Using Topic Models for Twitter Hashtag Recommendation

Fréderic Godin, Viktor Slavkovikj, Wesley De Neve,
Benjamin Schrauwen and Rik Van de Walle

Multimedia Lab, Ghent University – iMinds, Belgium

Reservoir Lab, Ghent University, Belgium

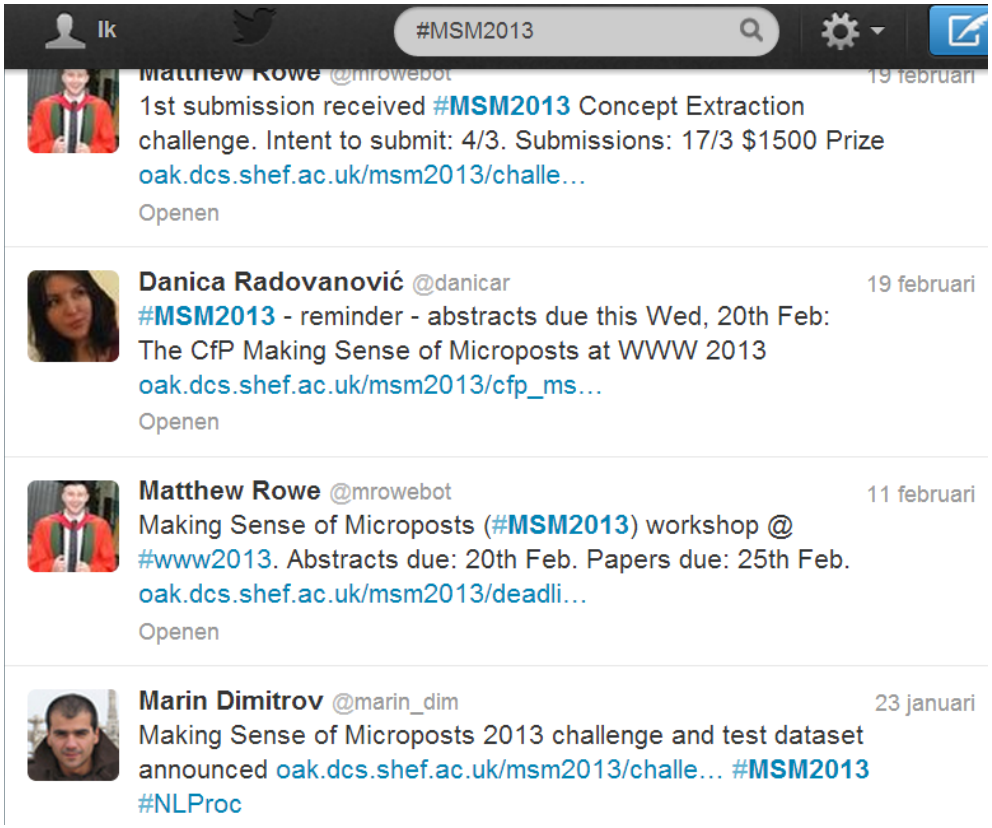
Image and Video Systems Lab, KAIST, South Korea

Introduction (1)

Search

Linking

Memes



General Topic

Indexing

Grouping

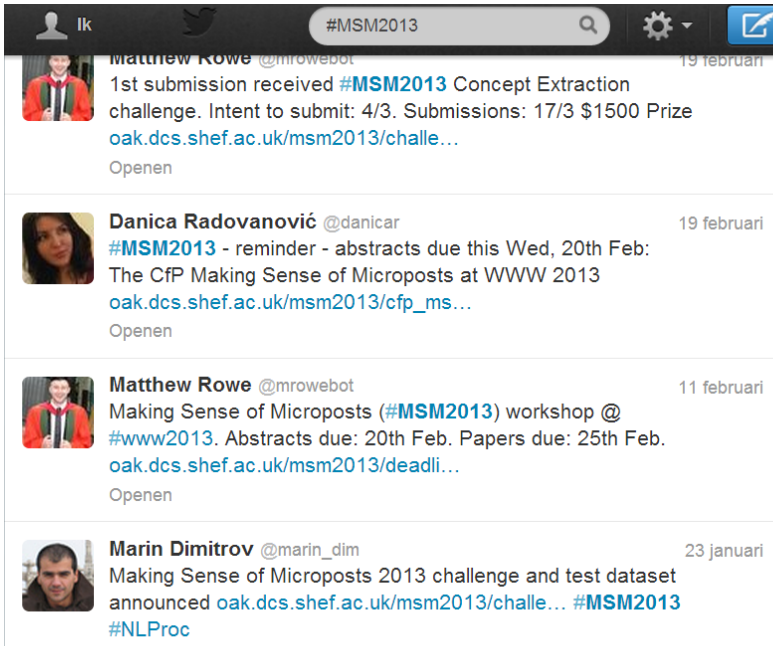
Information retrieval

Introduction (2)

Search

Linking

Memes



General Topic

Indexing

Grouping

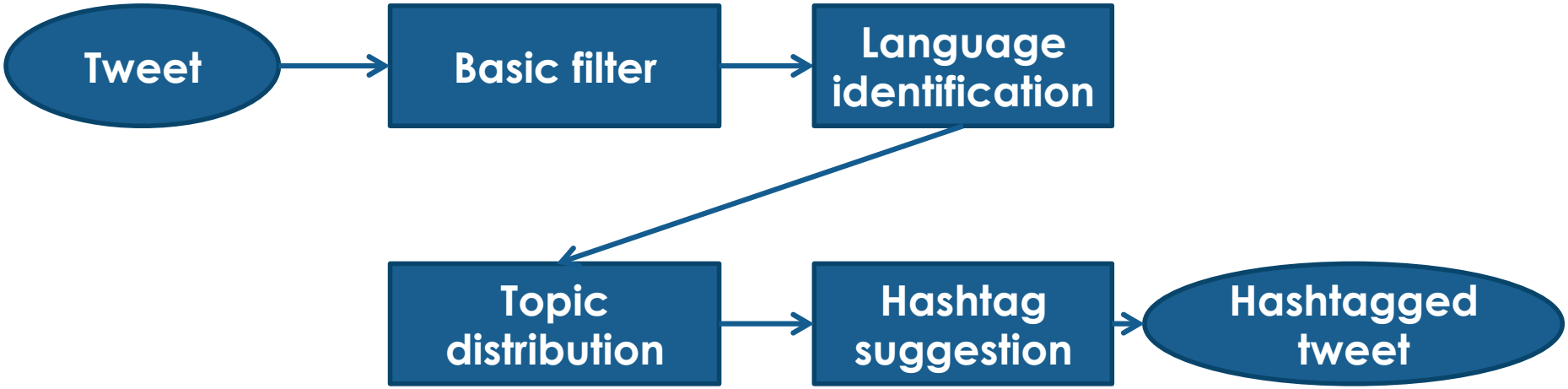
- ±10% of tweets contain a hashtag
- 3% of the hashtags are used more than 5 times

Goal

Suggest keywords that resemble the general topic of a tweet and that could be used as a hashtag

- Promote hashtags for effective indexing
- Allow for effective search of tweets through hashtags
- Reduce the use of sparse hashtags

Architectural overview



Basic filter

Clean up the tweet: URLs, special HTML entities, digits, punctuations, the hash character, …

During training:

- Remove tweets with just one word
- Remove retweets

Language identification

Why — We need to build a language-dependent topic model.

Goal — Build unsupervised classifier that discriminates between English and non-English tweets.

How — Using Naive Bayes and the Expectation-Maximization algorithm + character n-gram features

Result — Evaluation on a test set of 1000 randomly selected tweets

	Lui & Baldwin (LangID.py)	Our algorithm
Precision	97.9%	97.0%
Recall	91.8%	97.8%
F1	94.8%	**97.4%**

Calculating the topic distribution

Idea	Find the general topic(s) of a tweet
How	Using Latent Dirichlet Allocation to find the topic distribution in an unsupervised manner
Training	1.8 million tweets pre-filtered on 4000 keywords 200 topics, α=0.1, β=0.1
Example	"Please RT!! sign Bernie Sanders petition for the fiscal cliff! http://.."

0	1	2	3		57		199
[0.1;	0.0 ;	0.0 ;	0.0 ;	… ;	0.8 ;	… ;	0.05]

Topic 57:

1. Fiscal
2. Political
3. President

…

Hashtag suggestion (1)

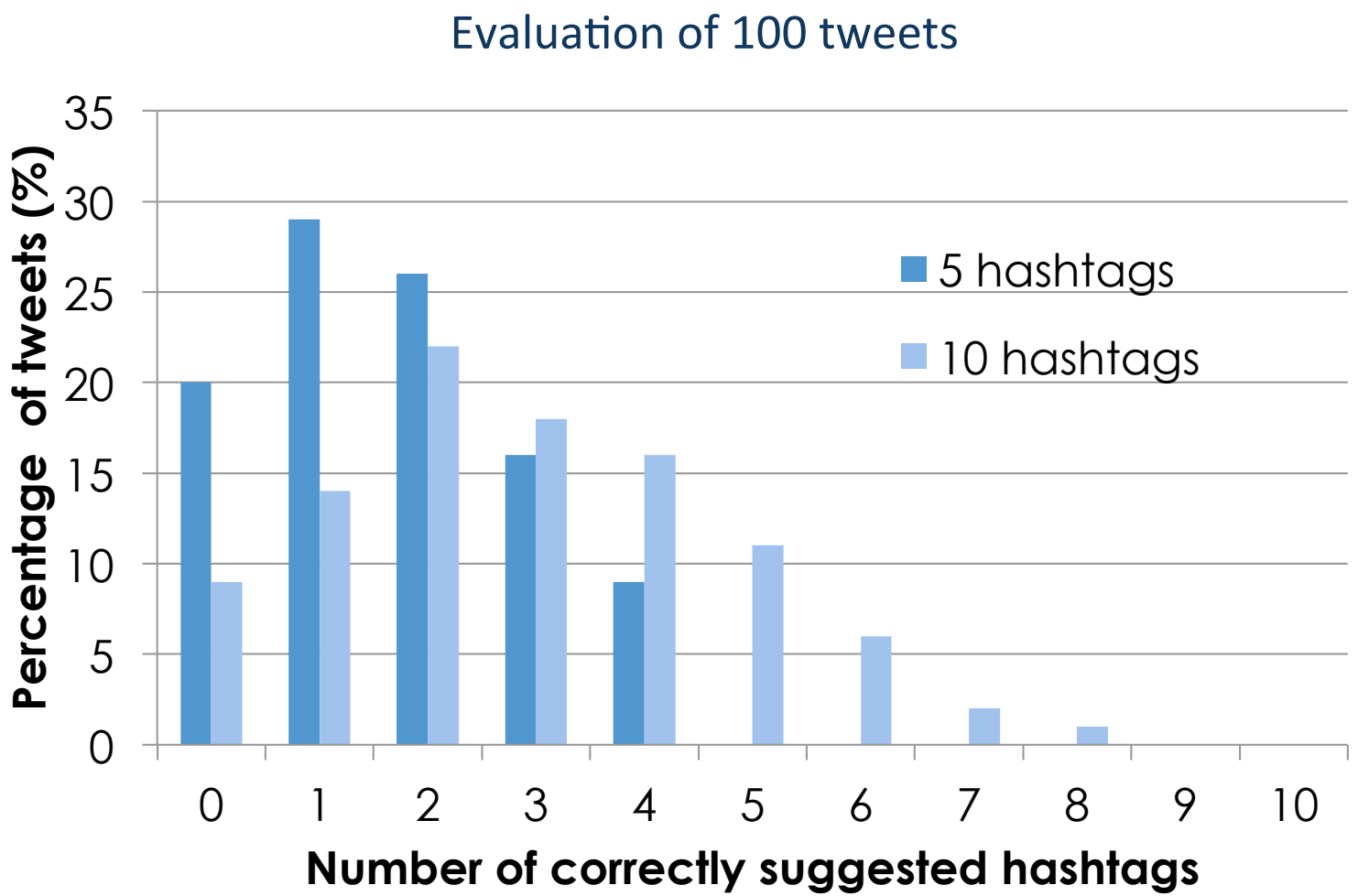
Idea — Suggest a number of hashtags based on the topic distribution of the tweet

How — Sample the topic distribution and suggest the top ranked keywords

Example

Yay, we got sixth period today	**school**	business	light	**time**	**period**
Please RT!! Sign Bernie Sanders petition for the fiscall! Http://..	**fiscal**	**political**	traffic	**president**	**policy**
comfort, elegance, prettiness	little	good	love	relationship	god

Hashtag suggestion (2)



Conclusions and Future Work

We built a hashtag recommendation system:

- Suggests general keywords
- Unsupervised

In the future:

- Use more context information: semantic web, social graph,…
- Adopt a hybrid approach between general and specific hashtags

#Questions @frederic_godin