2012
Z. Szlávik

Text Mining

Where's the interesting content in this slide?!
And what is it?

Finish the sentence: *"The three most frequently mentioned celebrities in the Netherlands are..."*

Lagos, Nigeria.
Attention: The President/CEO
Dear Sir,
Confidential Business Proposal
Having consulted with my colleagues and based on the information gathered from the Nigerian Chambers Of Commerce And Industry, I have the privilege to request for your assistance to transfer the sum of $47,500,000.00 (forty seven million, five hundred thousand United States dollars) into your accounts. The above s...

What's missing? ?.I.?.L.E.?.?.A.N.N.A.?.?.?.V.

Task A: continue the list
Task B: give a name to the group

Robin van Persie
Dirk Kuyt
Nigel de Jong
Tim Krul
Roland Bergkamp
Ronnie Stam
Marvin Emnes
Pim Balkestein

B: Dutch footballers playing in England

Source: http://eurorivals.net/abroad/dutch-players-in-england.html

Unstructured text?

- Please fill in the handout *(task 2)*
- Is text unstructured then?
  - **YES** – in the sense that it's not in rows and columns
  - **NO** – it has logical structure, lexical structure, etc.
  - **KIND OF** – XML document with free text in the nodes...
- Nevertheless, we can mine text, create, ask and answer questions based on text

Text mining – some definitions

- "The discovery of knowledge from database sources containing free text is called text mining." *(Kroeze et al., 2003)*
- "Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources." *(Hearst, 2003)*
- "A collection of methods used to find patterns and create intelligence from unstructured data" *(Francis, 2006)*

Text mining vs. Information retrieval

- **Text mining** =
  Data mining  Information retrieval

- **IR**: finding a needle in a (textual) haystack
- **TM**:
  - Finding something **useful** in the (textual) haystack that **solves your problem**

Further relations

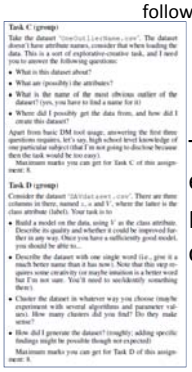
adapted from Kroeze et al. (2003) – see paper in BB

	Non-novel investigation *Retrieve*	Semi-novel investigation *Discover*	Novel investigation *Create*
Numeric data	Database queries	Standard DM	Intelligent DM
Textual data	IR	Standard text mining	Intelligent TM

- TM task examples:
  - Document clustering
  - Document classification *(e.g. spam/not spam)*
  - Summarisation
  - "Trending topics" detection
  - Citation analysis
  - Etc.

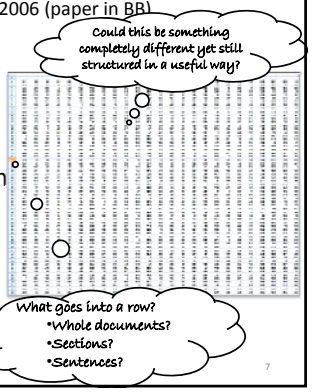
Unstructured2structured

following Francis, 2006 (paper in BB)



Term extraction

Feature creation



7

Example data
(Francis, 2006)

Textual attribute

Injury description
Broken ankle and sprained wrist
Foot contusion
Unknown
Mouth and knee
Head, arm lacerations
Lower back and legs
Back strain
Knee

Often the textual content is much longer than this

Often this is your only attribute to start with

Sometimes you only have one long piece of text

The following are synonyms: case, instance, example, record, ***document***

8

Term extraction

- Text is parsed into single words...
  - Or (proper) names, e.g. Lionel Messi
  - Or bigrams, trigrams, etc.
  - Or urls are identified
  - Etc.
- Basic string manipulation
  - Find delimiters, remove white space, convert all to lower case (?), I'll → I will, correct simple typoes, etc.

9

Matrix representation

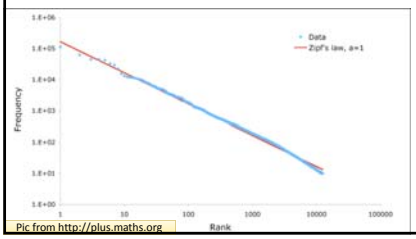
Injury desc.	Broken	Ankle	And	Sprained	Wrist	Foot	Contu-sion	Unknown	Neck	Back	Strain
Broken ankle and sprained wrist	1	1	1	1	1	0	0	0	0	0	0
Foot contusion	0	0	0	0	0	1	1	0	0	0	0
Unknown	0	0	0	0	0	0	0	1	0	0	0
Neck and back strain	0	0	1	0	0	0	0	0	1	1	1

- Notice:
  - The matrix is sparse *(efficient storage)*
  - It can be very big *(storage, access)*
  - There are only 1-s and 0-s *(anything else?)*
  - You don't have a word order *(is it a big problem?)*

10

Word frequency distribution

- Low number of words with very high occurrence
  - Stopwords
- High number of words with very low occurrence
  - Do you want to keep them?



Pic from http://plus.maths.org

Stopwords
A
The
And
While
Not (?)
Twilight (?)

Context trendy or stopword?

11

What else can you do?

- Stemming
  - Linguistic type: e.g. estimating → estimate
  - Rule-based stem finding: e.g. ... → estimat

- A number of linguistic operations
  - Disambiguation, synonym detection, etc.,etc.
  - That's not our focus now *(another course, another tools)*
- What we want now is to
  - Use standard DM methods
  - Reduce the dimensionality of the data

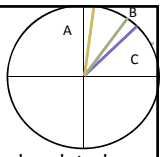
12

Dimensionality reduction

- Usual question: rows or columns?
- Rows *(docs)*:
  - Document clustering *(whole TM task on its own)*
  - Summarisation *(keep sentences classified as "summary-worthy" or "representative", "informative")*
- Columns *(words)*:
  - Methods from the previous two slides
  - Word clustering
    - To arrive at a managable data size
    - Keep "important" words
    - "Merge" statistically related words

13

Relatedness measures



- Similarity/dissimilarity
  - High/low number if two items are stongly related
- Relatedness measure examples
  - Euclidean distance (squared)
  - Manhattan distance (absolute)
  - Agreement (e.g. kappa)
  - Cosine similarity
  - Jaccard coeff., Dot coeff., etc.

distances
1.7320
3
0.4285
0.5773
f(A,B) = ?

A	B
0	0
0	1
1	1
1	0
0	0
0	1
1	1

14

Non-binary fields

(a `binary2numeric` operation)

- Present (1) – Not present (0) is often not enough
- Present how many times?
  - occurrences in a doc
- Normalise a doc vector
  - Length will be one *(cosine will make more sense)*
  - But depending on how length is measured, different normalisation can be applied
- You get a **TF** *(term frequency)* number

15

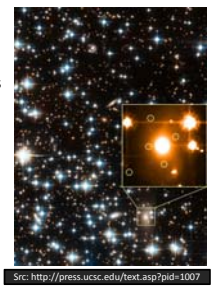
Consider a term *(word)* in relation to occurrences in other documents

- Df = number of documents in the collection in which a particular term occurs
- **Idf** (inverse document frequency)= log(D/Df)
  - Where D = number of docs in the collection
  - Why log? – because it works
- New number in cell = **tf** * **idf**
- What if we had i*sf* (*s* = sentence, section, etc)?

16

Clustering

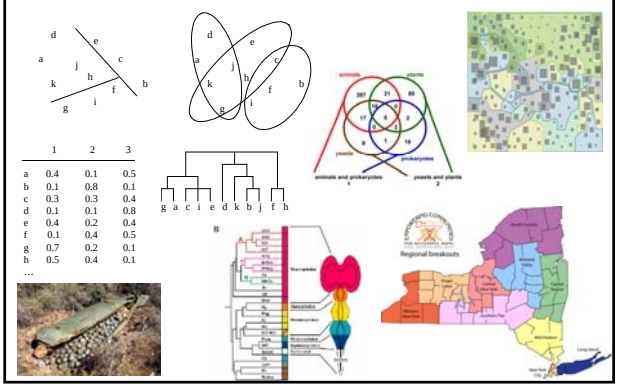
- **Given**
  - A collection of input vectors
    - From row or column data
    - With binary, nominal or numeric values
- **You choose**
  - A distance/similarity measure
  - Number of clusters (if needed)
  - Measure of cluster quality
  - An algorithm that forms clusters
- **This will give you**
  - Groups of similar vectors


Src: http://press.ucsc.edu/text.asp?pid=1007

- Note the similarity between clustering and instance-based learning

17

Types & (re)presentations



Document clusters example

Cluster ID	back	contusion	head	knee	strain	unknown	laceration
1	1.00	0.11	0.11	0.05	0.40	0.00	0.00
2	0.00	0.04	0.04	0.48	0.09	0.00	0.05
3	0.00	0.17	0.14	0.04	0.05	0.16	0.19

- Handout time!
- These were 3 clusters *(k=3)*
- How many do we need, really?
  - A fixed number you determine up front
  - Measure cluster 'goodness', and chose the best *k*
  - Start with high/low number of clusters, then stop when no significant change happens in the clusters' goodness
  - Check with how many clusters you get the best accuracy in a classification task *(if applicable)*

19

Clusters of words (*or documents*)

- How do you name a cluster?

Cluster ID	Back	Contusion	Head	Knee	Strain	Unknown	Laceration	Leg
1	0.000	0.000	0.000	0.095	0.000	0.277	0.000	0.000
2	0.022	1.000	0.261	0.239	0.000	0.000	0.022	0.087
3	0.000	0.000	0.162	0.054	0.000	0.000	1.000	0.135
4	1.000	0.000	0.000	0.043	1.000	0.000	0.000	0.000
5	0.000	0.000	0.065	0.258	0.065	0.000	0.000	0.032
6	0.681	0.021	0.447	0.043	0.000	0.000	0.000	0.000
7	0.034	0.000	0.034	0.103	0.483	0.000	0.000	0.655

'Strain' occurs in 48.3% of the documents of Cluster 7

20

Naming

- A name is
  - Most frequent word in a cluster
  - Most frequent *n* words in a cluster that are above *x* occurrence ratio
  - Most frequent words that are not frequent in other clusters
  - Some tf-idf like measure for naming?
  - Etc.
- But should I really care about naming?
  - **Yes**, if you want to gain insight into what clusters have been created, or display the labels somewhere, somehow
  - **No**, if you can just assign a new doc to an existing cluster and use only the cluster number in, say, a classification task

21

What's next

(a possible scenario)

- You replace the text of a record by its **Cluster ID**
  - Serious dimension reduction
  - Extra computations
- New record is assigned to whichever cluster it is closer to
  - Instance-based methods?
  - Many new records? → reclustering
- Now you might have several nominal/numeric attributes, plus this **Cluster ID**
- Train a classifier and, e.g., predict if the medical expenses are above $€10.000$ or not

22

A practical text mining example: humour detection

- What's funny?
- Why do we laugh?
- What's humour?
- There are some theories but there isn't a unified one
- Still, can we still identify or generate humour using a computer?

Humour detection as a classification problem

- Situation:
  - Here's a piece of text
  - Is it funny? (the machine has to tell)
- An example:
  - Detecting humourous one-liners
- An example of a one liner:
  - "We live in a society where pizza gets to your house before the police."

24

Dataset

- Short humorous sentences (one-liners)

- Titles of Reuters news articles
- Proverbs
- Sentences from the British National Corpus
- Statements from the Open Mind Common Sense Corpus

One line – one document

Features

Text2columns

- Stylistic features:
  - number of alliterations
  - number of antonym pairs
  - number of words with sexual connotation
- Content based features:
  - bag of words
  - n-grams (unigrams, bigrams, trigrams)

Marriage is like pi - natural, irrational, and very important.

26

Classification methods

- On stylistic features
  - Decision tree
    - Alliteration is most indicative of humour among the three features
- On content-based features
  - Naïve Bayes
  - Support Vector Machines
  - Comparable results when different datasets are used as negative examples

27

Content-based findings

- One-liners tend to:
  - refer to human related scenarios
  - display negative scenarios
  - refer to professional communities
  - refer to situations that display human weakness

28

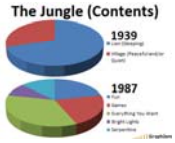
Today we had

- Introduction to text mining
- Short introduction to clustering
- A text classification example

29

Last slide today

- Questions?
- Videos?
- Lunch?



30