Text mining

DR. UROS GODNOV

Structured Data

0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628

Unstructured Data

Structured Data

0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628

Unstructured Data

Why text mining?

- 85-90 percent of all corporate data is in some kind of unstructured form (e.g., text)
- Unstructured corporate data is doubling in size every 18 months.

Why text mining?

- Text Mining Concepts Benefits of text mining are obvious especially in text-rich data environments e.g.:
- law (court orders)
- academic research (research articles)
- finance (quarterly reports)
- medicine (discharge summaries)
- marketing (customer comments)
- Electronic communication records (e.g., Email) Spam filtering Email prioritization and categorization Automatic response generation

Concepts

- Bag of words concepts
- NLP (natural language processing)

Concepts

- Bag of words concepts
  - mostly used tehnique
  - every word is independet (mostly); there are exceptions (e.g. Naive Bayes)
  - stemming (tourist, tourists and tourism may be the same word)
  - Sentiment analysis (different lexicons)
  - Entities extraction
  - Topics identification (e.g. LDA algorithm)
- NLP (natural language processing)
  - Uses dictionaries to learn (e.g. Stanford NLP)
  - a subfield of artificial intelligence and computational linguistics. the study of "understanding" the natural human language

Sentiment analysis

- Lexicons
- Simple→advanced
- Simple (e.g. Liu&Hu): -1 negative, 0 neutral, +1 positive
- Advanced (e.g. AFINN): -5<->5
- Different emotions (e.g. NRC)

NO STEMMING!

Topic analysis

- Key words
- LDA

STEMMING!

Topic analysis

- Key words
- LDA
  - Document consist of topics
  - Topic consist of words

- I like to eat broccoli and bananas.
- I ate a banana and spinach smoothie for breakfast.
- Chinchillas and kittens are cute.
- My sister adopted a kitten yesterday.
- Look at this cute hamster munching on a piece of broccoli.

LDA

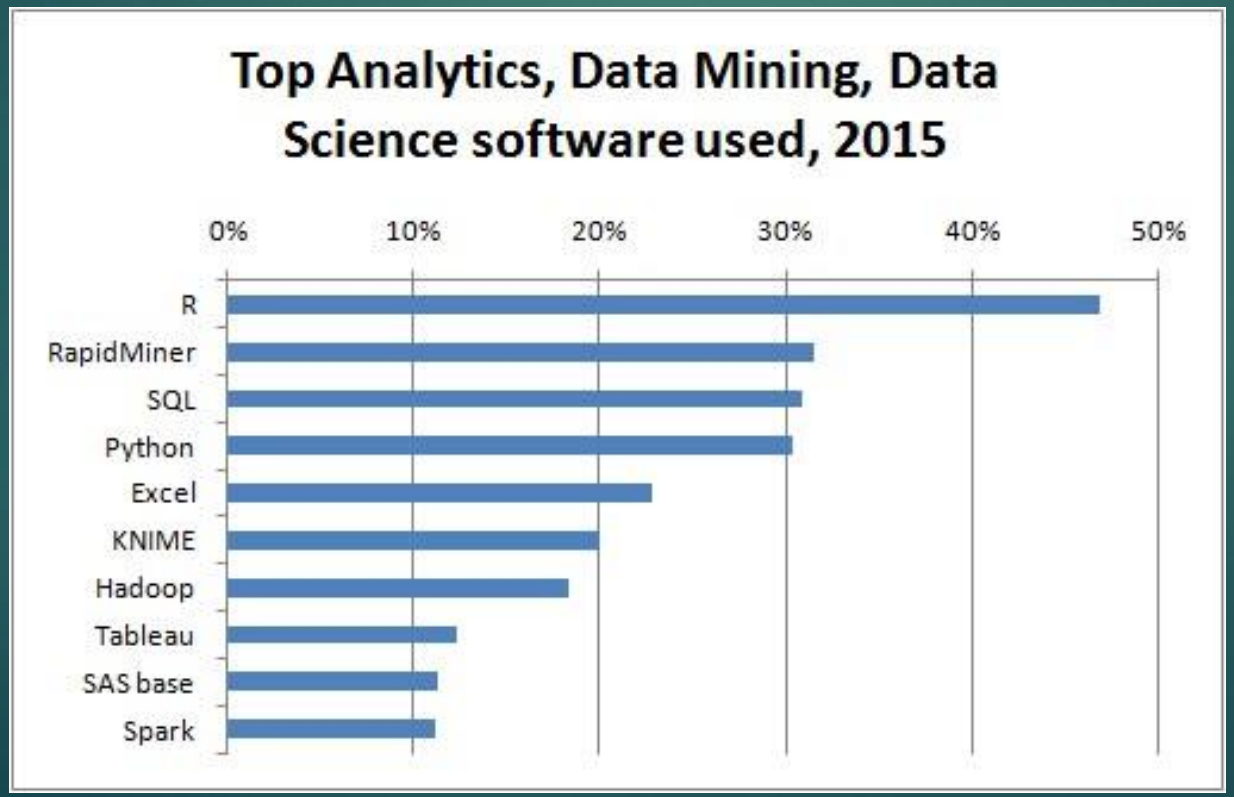
- **Sentences 1 and 2**: 100% Topic A
- **Sentences 3 and 4**: 100% Topic B
- **Sentence 5**: 60% Topic A, 40% Topic B
- **Topic A**: 30% broccoli, 15% bananas, 10% breakfast, 10% munching, … (at which point, you could interpret topic A to be about food)
- **Topic B**: 20% chinchillas, 20% kittens, 20% cute, 15% hamster, … (at which point, you could interpret topic B to be about cute animals)

Tools

- Open source tools
  - R
  - Rapidminer
- Commercial tools
  - SPSS (text mining model)
  - Semantria

R vs. Rapidminer

- R → programming oriented / Rapidminer→ GUI
- R – more flexible and fastest growing statistical software in the world



Terms associated with text mining

- Word frequency
- Stemming
- Term-document matrix / document-term matrix:
  - Removing stop words
  - Removing punctuations
  - Removing numbers
  - Removing whitespace

Why Rapidminer and not R?

Rapidminer is easier for you!

Rapidminer

- Which extensions do we need for text mining?
- Wordnet
- Textprocessing
- Aylien (make sure you are registered and got API key):
  - API
  - Commercial/free 1000 calls per day
  - Sentiment analysis

Configure Local repository

- Create Local repository on a friendlier place on disk
- Save process to repository

Identifying frequent words in phrases

Importing demo data

- Import TA.csv into local repository
- Transform date column
- Save data as demodata

Creating process

- Add following tasks:
- Retrieve
- Select attributes
- Nominal to text
- Process documents from data:
  - Tokenize
  - Transform cases
  - Filter tokens by length
  - Filter stopwords
  - Generate n-grams
- Wordlist to data
- Write Excel (thr)

Custom stoplist

- Adding custom stoplist:
- Filter stoplist(dictionary)
  - One word per line
  - First row doesn‘t count

Export result into excel and further analysis

- Convert wordlist to
- Write to excel
- Analyze results with pivot table in Excel

Identifying association rules

Create document term matrix and rules creation

- Document term matrix=create word vector
- Vector creation→binary occurences (0/1)
- Converting 0/1 to binominal
- FP growth
- Create association rules

FP Growth

- Calculates frequent itemsets (e.g. room-balcony)
- Find min number of itemsets
- Max number of retries
- Min support

Create association rules

- If→then rules
- Support
- Confidence

Sentiment analysis

Sentiment analysis – Part 1

- Wordnet:
  - Wordnet dictionary
  - Princeton university
  - 5000 words
  - **It has to be wordnet 3.0!**

Sentiment analysis – Part 2

- Custom extensions, e.g.: Aylien
  - Free plan (1000 API cals/day)
  - Sentiment analysis
  - Entity recognition
  - Hashtag recommender

Named entity recognition

Microsoft azure machine learning

- NLP
- organization names
- personal names
- locations
- English sentences
- Source (store in csv)
- convert result in csv and download
- further analysis in Excel

Documents similarity – preparing data

- How are documents similar (e.g. fake reviews)
- It uses TF-IDF:
  - Term frequency in a document: 0.3
  - Documents count: 500
  - Count of documents containig this word+1: 11
  - Inverse document frequency: 500/11=45.45
  - Log of inverse document frequency=1.66
  - Product:0.3*1.66=0.467 (relative importance of your word/phrase)

Documents similarity – calculating similarity

- Different measures of similarity (e.g. Jaccard, Jaro-Winkler, cosine similarity)
- Calculating angle between terms (smaller more similar)