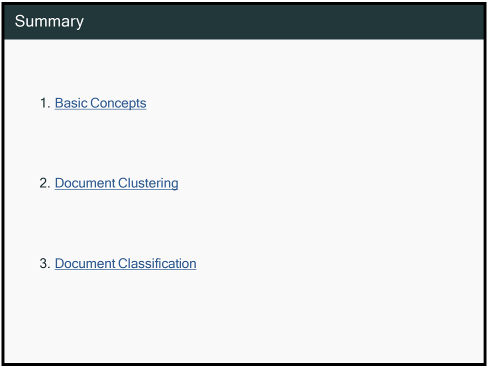
Data Mining II / Adv. Topics in Data Science

Text Mining

Álvaro Figueira

Rita Ribeiro

1



Summary

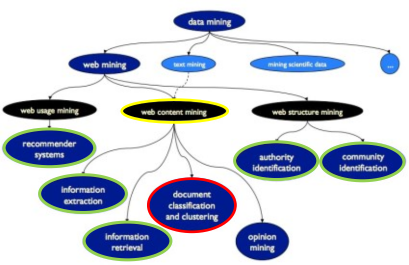
1. Basic Concepts
2. Document Clustering
3. Document Classification

2

Basic Concepts

3

Data Mining - a structured view



4

Information Retrieval and Text Mining

- Information Retrieval (IR)
  Tasks include: crawling, indexing documents, query processing, document ranking, relevance feedback, search and retrieval of documents.

- Text Mining
  Tasks include: document classification, document clustering, building an ontology, sentiment analysis, document summarization, keyword extraction, NER, co-reference resolution, text generation (among others).

- Think of it this way
  IR is like searching a library catalog to find relevant books, while text mining is like reading and analyzing those books to uncover hidden themes and connections.

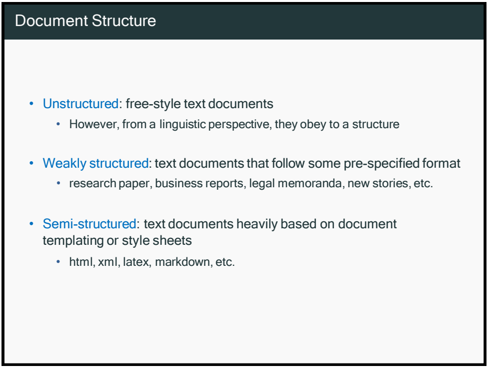
5

Corpus and Documents

- Corpus: a collection of documents; it can be static or dynamic.
  - Examples: PubMed, Reuters, Wikipedia, Jornal de Notícias, etc.
  - Several corpus with a common property form a Corpora

- Document: a unit of discrete text within a corpus.
  - Examples: a research paper, news story, business report, email, tweet, Facebook post, etc.

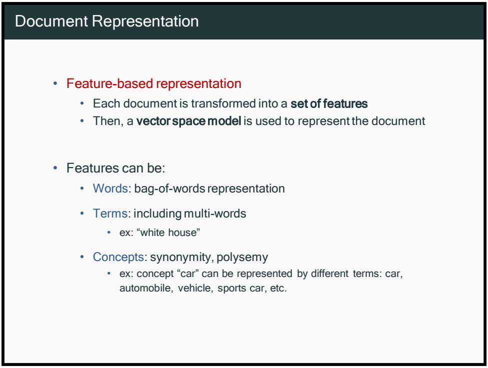
6



Document Structure

- Unstructured: free-style text documents
  - However, from a linguistic perspective, they obey to a structure
- Weakly structured: text documents that follow some pre-specified format
  - research paper, business reports, legal memoranda, new stories, etc.
- Semi-structured: text documents heavily based on document templating or style sheets
  - html, xml, latex, markdown, etc.

7



Document Representation

- Feature-based representation
  - Each document is transformed into a **set of features**
  - Then, a **vector space model** is used to represent the document
- Features can be:
  - Words: bag-of-words representation
  - Terms: including multi-words
    - ex: “white house”
  - Concepts: synonymity, polysemy
    - ex: concept “car” can be represented by different terms: car, automobile, vehicle, sports car, etc.

8

Document Representation (cont.)

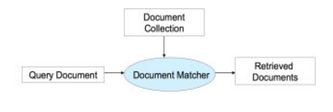
- **Problem:**
  "The curse of (high) dimensionality"
  - Structured representations of natural language documents →
    very large number of features.

    Example: one small Reuters collection of 15000 documents contains +25000 unique features (**word stems**)
  - Some algorithms do not deal very well with large numbers of features
    → We need to use **feature reduction techniques**.
  - Each document contains only a small number of all these potential features
    → **feature sparcity**.

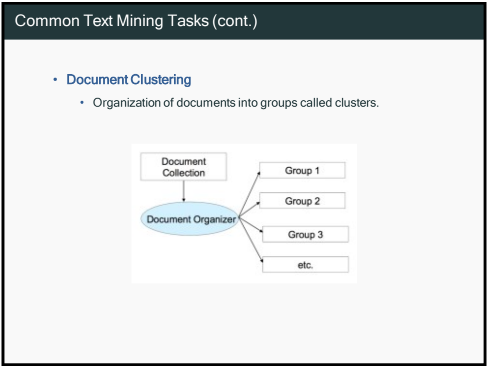
9

Common Text Mining Tasks

- **Information Retrieval**
  - Retrieval of documents in response to a query.



10



Common Text Mining Tasks (cont.)

- **Document Clustering**
  - Organization of documents into groups called clusters.

Document Collection

Document Organizer

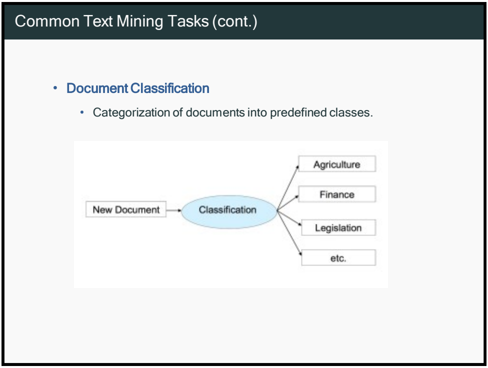
Group 1

Group 2

Group 3

etc.

11



Common Text Mining Tasks (cont.)

- **Document Classification**
  - Categorization of documents into predefined classes.

New Document

Classification

Agriculture

Finance

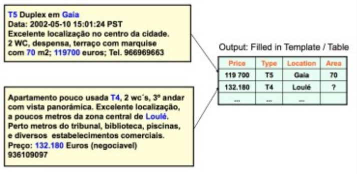
Legislation

etc.

12

Common Text Mining Tasks (cont.)

- Information Extraction
  - Identification of certain entities in the text, their extraction and representation in a pre-specified format (e.g. table).



T5 Duplex em Gaia
Data: 2002-05-10 15:01:24 PST
Excelente localização no centro da cidade.
2 WC, despensa, terraço com marquise
com 70 m2; 119700 euros; Tel. 966969663

Apartamento pouco usada T4, 2 wc´s, 3º andar
com vista panorámica. Excelente localização,
a poucos metros da zona central de Loulé.
Perto metros do tribunal, biblioteca, piscinas,
e diversos estabelecimentos comerciais.
Preço: 132.180 Euros (negociavel)
936109097

Output: Filled in Template / Table

Price	Type	Location	Area
119 700	T5	Gaia	70
132.180	T4	Loulé	?
...	...	...	

13

Advanced Text Mining Tasks

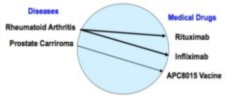
- Concept co-occurrence
- Identification of new or disappearing topics
- Summarization
- Keyword Extraction
- Sentiment Analysis and Opinion Mining

Examples in the next slides

14

Advanced Text Mining Tasks

- Concept co-occurrence
  - Quantification of co-occurrence
  - Association mining with terms or concepts in texts
  - Example: Disease - Medical Drug (based on BioWorld articles). A representation: circle graph where the width of the line indicates the strength of the connection



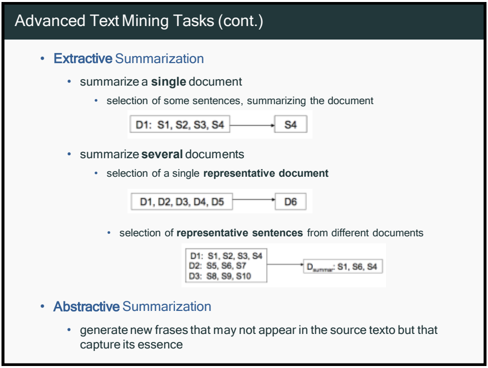
  - Example: President of The United States | POTUS | Biden | Joe Biden

15

Advanced Text Mining Tasks (cont.)

- Identification of **new topics** in the data
  - Did any news articles appear concerning a certain type of company?
    - e.g. a farmaceutical company
  - and a particular type of product?
    - e.g. a medical drug useful for treating lung cancer
- Identification of **disappearing topics** in the data
  - Example: Paris
    - Bataclan terrorist attack, fear, sadness, anger Vs. the city of love, joy
- Identification of a **period covered by a certain topic**

16



Advanced Text Mining Tasks (cont.)

- **Extractive** Summarization
  - summarize a **single** document
    - selection of some sentences, summarizing the document

D1: S1, S2, S3, S4 → S4

  - summarize **several** documents
    - selection of a single **representative document**

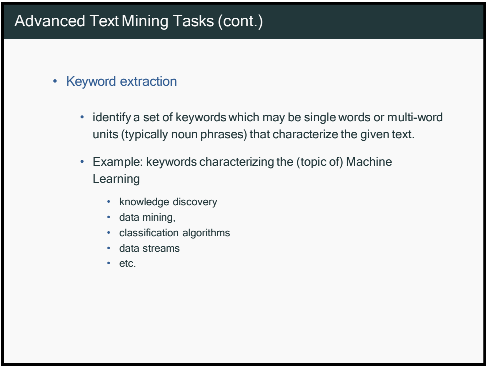
D1, D2, D3, D4, D5 → D6

    - selection of **representative sentences** from different documents

D1: S1, S2, S3, S4
D2: S5, S6, S7
D3: S8, S9, S10
→ D_{summar}: S1, S6, S4

- **Abstractive** Summarization
  - generate new frases that may not appear in the source texto but that capture its essence

17



Advanced Text Mining Tasks (cont.)

- Keyword extraction
  - identify a set of keywords which may be single words or multi-word units (typically noun phrases) that characterize the given text.
  - Example: keywords characterizing the (topic of) Machine Learning
    - knowledge discovery
    - data mining,
    - classification algorithms
    - data streams
    - etc.

18

Advanced Text Mining Tasks (cont.)

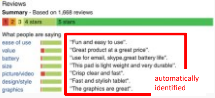
Sentiment Analysis / Opinion Mining

- Applications to reviews-related websites
  - use written film reviews
  - # positive words vs # negative words

*"The story of the movie is **good** but the acting of the actors is **awful**. Songs of the movie **hit** the chartbusters, youngsters **like** the songs very much"*

- Applications in business and government
  - discussions in forums about a product

Apple iPad Review:



Reviews

Summary - Based on 1,668 reviews

1 2 3 4 stars 5 stars

What people are saying

ease of use	"Fun and easy to use".
value	"Great product at a great price".
battery	"use for email, skype,great battery life".
size	"This pad is light weight and very durable".
picture/video	"Crisp clear and fast".
design/style	"Fast and stylish tablet".
graphics	"The graphics are great".

automatically identified

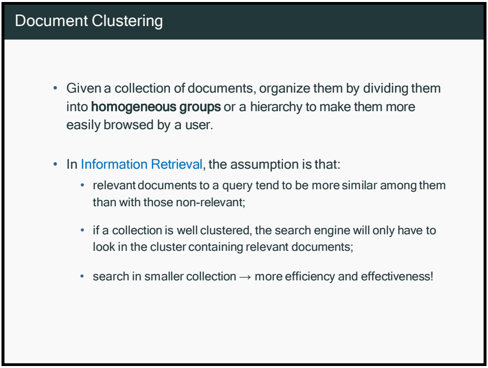
- Applications in business and government intelligence
  - opinions matter a great deal in politics
  - e.g., government-regulation proposals

"Government regulations are necessary to encourage businesses and consumers to rely more on renewable energy sources"

19

Document Clustering

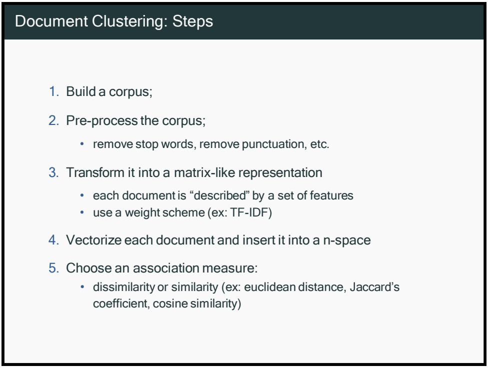
20



Document Clustering

- Given a collection of documents, organize them by dividing them into **homogeneous groups** or a hierarchy to make them more easily browsed by a user.
- In Information Retrieval, the assumption is that:
  - relevant documents to a query tend to be more similar among them than with those non-relevant;
  - if a collection is well clustered, the search engine will only have to look in the cluster containing relevant documents;
  - search in smaller collection → more efficiency and effectiveness!

21



Document Clustering: Steps

1. Build a corpus;
2. Pre-process the corpus;
   - remove stop words, remove punctuation, etc.
3. Transform it into a matrix-like representation
   - each document is “described” by a set of features
   - use a weight scheme (ex: TF-IDF)
4. Vectorize each document and insert it into a n-space
5. Choose an association measure:
   - dissimilarity or similarity (ex: euclidean distance, Jaccard’s coefficient, cosine similarity)

22

Document Clustering: Steps (cont.)

6. Choose a clustering algorithm:
   - partitional → decide the number of clusters (e.g., k-means)
   - hierarchical (agglomerative) → decide the pair of clusters to merge at each step
     - MIN (single link) proximity: shortest distance between two points that are in different clusters;
     - MAX (complete link) proximity: furthest distance between two points that are in different clusters;
     - Group Average Proximity: average distance between each two points that are in different clusters;
     - Ward's method: clusters are represented by its centroids; proximity between two clusters in terms of the increase in sum of the squared error that results from merging the two clusters; it minimizes within-group dispersion at each binary fusion.

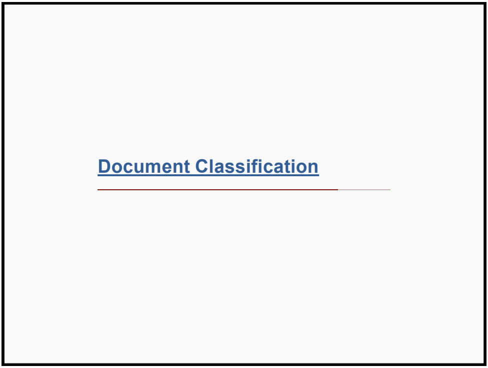
23

Document Clustering

- Characterization of clusters
  - Example technique: word clouds

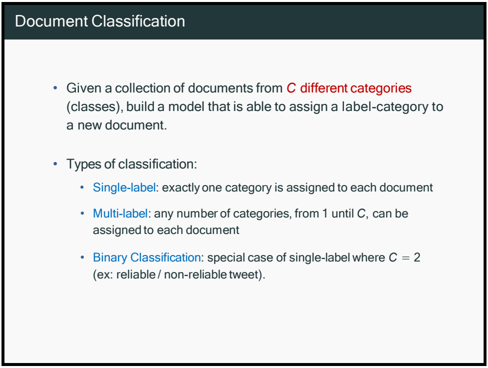
problems
lcsp
similarity
agreement
process
space
episode
novel
measure
sets
preference
efficient
automatic
partition-model
preference-data
properties
social-media
real
search
query
object
experiments
cascades
different
world
dags
method
important
multiscale
application
graphs
study
tool
temporal-intervals
graph
problem
community-search
many
social
communities
model
event
large
impact
mining
algorithm
distance
subspace
campaign
information
detection
visual
temporal
network
boolean
event-detection
technique
new
pattern
intervals
support
analysis
partition
order
measures
tensor
user
brain
real-world
experiment
range
roles
approach
factorizations
synthetic
social-roles
high
time
aspects
effective

24



Document Classification

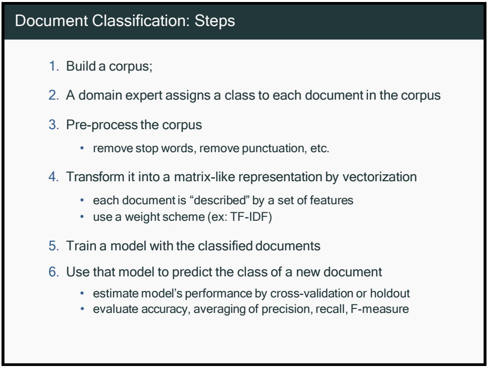
25



Document Classification

- Given a collection of documents from *C* different categories (classes), build a model that is able to assign a label-category to a new document.
- Types of classification:
  - Single-label: exactly one category is assigned to each document
  - Multi-label: any number of categories, from 1 until *C*, can be assigned to each document
  - Binary Classification: special case of single-label where $C = 2$ (ex: reliable / non-reliable tweet).

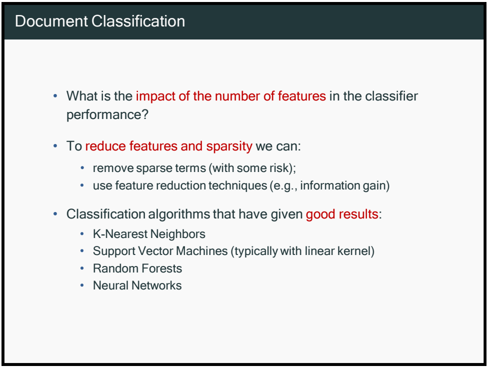
26



Document Classification: Steps

1. Build a corpus;
2. A domain expert assigns a class to each document in the corpus
3. Pre-process the corpus
   - remove stop words, remove punctuation, etc.
4. Transform it into a matrix-like representation by vectorization
   - each document is “described” by a set of features
   - use a weight scheme (ex: TF-IDF)
5. Train a model with the classified documents
6. Use that model to predict the class of a new document
   - estimate model’s performance by cross-validation or holdout
   - evaluate accuracy, averaging of precision, recall, F-measure

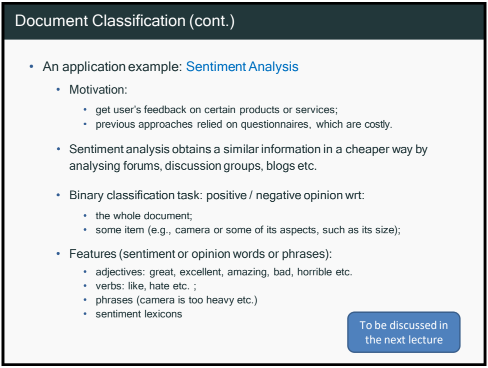
27



Document Classification

- What is the impact of the number of features in the classifier performance?
- To reduce features and sparsity we can:
  - remove sparse terms (with some risk);
  - use feature reduction techniques (e.g., information gain)
- Classification algorithms that have given good results:
  - K-Nearest Neighbors
  - Support Vector Machines (typically with linear kernel)
  - Random Forests
  - Neural Networks

28

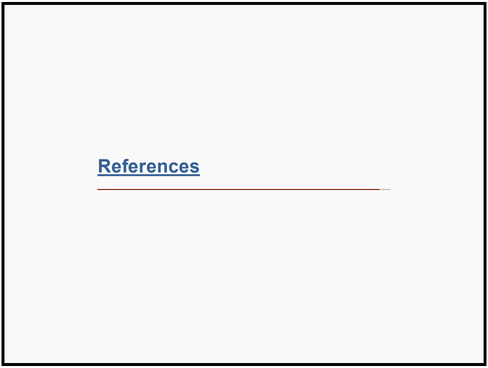


Document Classification (cont.)

- An application example: Sentiment Analysis
  - Motivation:
    - get user's feedback on certain products or services;
    - previous approaches relied on questionnaires, which are costly.
  - Sentiment analysis obtains a similar information in a cheaper way by analysing forums, discussion groups, blogs etc.
  - Binary classification task: positive / negative opinion wrt:
    - the whole document;
    - some item (e.g., camera or some of its aspects, such as its size);
  - Features (sentiment or opinion words or phrases):
    - adjectives: great, excellent, amazing, bad, horrible etc.
    - verbs: like, hate etc. ;
    - phrases (camera is too heavy etc.)
    - sentiment lexicons

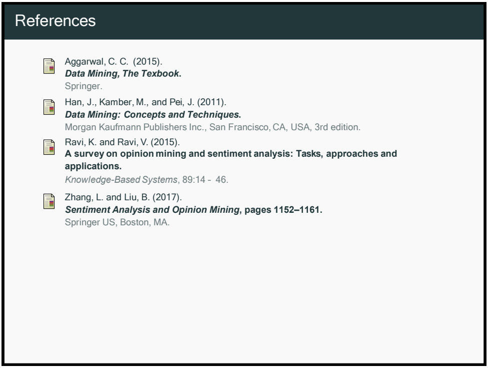
To be discussed in the next lecture

29



References

30



References

- Aggarwal, C. C. (2015).
  Data Mining, The Texbook.
  Springer.
- Han, J., Kamber, M., and Pei, J. (2011).
  Data Mining: Concepts and Techniques.
  Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.
- Ravi, K. and Ravi, V. (2015).
  A survey on opinion mining and sentiment analysis: Tasks, approaches and applications.
  Knowledge-Based Systems, 89:14 - 46.
- Zhang, L. and Liu, B. (2017).
  ***Sentiment Analysis and Opinion Mining*, pages 1152–1161.**
  Springer US, Boston, MA.

31