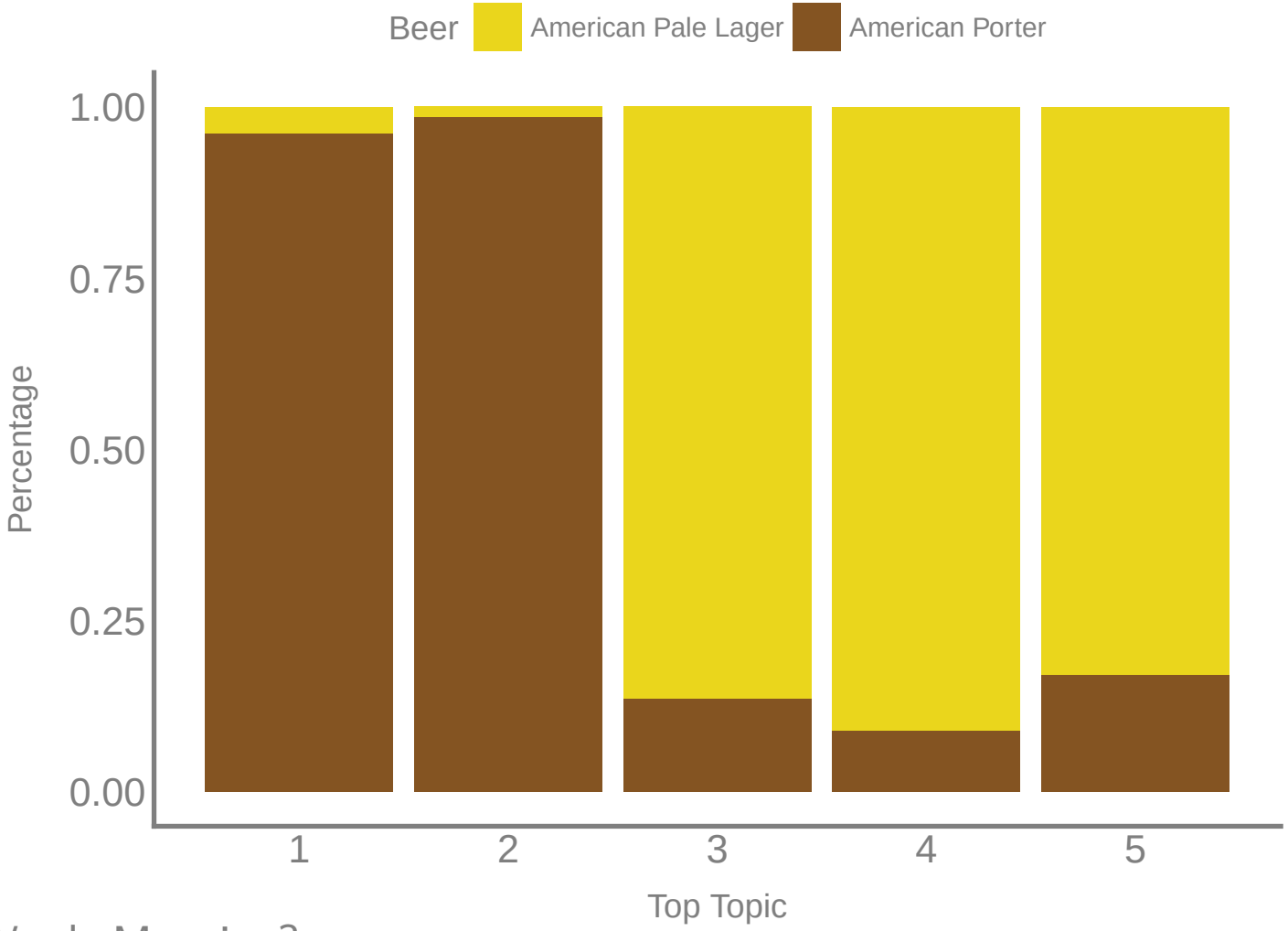
True \ Pred	Porter	Lager
Porter	10,559	616
Lager	859	5,991

Top Terms in Topics - 2-grams

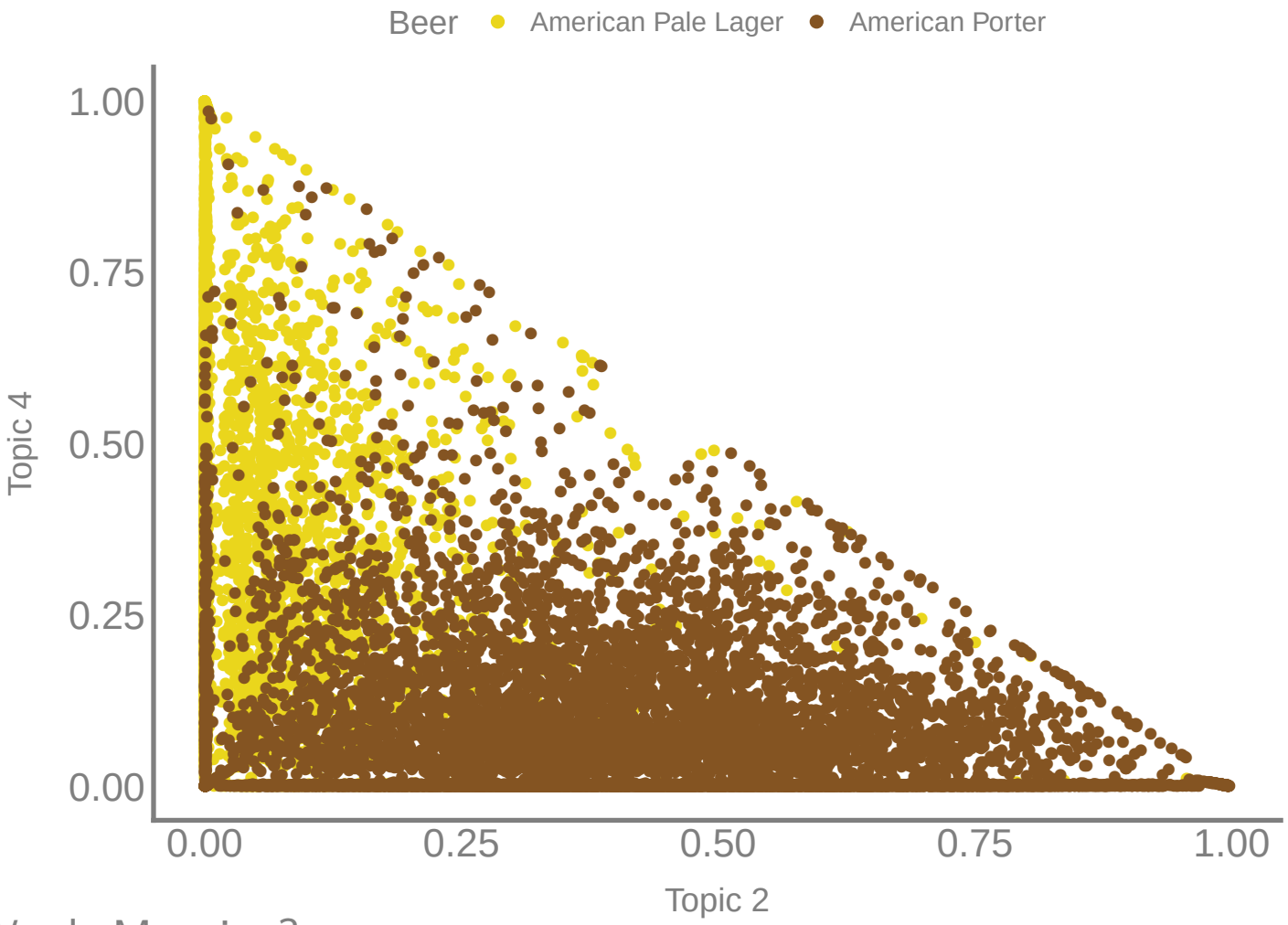
Topic	t1	t2	t3
1	tan head	dark fruit	dark brown
2	dark chocolate	roasted malt	peanut butter
3	white head	light body	pale lager
4	white head	floral hop	earthy hop
5	pint glass	good beer	half

- Topic 1: dark beers
- Topic 3: light beers
- Topic 2, 4: taste and color
- Topic 5: general

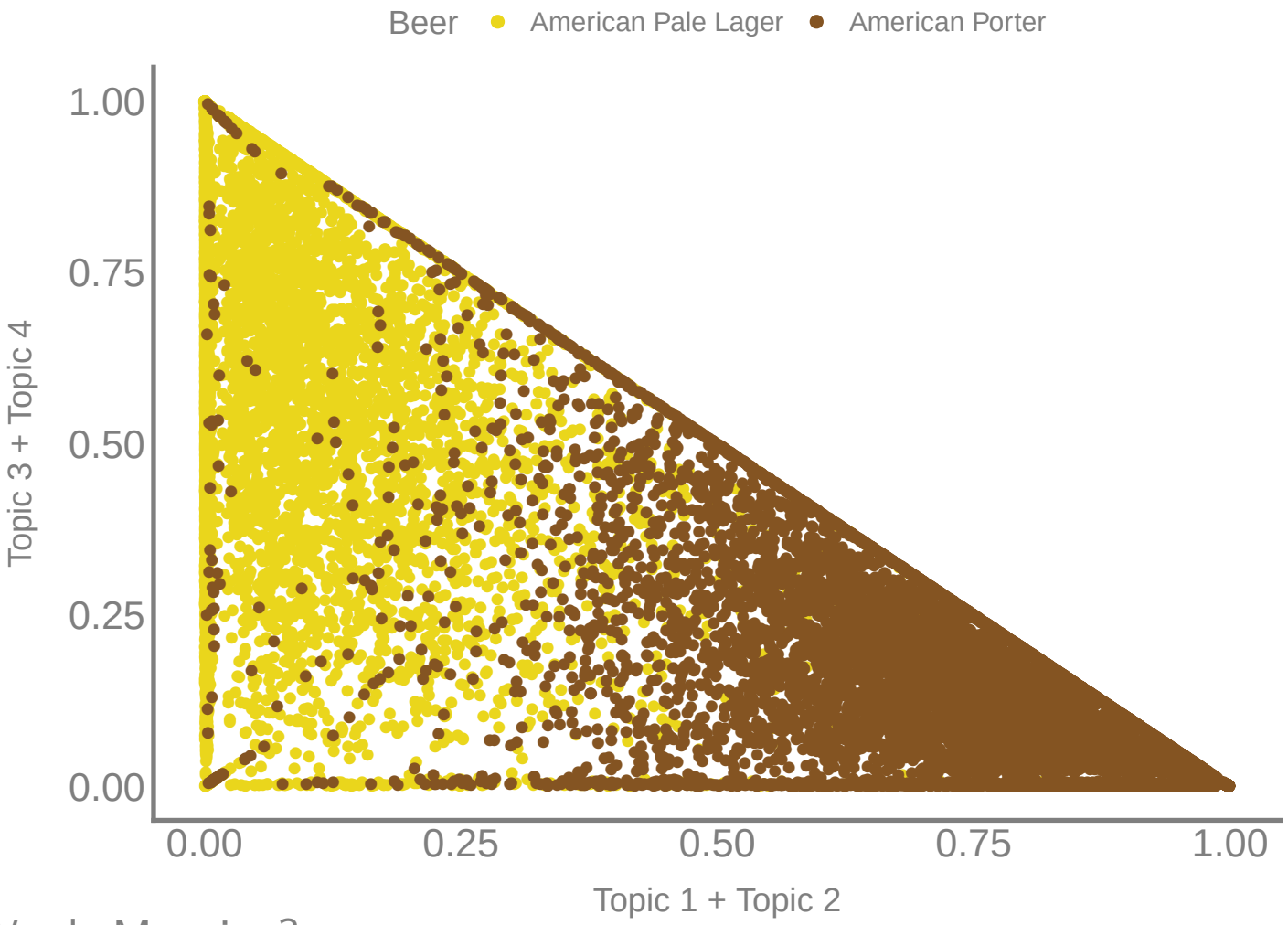
Top Topic - 2-grams



Topic Values - 2-grams



Topic Values - 2-grams ctd



Discrimination Possible? 2-grams

- Classify style by topic value
- Discrimination rule

Lager if Topic 3 + Topic 4 > Topic 2 + Topic 1

True \ Pred	Lager	Porter
Lager	10,723	452
Porter	616	6,234

Drawbacks of LDA

- Fixed number of topics
- Non-hierarchical
- Static
- Number of documents and length is crucial

Sentiment Classification

- Texts not always factual
- Sentiment, polarity, subjectivity, opinion, attitude
- Can we
  - classify the rating of a product review?
  - use news to predict stock movements?
  - predict election results based on social media?
- Approaches
  - Lexicon based
  - Supervised learning

Setup

- 2000 reviews with highest and lowest ratings, respectively
- Language Processing
  - Simple negation handling (not good: good_not)
  - Lemmatization: keep nouns, adjectives and verbs
  - Noun Chunks, 1-grams and 2-grams
- Dimension Reduction
  - 46,208 unique terms
  - LDA with 100 Topics
- Modeling
  - Support Vector Machines
  - Kernels: linear, polynomial, sigmoid, radial basis function

Evaluation

- Accuracy training set: 0.857
- Accuracy validation set: 0.871
- Confusion matrix validation set:

True \ Pred	Negative	Positive
Negative	842	139
Positive	116	883

Lexical Projection with BL

- Beats random classification
- Not surprisingly, worse than supervised approach
- Accuracy: 0.603
- Confusion matrix validation set:

True \ Pred	Negative	Positive
Negative	1,023	1,977
Positive	403	2,597

Conclusion

- Text data is noisy
- Basic text mining setup is "easy"
  - Usually sufficient for reviews
- Gets harder with increasing complexity
- Bottleneck: labeled training data

Texts Words Meaning?

Elisabeth Bommes

mbr targeting / Ströer Labs
and
Ladislaus von Bortkiewicz Chair of Statistics
Humboldt-Universität zu Berlin
http://lvb.wiwi.hu-berlin.de

Bibliography

Gantz, J and Reinsel, D
The digital universe in 2020
IDC iView: IDC Analyze the future, 2012

Wissler, L. and Almashraee, M. and Díaz, D. M. and Paschke, A.
The Gold Standard in Corpus Annotation
IEEE GSC, 2014

Kristianto, G Y, Ngiem, M Q, Matsubayashi, Y and Aizawa, A
Extracting definitions of mathematical expressions in scientific papers
Proc. 26th JSAI, 2012

Lesk, M
Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone
SIGDOC '86, 1986

Blei, D M and Ng, Andrew Y and Jordan, M I
Latent Dirichlet Allocation
J of Machine Learning Research, 2004

Appendix

Google's robots.txt

```
User-agent: *
Disallow: /search
Allow: /search/about
Allow: /search/howsearchworks
Disallow: /sdch
Disallow: /groups
Disallow: /index.html?
Disallow: /?
Allow: /?hl=
Disallow: /?hl=*&
Allow: /?hl=*&gws_rd=ssl$
Disallow: /?hl=*&*&gws_rd=ssl
Allow: /?gws_rd=ssl$
Allow: /?pt1=true$
Disallow: /imgres
Disallow: /u/
Disallow: /preferences
Disallow: /setprefs
Disallow: /default
Disallow: /m?
Disallow: /m/
Allow:    /m/finance
```

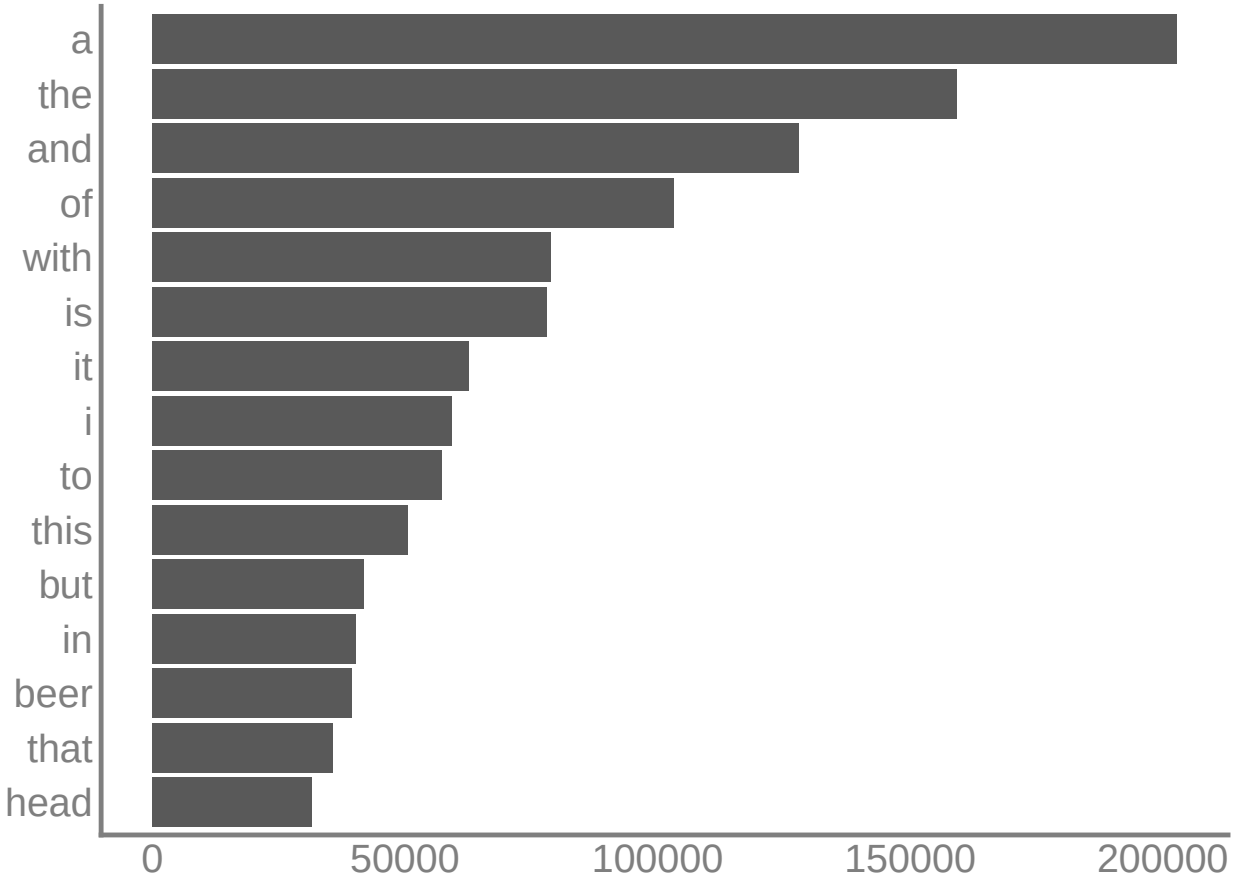
Back

Grammar based Dependency Tree

```
                          Stacked
                             |
              ________________|_____________________________
            against                                         |
               |                                            |
            pilsners                                        |
               |_______                                     |
  ____________|       |                                    |
 |     |              of                                   |
 |     |              |                                    |
 |     |            world                                 is
 |     |       _______|_______________              ______|___
 |     |      |       |            offerings        |      |   beer
 |     |      |       |        ________|______      |      |    |
the great   the      or   similar            ,  mediocre  .   this
```

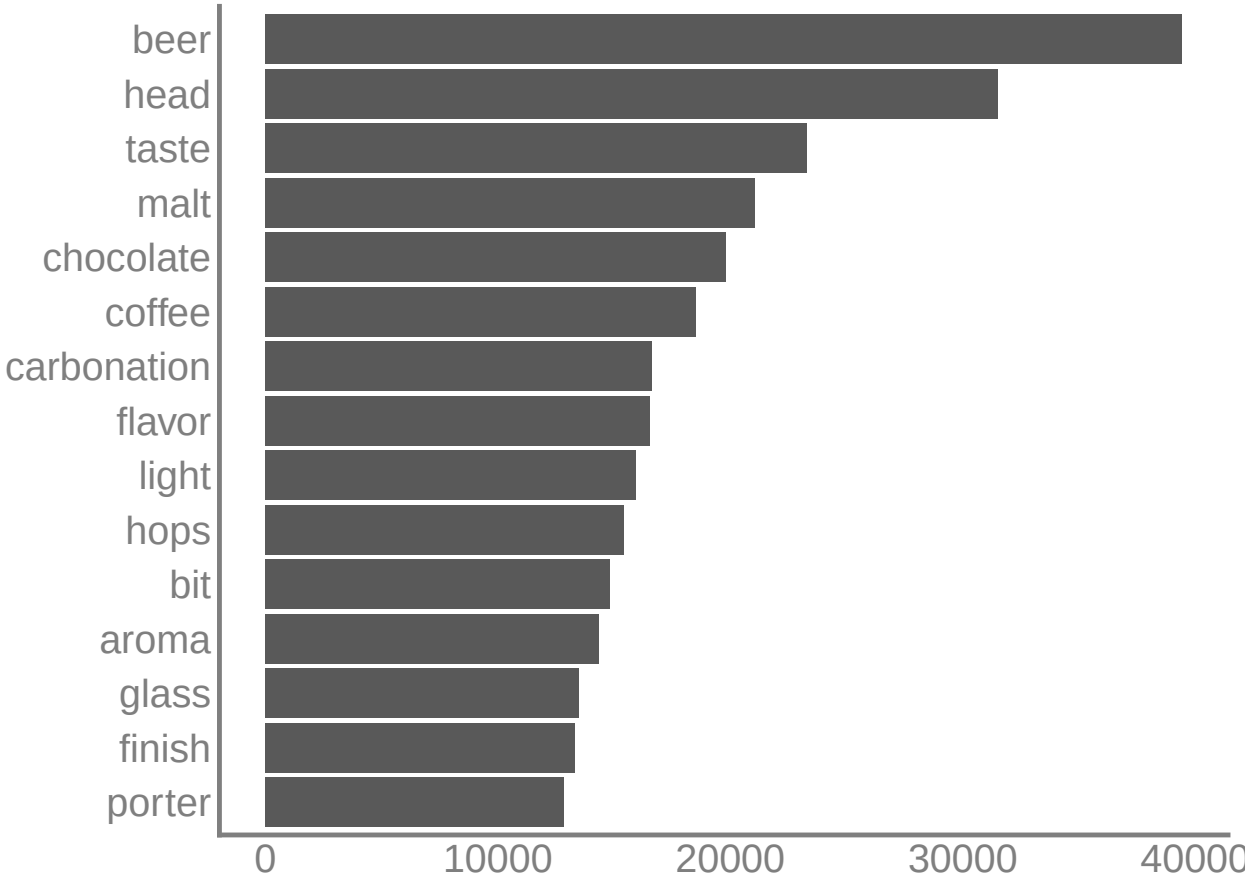
Back

Summary - All Words



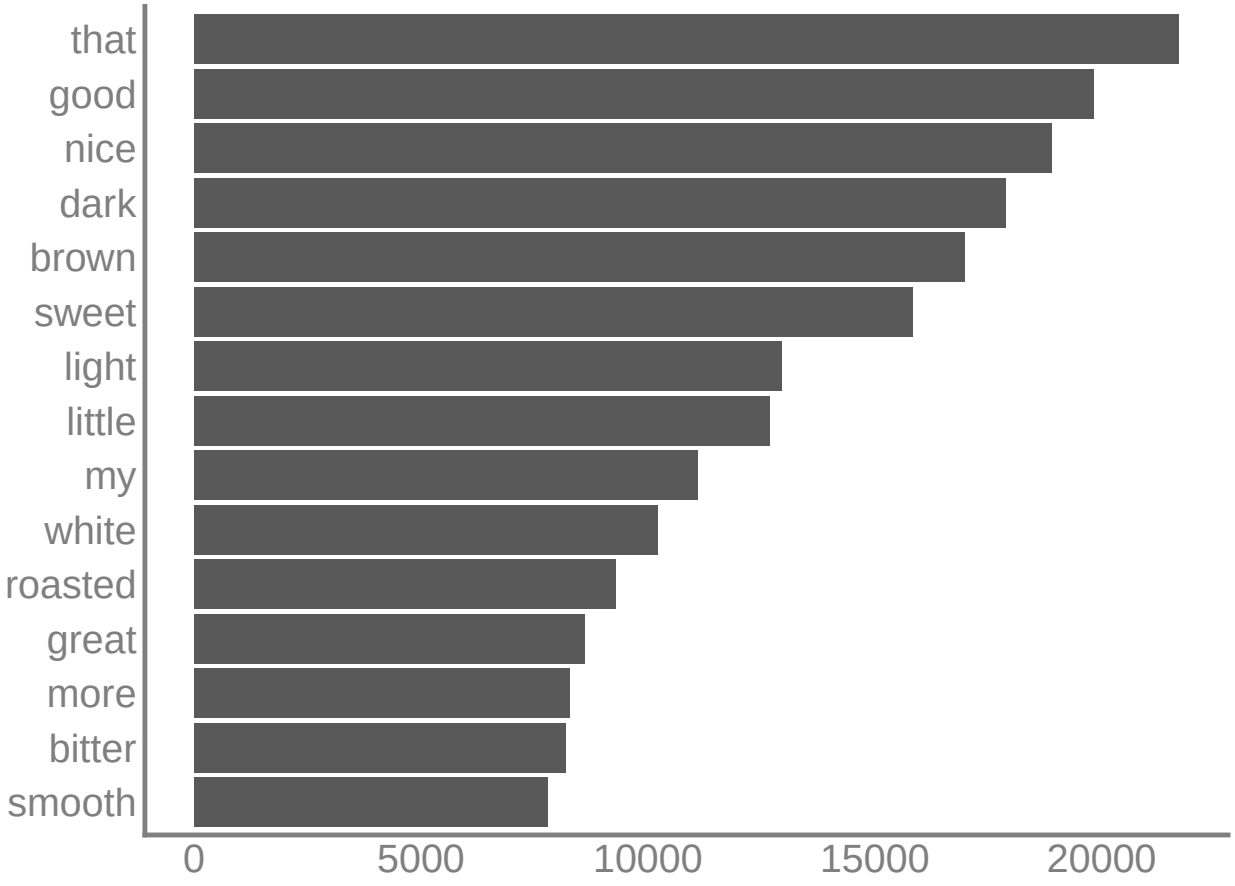
Back

Summary - All Nouns



Back

Summary - All Adjectives



Back