TEXT MINING

L12. CONCLUSIONS

SUZAN VERBERNE 2021

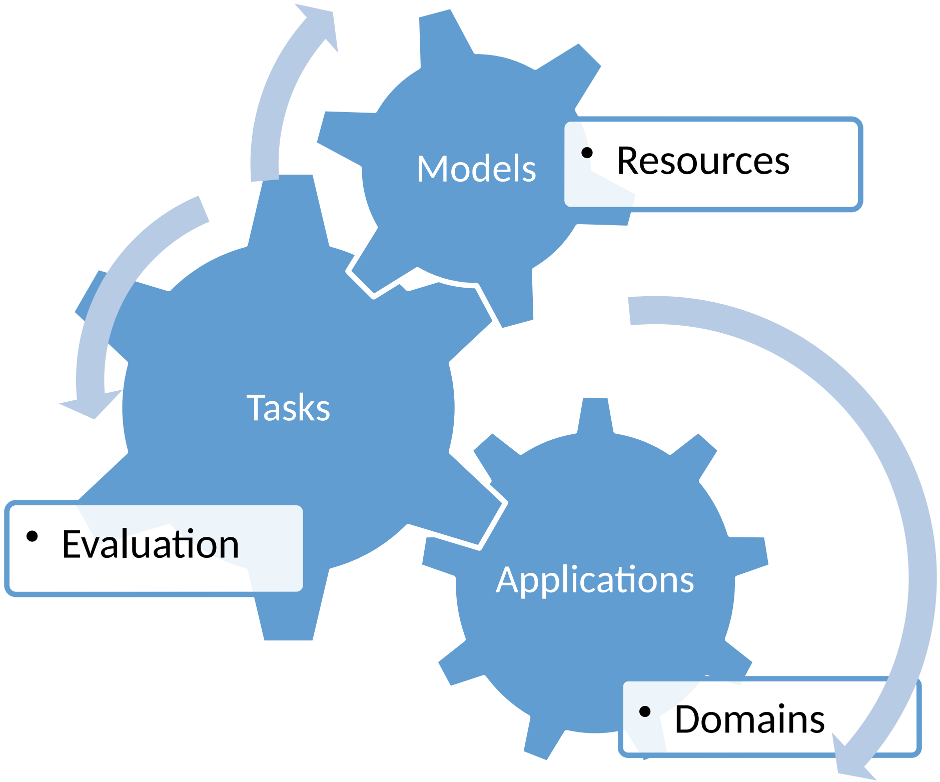
Universiteit Leiden

TODAY'S LECTURE

- Course summary
  - Tasks
    - and evaluation
  - Models
    - and resources
  - Applications
    - and domains
- Exercises to practice for the exam
- Schedule of tests (exam/assignment)

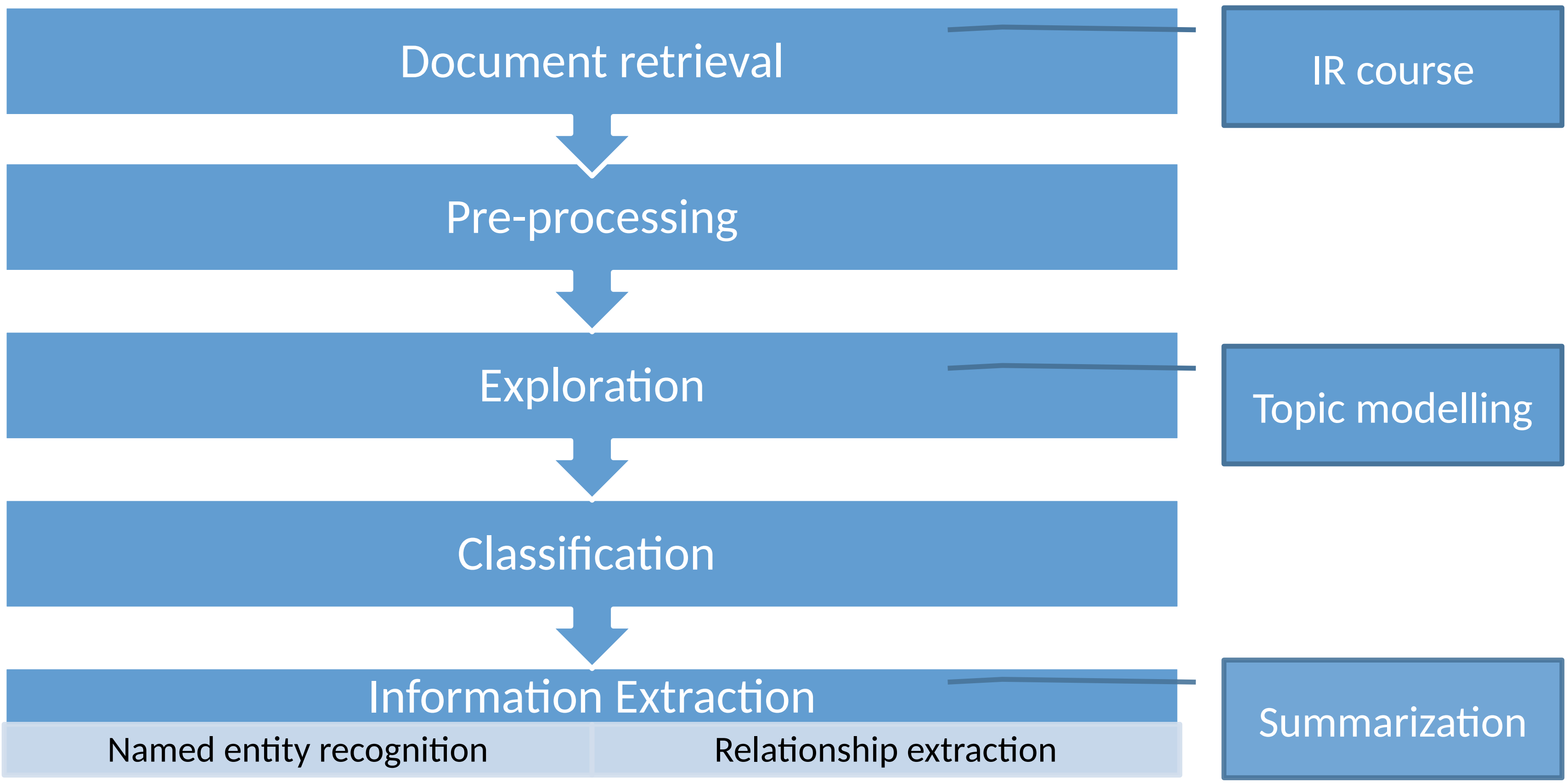
COURSE SUMMARY

MODELS, TASKS, APPLICATIONS



TASKS

TASKS IN THE TM PIPELINE



PRE-PROCESSING

Cleaning

- PDF/docx/HTML to text
- Language filtering
- Encoding issues
- Regex patterns
- Spelling correction

Minimal edit distance

Linguistic pipeline

- Tokenization
- Stop word removal
- Lemmatization/stemming
- POS-tagging

CLASSIFICATION

- Multi-class vs multi-label
- Feature selection
- Term weighting: tf-idf

INFORMATION EXTRACTION

- Named entity recognition
  - Segmentaton & classification
  - Sequence labelling task with IOB-labels
  - Rule-based, feature-based, neural-network based
  - Help of dictionaries

- Relation extraction
  - Co-occurrences, patterns, classification
  - Distant supervision

SUMMARIZATION

- Extractive: sentence classification
- Abstractive: sequence-to-sequence (compare with translation)

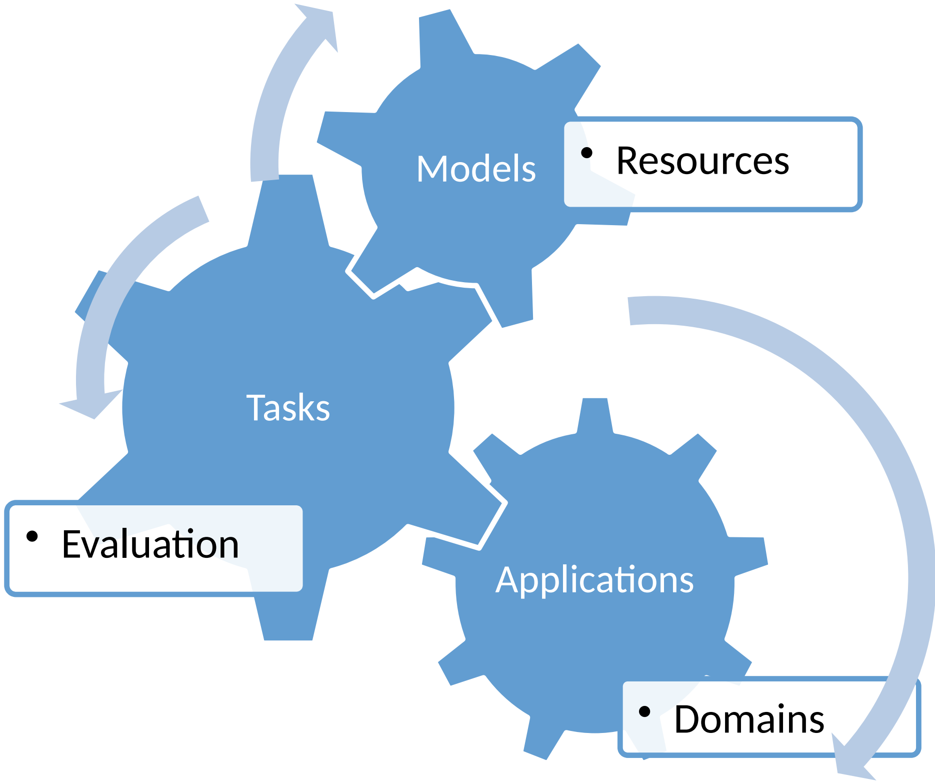
- Challenges:
  - Training data (ground truth, humans disagree)
  - Evaluation

EVALUATION

- Intrinsic: comparison against ground truth (=human) for task
  - Metrics:
    - Accuracy
    - Precision, recall, F1
    - ROUGE for summarization
  - Train-test split to prevent overfitting, or cross validation
    - Hyperparameter tuning on train-tune-set, or cross validation (GridSearchCV)

- Extrinsic: effectiveness in context

MODELS, TASKS, APPLICATIONS



MODELS

MODELS

- Neural language models (embeddings)
  - Traditional: word2vec (a NN with 1 hidden layer) and others
  - Transformer-based: BERT
- Goal: from high-dimensional sparse vectors (10,000s) to lower-dimensional (~100-800) dense vectors
- Pre-trained as a word/sentence prediction task
  - = Language modelling
- The hidden layer has the dimensionality of the embeddings

Distributional hypothesis

MODELS

- Classification
  - Vector space model, bag of words
  - Dimensionality reduction
  - Machine learning methods
    - Naïve Bayes (probabilistic)
    - Support Vector Machines (vector space)
    - Feedforward neural networks (compare with logistic regression, multi-node, multi-layer)
    - Transformer models (BERT)

MODELS

- Sequence labelling
  - Conditional Random Fields
  - Recurrent Neural Networks / Bi-LSTMs
  - Transformer models: BERT

- Sequence-to-sequence
  - Encoder-decoder models
  - Transformer models

TRANSFER LEARNING

- Pre-trained neural language models allow for transfer learning
- Inductive transfer learning: transfer the knowledge from pretrained language models to any text mining task
  - Fine-tuning
- Current state-of-the-art for many text mining tasks

RESOURCES

- Labeled data
  - For training and evaluating task-specific models
  - Typically small (1000s of examples, or even 100s)
  - Supervised learning
  - How to obtain labelled data:
    - Benchmark data
    - Existing expert labels
    - User-generated content
    - Data annotation (crowdsourcing)

Inter-rater agreement (Cohen's Kappa)

RESOURCES

- Unlabeled data
  - For pre-training language models (language modelling = word prediction/sentence prediction)
  - General vs domain-specific
  - Typically large (Millions of words)

RESOURCES

- Dictionaries (gazetteers)
- Ontologies / taxonomies
- General domain (names, places) or specific domain (e.g. bio)

Controlled vocabulary

APPLICATIONS

APPLICATIONS

- Sentiment analysis
  - Classification / ordinal regression / linear regression
  - Extraction: aspect-based sentiment analysis (E,A,S,H,C)
- Argument mining
  - Sentence classification (classifying argument components into different types such as claims and premises)
  - Structure identification focuses (linking arguments or argument components)

APPLICATIONS

- CV-vacancy matching
  - Extracting the structure of a CV
  - Extracting structured fields (e.g. name, education) from a CV
  - Mapping words with similar meaning (e.g. function titles) between CV and vacancy

DOMAINS

- Social media analytics (classification, extraction, sentiment)
- E-commerce (sentiment classification/extraction, ontologies)
- Biomedical text mining (classification/retrieval, extraction)
  - Clinical applications (EHRs)
- Humanities, historical documents (pre-processing, classification/retrieval, extraction)

EXERCISES

EXERCISES

- Minimal edit distance
- Tf-idf
- Naïve Bayes
- Inter-rater agreement
- Precision and recall

MINIMAL EDIT DISTANCE

- Compute the Levenshtein distance between ‘where’ and ‘hear’. Show your computation.

MINIMAL EDIT DISTANCE

		H	E	A	R
	0	1	2	3	4
W	1				
H	2				
E	3				
R	4				
E	5				

MINIMAL EDIT DISTANCE

cost	operation	input	output
1	delete	w	
0	(copy)	h	h
0	(copy)	e	e
1	substitute	r	a
1	substitute	e	r
Total: 3			

TF-IDF

- We have a collection of 100,000 movie scripts.
  a. The term *Elsa* occurs in 10 scripts. What is the inverse document frequency for *Elsa*? Show your computation.
  b. We have a film script *s* in which *Elsa* occurs 2 times. What is the tf-idf weight for *Elsa* in *s*?

TF-IDF

a. idf = $\log_{10}(100{,}000/10) = \log_{10}(10{,}000) = 4$

b. tf = $1+ \log_{10}(2) \approx 1.3$
   tf*idf = $1.3 * 4 = 5.2$

NAÏVE BAYES

- Consider this toy training set for a text classification task with Naïve Bayes:

Doc id	Content	Class
1	make our garden grow	relevant
2	we make the best of it	not relevant
3	together we can grow	not relevant
4	we make the best plans	not relevant

Doc id	Content	Class
1	make our garden grow	relevant
2	we make the best of it	not relevant
3	together we can grow	not relevant
4	we make the best plans	not relevant

a. What is the prior probability of the ‘relevant’ class?

b. What is the vocabulary size of the training set? Assume that we do not remove stop words.

c. Estimate *P(‘make’, not relevant)* using the maximum likelihood estimate on the train set.

d. Why is add-one smoothing needed when we estimate the probability of an unseen document? Provide an example test document given the toy training set for which add-one smoothing is needed.

Doc id	Content	Class
1	make our garden grow	relevant
2	we make the best of it	not relevant
3	together we can grow	not relevant
4	we make the best plans	not relevant

a. 1/4

b. 12

c. (2+1)/(15+12)

d. Because a word in the test document that does not occur in the training set will have a zero probability and the multiplication of zero probabilities will lead to a combined probability of zero; a correct example would be any text with a word that does not occur in training set.

INTER-RATER AGREEMENT

- Compute Cohen's Kappa for this agreement table. Show your computation. (You can keep the last fraction of your computation as it is, without estimating the decimal numbers.)

Agreement table		Annotator 2		
		Positive	Negative	Neutral
Annotator 1	Positive	25	10	5
	Negative	0	25	15
	Neutral	5	5	10

INTER-RATER AGREEMENT

Agreement table		Annotator 2			
		Positive	Negative	Neutral	
	Positive	25	10	5	**40**
	Negative	0	25	15	**40**
	Neutral	5	5	10	**20**
		30	**40**	**30**	**100**

Pr(a)	60 / 100		0.6
Pr(e,pos)	40 / 100 * 30 / 100		0.12
Pr(e,neg)	40 / 100 * 40 /100		0.16
Pr(e,neut)	30 / 100 * 20 / 100		0.06
Pr(e)	(sum of the 3 rows above)		0.34
kappa	(0.6 - 0.34) / (1 - 0.34) = 0.26 / 0.66		

Universiteit Leiden

PRECISION AND RECALL

- Consider the following output table of an automatic classifier for 10 documents. Compute (please show the fractions):

a. the recall for the A class

b. the precision for the A class

doc id	class assigned by classifier	ground truth class
1	A	C
2	A	A
3	B	B
4	B	A
5	C	C
6	A	A
7	D	A
8	A	D
9	B	B
10	C	A

PRECISION AND RECALL

- Consider the following output table of an automatic classifier for 10 documents. Compute (please show the fractions):

a. Recall(A) = 2/5

b. Precision(A) = 2/4

doc id	class assigned by classifier	ground truth class
1	A	C
2	A	A
3	B	B
4	B	A
5	C	C
6	A	A
7	D	A
8	A	D
9	B	B
10	C	A

CONCLUSIONS

SUZAN VERBERNE 2021

Universiteit Leiden

HOMEWORK

- Work on the final assignment

- Prepare for the exam
  - Exam Text mining, Thursday January 13,
    10.15 – 13.15
    - (Students with a provision card get an addition 30 minutes)
  - Location: GORL / 04/5
  - The exam is closed book, individual

  - Practice materials are on Brightspace

THANK YOU

- Thank you for the participation in this course,
- And a big thanks to the TAs:
  - Michiel van der Meer
  - Juan Bascur Cifuentes
  - Cheyenne Health
  - Hainan Yu

TIME FOR XKCD