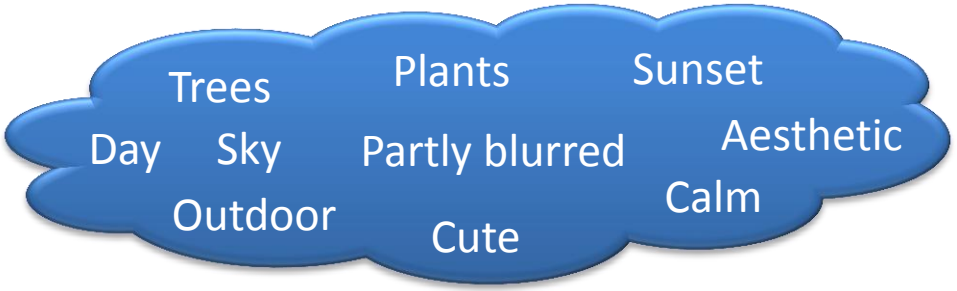
MLKD's Participation at the ImageCLEF 2011 Photo Annotation and Concept-Based Retrieval Tasks

Eleftherios Spyromitros-Xioufis, Konstantinos Sechidis,
Grigorios Tsoumakas and Ioannis Vlahavas
Machine Learning and Knowledge Discovery Group,
Department of Informatics, Aristotle University of Thessaloniki, Greece

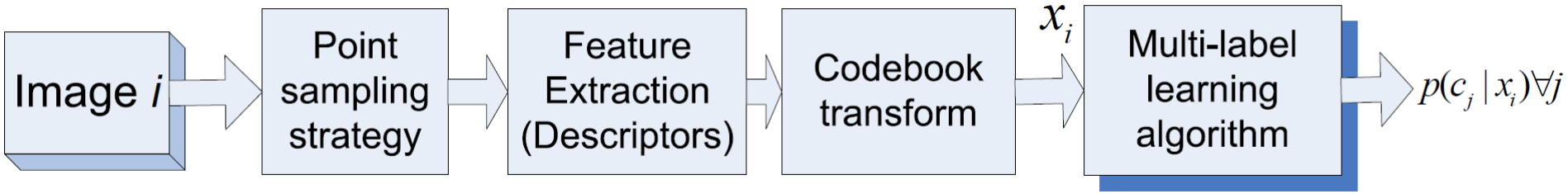
Photo annotation task



- A multi-label classification problem (each image belongs to many concepts)
- Evaluation measures:
  1. Mean interpolated average precision (**MIAP**)
  2. Example-based F-measure (**F-ex**)
  3. Semantic R-precision (**SR-Precision**)
- Model selection: based on Mean Average Precision (MAP)
- MAP estimation: 3 fold cross-validation on the 8000 training images
- 5 submissions in total:
  - Visual
  - Textual
  - Multi-modal (3 variations)

Eleftherios Spyromitros–Xioufis | espyromi@csd.auth.gr

Visual model – feature extraction



- The ColorDescriptor [van de Sande et al., 2010] software was used for visual feature extraction
- 2 point detection strategies: Harris-Laplace, Dense Sampling
- 7 descriptors: SIFT, HSV-SIFT, HueSIFT, OpponentSIFT, C-SIFT, rgSIFT and RGB-SIFT
- Codebook generation
  - K-means (other?) clustering on 250,000 randomly sampled points (more points?)
  - Codebook size (k) fixed to 4096 words (more words?)
  - Hard assignment of points to clusters
- 14 multi-label training datasets in total
  - #features: 4096
  - #labels: 99

Visual model – learning method

- The **Binary Relevance** (problem transformation) method was used:
  - Transforms the multi-label classification task into multiple binary classification tasks
  - Any single-label classifier can be used (Random Forest #trees:150 #features:40)
  - Instance weighting to deal with class imbalance:

$$w_{min} = \frac{min+maj}{min} \qquad w_{maj} = \frac{min+maj}{maj}$$

Training set for $\boldsymbol{\lambda_1}$

	f_1	f_2	…	f_{4096}	λ_1	λ_2	…	λ_{99}
x_1	1	0	…	1	0	1	…	1
x_2	0	1	…	0	1	0	…	0
…	…	…	…	…	…	…	…	…
x_{8K}	0	0	…	1	0	0	…	1

Feature Space | Target

Eleftherios Spyromitros–Xioufis | espyromi@csd.auth.gr

Visual model – learning method

- The **Binary Relevance** (problem transformation) method was used:
  - Transforms the multi-label classification task into multiple binary classification tasks
  - Any single-label classifier can be used (Random Forest #trees:150 #features:40)
  - Instance weighting to deal with class imbalance:

$$w_{min} = \frac{min+maj}{min} \qquad w_{maj} = \frac{min+maj}{maj}$$

Training set for $\boldsymbol{\lambda_2}$

	f_1	f_2	…	f_{4096}	λ_1	λ_2	…	λ_{99}
x_1	1	0	…	1	0	1	…	1
x_2	0	1	…	0	1	0	…	0
…	…	…	…	…	…	…	…	…
x_{8K}	0	0	…	1	0	0	…	1

Feature Space

Target

Eleftherios Spyromitros–Xioufis | espyromi@csd.auth.gr

Visual model – learning method

- The Binary Relevance (problem transformation) method was used:
  - Transforms the multi-label classification task into multiple binary classification tasks
  - Any single-label classifier can be used (Random Forest #trees:150 #features:40)
  - Instance weighting to deal with class imbalance:

$$w_{min} = \frac{min+maj}{min} \qquad w_{maj} = \frac{min+maj}{maj}$$

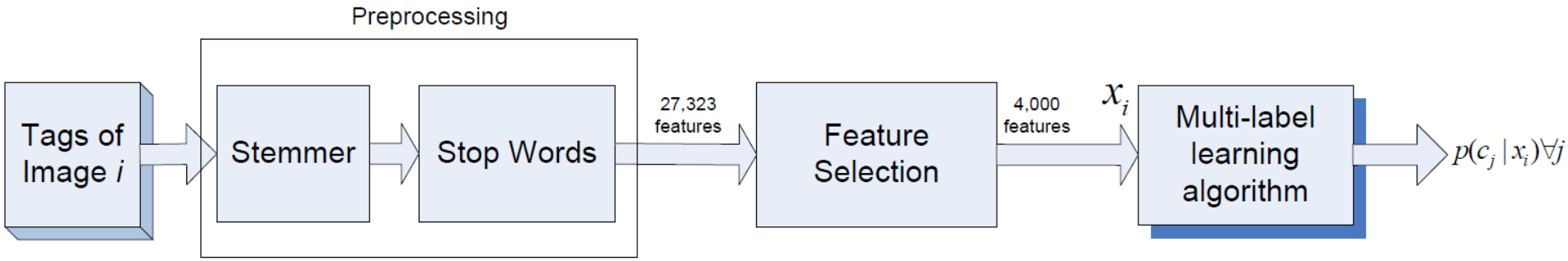
Training set for $\boldsymbol{\lambda_{99}}$

	f_1	f_2	…	f_{4096}	λ_1	λ_2	…	λ_{99}
x_1	1	0	…	1	0	1	…	1
x_2	0	1	…	0	1	0	…	0
…	…	…	…	…	…	…	…	…
x_{8K}	0	0	…	1	0	0	…	1

Feature Space

Target

Textual model – feature extraction



- Flickr user tags were used
- Initial vocabulary: the union of tag sets of the training images
- Stemming : porter stemmer (English..) & stop word removal -> 27000 stems
- Feature selection using χ^2_{max} criterion [Lewis et al., 2004]:
  - χ^2 statistic for each feature with respect to each label is calculated
  - Features are ranked according to their maximum χ^2 score across all labels
  - After evaluation of different sizes top 4000 features were selected

Textual model – learning method

- Ensemble of Classifier Chains (ECC) [Read et al., 2009]:
  - Random chains are created
  - Feature set for each label in the chains is augmented with the previous labels
  - Able to capture correlations, class imbalance is still a problem

Training set for λ_1

Chain order: 1,2,..,99

	f_1	f_2	…	f_{4000}	λ_1	λ_2	…	λ_{99}
x_1	1	0	…	1	0	1	…	1
x_2	0	1	…	0	1	0	…	0
…	…	…	…	…	…	…	…	…
x_{8K}	0	0	…	1	0	0	…	1

Feature Space | Target

Textual model – learning method

- Ensemble of Classifier Chains (ECC) [Read et al., 2009]:
  - Random chains are created
  - Feature set for each label in the chains is augmented with the previous labels
  - Able to capture correlations, class imbalance is still a problem

Chain order: 1,2,..,99

Training set for λ_2

	f_1	f_2	…	f_{4000}	λ_1	λ_2	…	λ_{99}
x_1	1	0	…	1	0	1	…	1
x_2	0	1	…	0	1	0	…	0
…	…	…	…	…	…	…	…	…
x_{8K}	0	0	…	1	0	0	…	1

Feature Space | Target

Textual model – learning method

- ECC is also a problem transformation method:
  - Again coupled with Random Forest as base classifier (#trees:10, #features:default)
  - Ensemble size: 15 (150 random trees in total for each label)
  - Again instance weighting for class imbalance

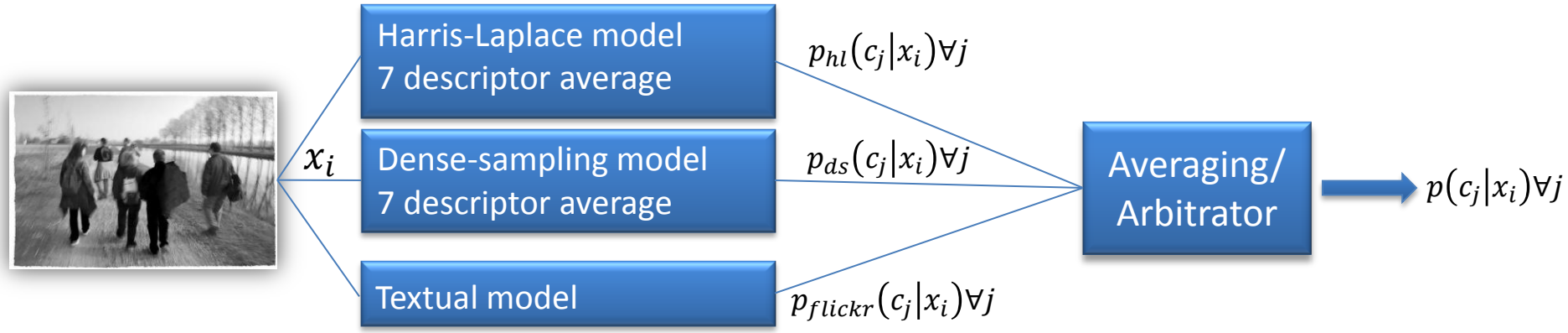
Training set for λ_{99}

Chain order: 1,2,..,99

	f_1	f_2	…	f_{4000}	λ_1	λ_2	…	λ_{99}
x_1	1	0	…	1	0	1	…	1
x_2	0	1	…	0	1	0	…	0
…	…	…	…	…	…	…	…	…
x_{8K}	0	0	…	1	0	0	…	1

Feature Space | Target

Multi-modal



- A hierarchical late fusion scheme:
  - 3 different views of the images:
    - Harris Laplace -> concepts related to objects (Fish and Ship)
    - Dense sampling -> concepts related to scenes (Night and Macro)
    - Textual -> concepts which are typically tagged by users (Dog , Insect, …)
  - 2 ways to combine the 3 different views:
    - Averaging
    - Arbitrator (the best view based on internal evaluation)

Thresholding – from scores to bipartitions

Horse	Sport	Sky	Plants	Happy	Graffiti
0.76	0.63	0.44	0.33	0.25	0.10

$t = 0.4$

Relevant	Irrelevant
Horse, Sport, Sky	Plants, Happy, Graffiti

- Scores are ok for evaluation on MIAP and SR-precision
- Example-based F-measure a bipartition of concepts to relevant and irrelevant
- The thresholding method described in [Read et al., 2009] was used:
  - A common threshold across all concepts
  - Provides a close approximation of the training set's label cardinality to the test set predictions:
    - $t = argmin_{\{t \in 0.00, 0.05, \ldots, 1.00\}} |LC(D_{train}) - LC(H_t(D_{test}))|$

Eleftherios Spyromitros–Xioufis | espyromi@csd.auth.gr

Concept-based retrieval

- 40 retrieval topics
  - Logical connections of the 99 concepts of the photo annotation task
  - E.g. “Find all images that depict a small group of persons in a landscape scenery showing trees and a river on a sunny day”
  - 2 to 5 example images are also given for each topic
- Goal:
  - A ranked list of the 1000 most relevant photos per topic
  - From a pool of 200.000 non-annotated images
- Evaluation measure:
  - Mean Average Precision, P@10, P@20, P@100, R-prec
- Two approaches:
  - Manual: Using the models learned on the training images
  - Automated: Using the example images

Manual approach

- Given
  - $I = 1, \dots, 200.000$ the collection of retrieval images
  - $Q = 1, \dots, 40$ the set of topics
- We apply out automated image annotation system to each image $i \in I$
  - Textual model + visual models built using only RGB-SIFT features
  - A 99-dimensional vector with relevance scores $S_i = [s_i^1, s_i^2, \dots, s_i^{99}]$
- For each topic $q \in Q$
  - $P_q \subseteq C$, $N_q \subseteq C$ the sets of positively/negatively correlated concepts
  - For each concept c in $P_q \cup N_q$
    - $m_q^c \geq 1$ is a real valued parameter denoting the influence of c to q
- Finally for each topic q and image i, the scores of the relevant concepts are combined:

$$S_{q,i} = \prod_{c \in P_q} (s_i^c)^{m_q^c} \prod_{c \in N_q} (1 - s_i^c)^{m_q^c}$$

- The selection of related concepts and the setting of values for the m_q^c parameters was done using a trial-and-error approach (examining the top 10 retrieved images)

Manual approach - example

Topic 5: rider on horse. *"Here we like to find photos of riders on a horse. So **no sculptures or paintings** are relevant. The rider and horse can be also only in parts on the photo. It is important that the person is riding a horse and not standing next to it."*

- Concepts 75 (**Horse**) and 8 (**Sports**) are positively related (rider on horse)
- Concept 63 (**Visual_Arts**) is negatively related (no sculptures or paintings)
- Therefore:
  - $P_5 = \{75,8\}, N_5 = \{63\}$
  - $m_5^{75} = m_5^8 = m_5^{63} = 1$ (equal strength to all related concepts for this topic)

Eleftherios Spyromitros–Xioufis | espyromi@csd.auth.gr

Automated approach – query by example

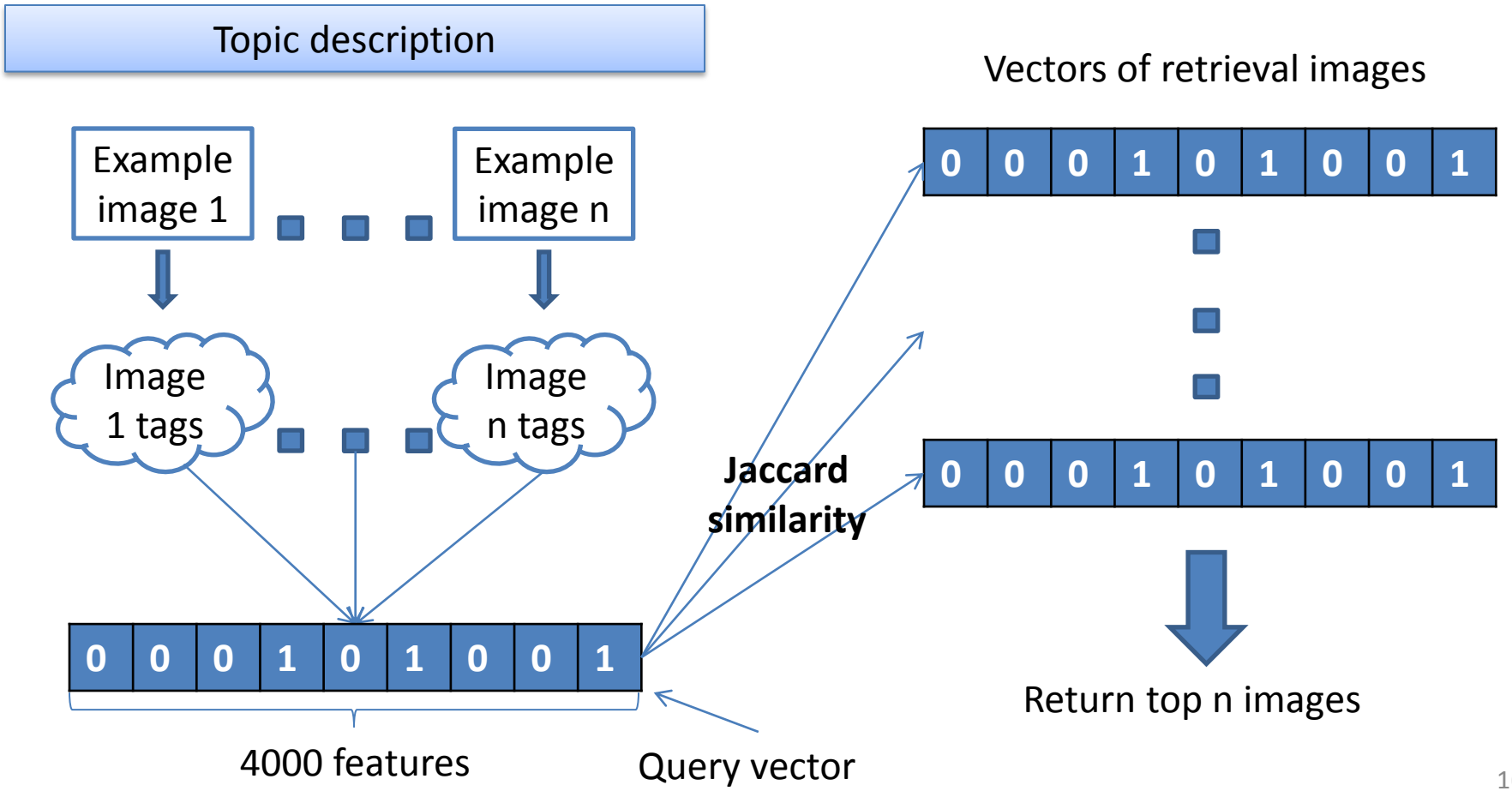


Photo annotation results

Approach	Team ranks - scores		
	MIAP	F-example	SR-Prec
Visual	9th/15 – 0.3114	5th/15 – **0.5595**	9th/15 – 0.6981
Textual	3rd/7 – 0.3256	2nd/7 – 0.5061	3rd/7 – 0.6257
Multi-modal	5th/10 – **0.4016**	5th/10 – 0.5588	7th/10 – **0.6982**
Overall	**5th/18 – 0.4016**	**7th/18 – 0.5595**	**10th/18 – 0.6982**

- Better in MIAP (model selection was based on Mean Average Precision)
- Averaging the multiple models worked better than arbitrating
- Good in textual – bad in visual – average overall

Concept-based retrieval results

Configuration	Submission ranks - scores				
	MAP	P@10	P@20	P@100	R-Prec
Automated	1st/16 – 0.0849	1st/16 – 0.4100	1st/16 – 0.2800	1st/16 – 0.2188	1st/16 – 0.1530
Manual	1st/15 – 0.1640	1st/15 – 0.4175	1st/15 – 0.3838	1st/15 – 0.3180	1st/15 – 0.2467
Overall	**1st/31 – 0.1640**	**1st/31 – 0.4175**	**1st/31 – 0.3838**	**1st/31 – 0.3180**	**1st/31 – 0.2467**

- He are ranked 1st both in the automated and the manual retrieval approach
- Manual performs much better than automated on average
- Surprisingly automated performed better on 9 topics!

Conclusions – Future work

- Lessons learned:
  - We need collaboration with a computer vision/image group
  - Binary multi-label classification approaches work well:
    - Coupled with strong base learners (Random Forest)
    - Class imbalance issues should be handled
  - Measure specific model selection is needed:
    - Suggestion: more submissions should be allowed to the annotation task
- Future directions:
  - Better preprocessing of textual information (e.g. translate non-English tags)
  - Other hierarchical late fusion schemes – more advanced arbitration techniques
  - Better thresholding approaches
  - Experiments with more multi-label methods and base classifiers
  - Explore why we performed so well in the concept-based retrieval task

Eleftherios Spyromitros–Xioufis | espyromi@csd.auth.gr

THANK YOU!
QUESTIONS?

Eleftherios Spyromitros–Xioufis | espyromi@csd.auth.gr

Software used - acknowledgements

- Software tools
  - Mulan (http://mulan.sourceforge.net/)
    - Multi-label classification, feature selection and thresholding methods
    - Evaluation Framework
  - ColorDescriptor (http://koen.me/research/colordescriptors/)
    - Image feature extraction
  - Weka (http://www.cs.waikato.ac.nz/ml/weka/)
    - Text preprocessing – codebook generation (k-means clustering)
- Acknowledgements
  - PetaMedia: student travel support
  - European Science Foundation: student registration

Key references

- Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., Vlahavas, I.: Mulan: A java library for multi-label learning. Journal of Machine Learning Research (JMLR) 12, 2411-2414 (July 12 2011)
- van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1582{1596 (2010)
- Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. In: Proc. 20th European Conference on Machine Learning (ECML 2009). pp. 254{269 (2009)
- Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: Rcv1: A new benchmark collection for text categorization research. J. Mach. Learn. Res. 5, 361{397 (2004)