Section 7

Text mining

Text mining

Definition

Text mining is about extracting high level information from textual data.

A joint effort of *Machine Learning* and *Natural Language Processing* (NLP).

Example 1: sentiment on brands

Is people on Twitter talking good or bad about brand X?

Example 2: topics in letters

In this corpus of letters to/from the front in WW1, which are the topics covered?

Common tasks

- text categorization → CLASSIFICATION
- text clustering
- entity extraction
- sentiment analysis
- summarization
- ...

Example 3: relevance of citations

Find a way to quantify relevance of a citation from a scientific paper A to a scientific paper B?

Step 0

- Define the nature of the solution:
  - input
  - output
  - learning data (if any)
- Define a way to asses a solution

Step 0: is it easy?

	INPUT	OUTPUT	LEARNING	ASSESSMENT
1	A TWEET	{GOOD, BAD}	Y	• USE COLLECTED DATA • COMPARE W/ SURVEYS • MEASURE MONEY
2	CORPUS	SET OF SETS OF WORDS (TOPIC)		"MANUAL" READING (POSSIBLY SAMPLE)

1. **Q:** for "sentiment on brands"?
2. **Q:** for "topics in letters"?
3. **Q:** for "relevance of citations"?