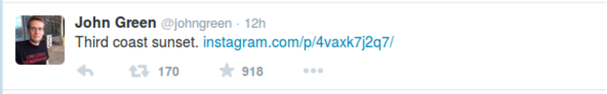
Location-aware online hashtag recommendation

Júlia Pap

Institute for Computer Science and Control
Budapest

July 6, 2015, Bagnères-de-Luchon

About Twitter



John Green @johngreen · 12h
Third coast sunset. instagram.com/p/4vaxk7j2q7/

- microblog service
- users can post short messages,
- and read posts of other users they follow
- other aspects:
  - *hashtag*: topic label (like *#TDF2015*, *#July4*, *#Google*)
  - *mention* another user
  - *retweet* a tweet
  - geographical information

Recommending hashtags online

The task:

- recommend new hashtags to users
- knowing the time and place of their tweets

$$\hat{r}(u, h, l, t) = ?$$

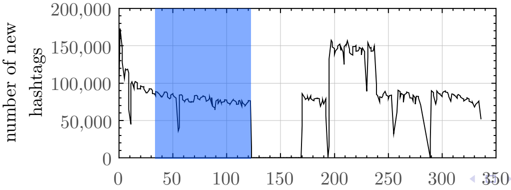
- implicit recommendation
- the location is not unique neither to the user, nor to the hashtags

Our dataset

- tweets from 2012
- through Twitter API
- filter: should contain geo info
- 1,266,004,930 tweets, 173,493,860 containing hashtags

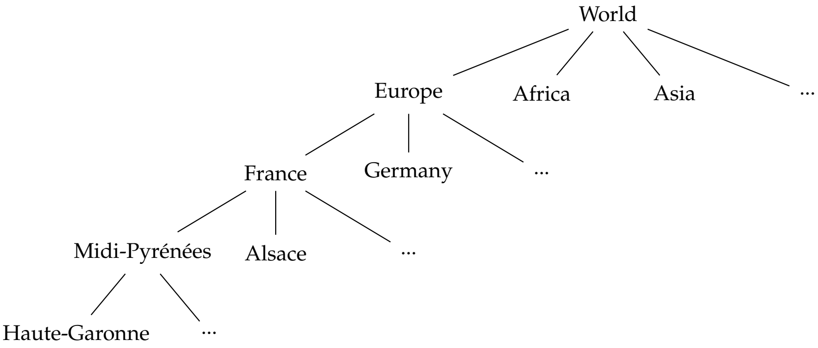
Cleaning the data

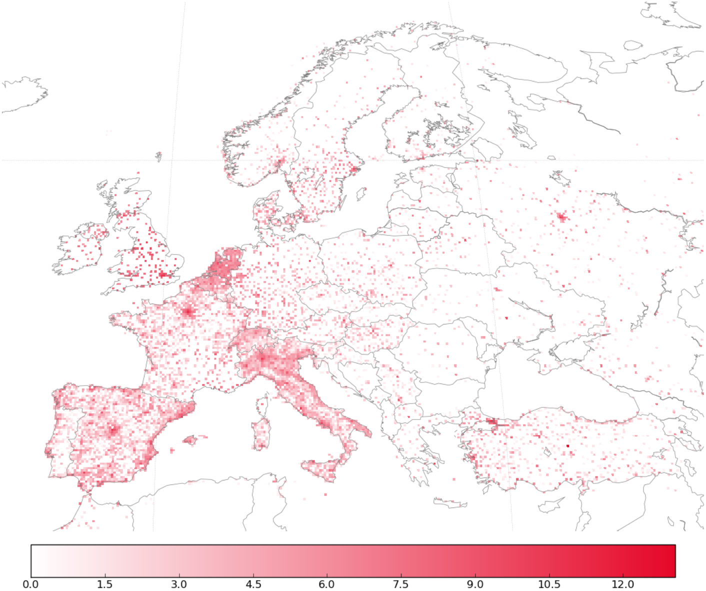
- ∀ (user, hashtag) pair only the first occurrence
- skip the first 3 weeks
- 3 months until a break in the dataset
- 2,993,183 (user, hashtag) pairs from 49 countries



Geographical hierarchy of regions

- idea: use a geographical partition with variable coarseness
- tree of regions from gadm.org
- 214,230 regions, among which 190,315 are leaves
- 17,000 leaves have tweets from the cleaned data
- 5 layers, +1 for continents




0.0
1.5
3.0
4.5
6.0
7.5
9.0
10.5
12.0

Model 1

Popularity by time and location

- count the hashtags in the nodes of the GADM tree,
- in the last time interval
- score: sum on the path from the root

$$\hat{r}(u,h,l,t) = \sum_{l' \in \mathrm{Path}(l)} \log(\mathrm{pop}(l',h,t))$$

- or: learn weights for the nodes:

$$\sum_{l' \in \mathrm{Path}(l)} w_{l'} \cdot \log(\mathrm{pop}(l',h,t))$$

Model 2

Using hashtag recency

- store the last appearance of the hashtags in the nodes.

$$\hat{r}(u, h, l, t) = \sum_{l' \in \mathrm{Path}(l)} w_{l'} \cdot f(t - t_{\mathrm{last}}(l', h))$$

  for time decay function $f(t) = 1 - \left(1 + \frac{\Delta t}{t}\right)^{(1-\alpha)}$
- we learn the w_l weights with SGD

Baseline models

Online matrix factorization

$$\hat{r}(u,h,l,t) = P_u Q_h$$

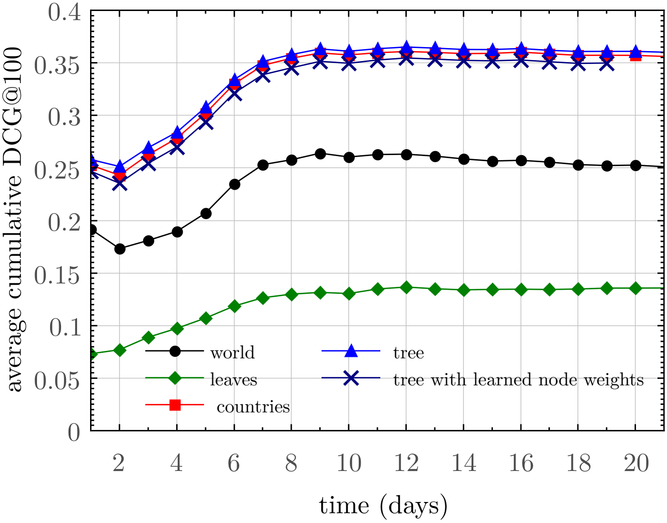
- optimize for MSE using SGD

Nearest neighbors

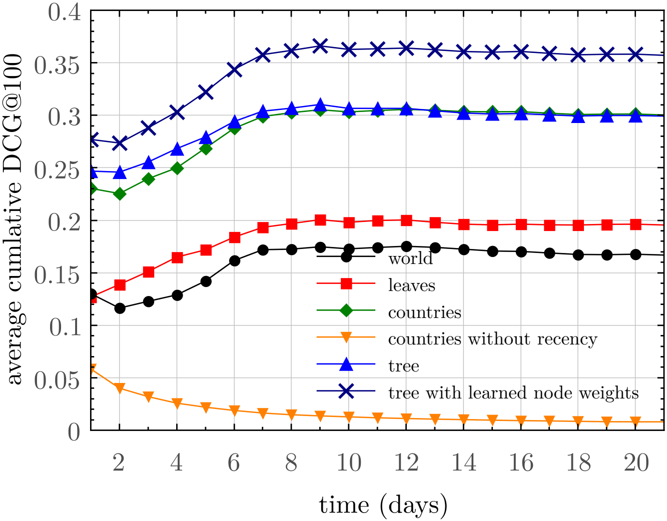
$$\hat{r}(u,h,l,t) = \sum_{(u',h,l',t') \in N_k(l,t,h)} \frac{f(t-t')}{d(l,l')^2}, \text{ where}$$

- f is a time decay function
- $N_k(l,t,h)$ is the set of k nearest tweets to l that uses hashtag h, until time t

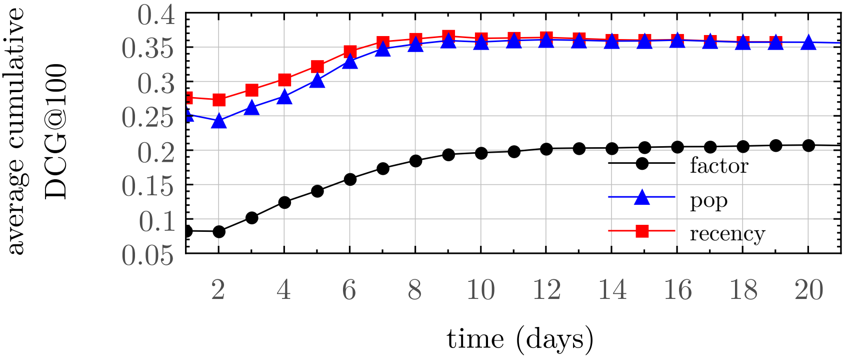
Popularity-based models



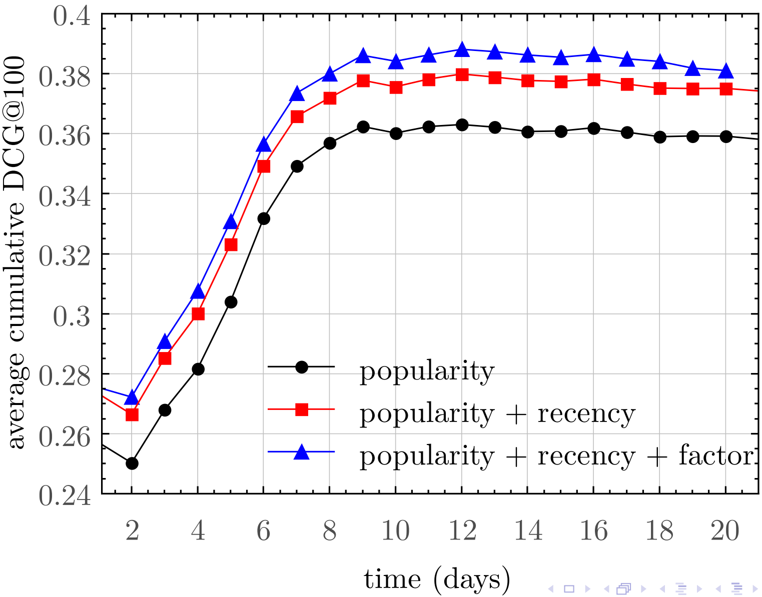
Recency-based models



Best performances



Combination


average cumulative DCG@100
time (days)
popularity
popularity + recency
popularity + recency + factor