Hierarchic Topic Models Visualization

Dmitriy S. Fedoriaka

Moscow Institute of Physics and Technology

November 30, 2016

Introduction

Global task

Given large ($10^3 - 10^6$) database of documents. User wants to explore new area. He isn't expert and don't know keywords. We want make this possible and fast.

Idea

Let's group documents by topics!

Problems

Topic model is set of matrix, useless for user. Each document can be member of more then one topics. We want to show all set of documents as interactive picture.

Topic modeling

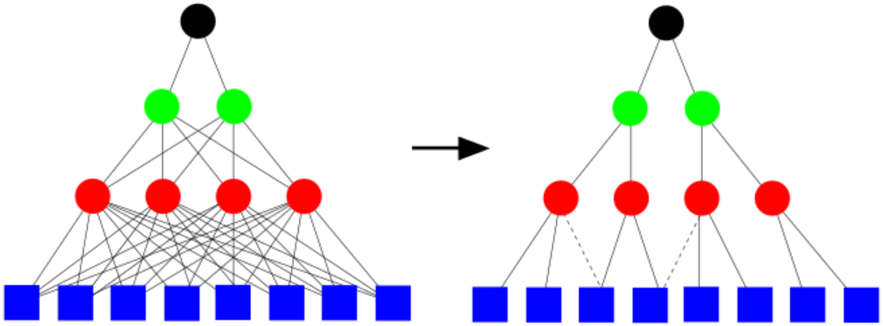
- $F_{w,d}$ — words-documents.
- $\Phi_{w,t}$ — words-topics.
- $\Theta_{t,d}$ — topics-documents.
- $F = \Phi\Theta$.
- $\sum_{d \in D} \sum_{w \in W} |d|_w \ln \sum_{t \in T} \Phi_{w,t}\Theta_{t,d} + R(\Theta, \Phi) \to \max$

Hierarchical tree building

- Nodes: documents, topics, root.
- Extracting tree from matrix:

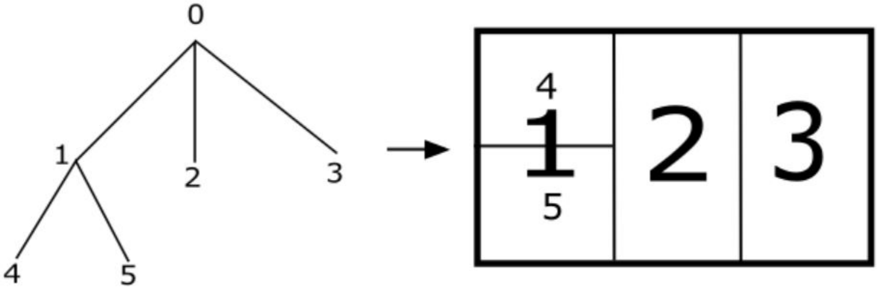
$$Topic(d) = \arg\max_t \Theta_{t,d}$$

$$MultiTopic(d) = \arg\max_t \Theta_{t,d} \cup \{t|\Theta_{t,d} > \varepsilon\}$$

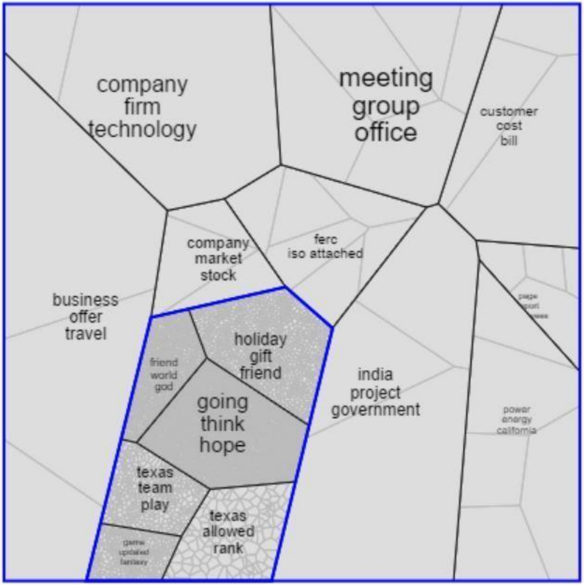


Visualization of hierarchical tree

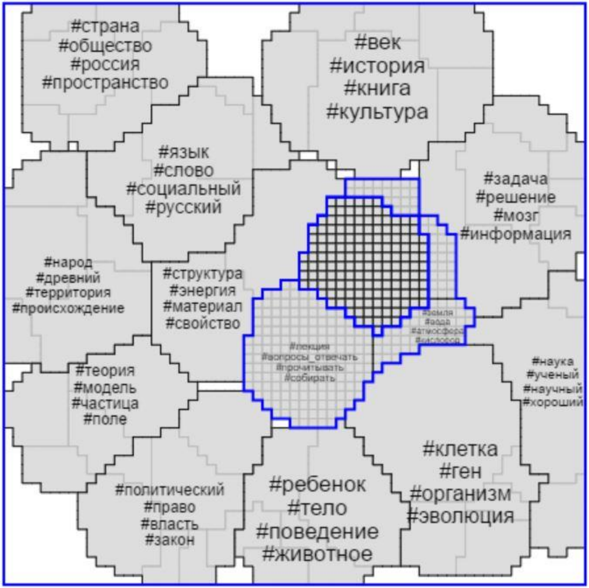
- *"Overview first, zoom and filter, details on demand"*.
- Solution: inserted polygons.



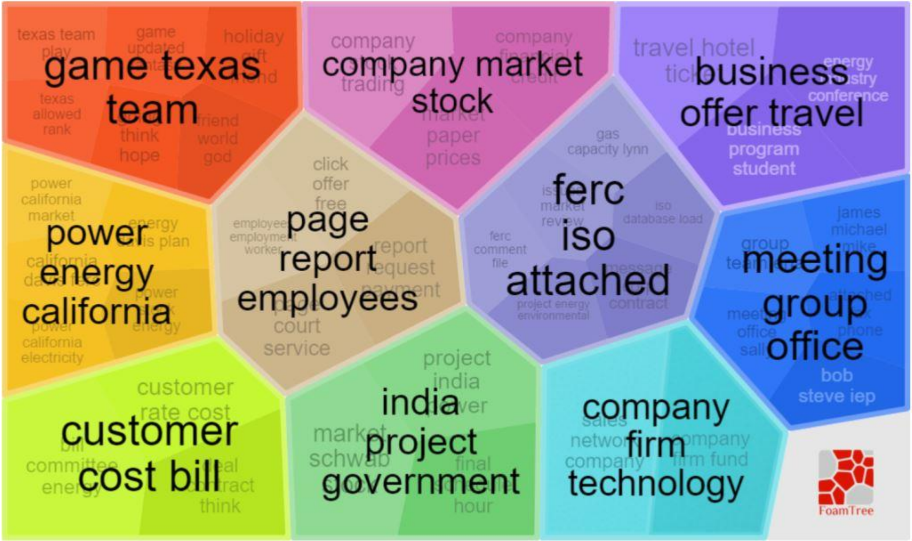
Random Voronoi map



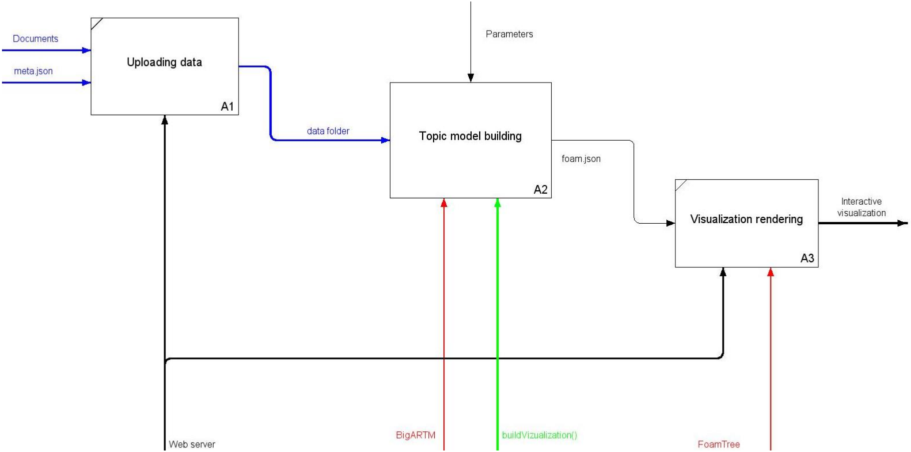
Grid visualization



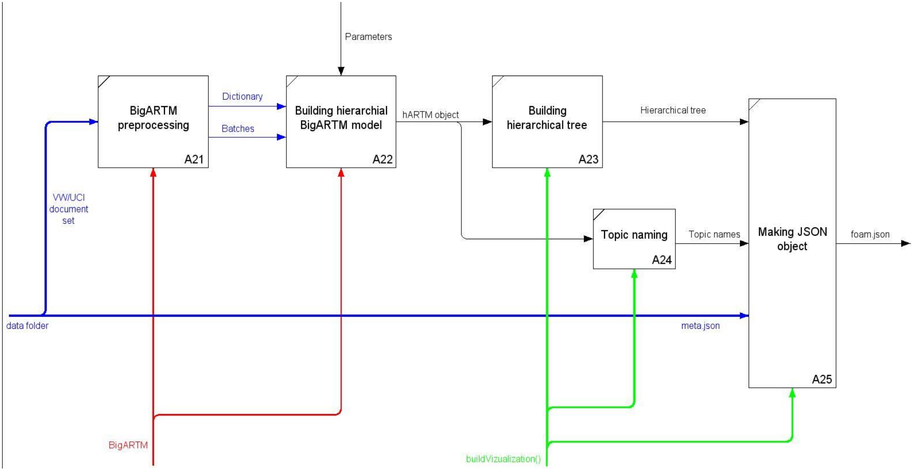
FoamTree


game texas team
texas team play
game updated
holiday gift
texas allowed rank
think hope
friend world god
company market stock
company trading
company
market paper prices
business offer travel
travel hotel
energy industry conference
business program student
power energy california
power california market
energy
power
california
power california electricity
page report employees
click offer free
employee employment worker
report request payment
court service
ferc iso attached
gas capacity lynn
iso database load
ferc comment file
project energy environmental
contract
meeting group office
james michael mike
group team
attached
phone
office
bob steve iep
customer cost bill
customer rate cost
committee energy
think
india project government
project india
market schwab
final hour
company firm technology
sales network company
company firm fund
FoamTree

Implementation



Implementation



Quality measurement

- Main goal: make search and exploration **faster**. So, metric should be **time** of search by user.
- Improvement coefficient:

$$\alpha = \frac{t_0}{t}$$

(t_0 — basic system time, t — time with our system)

Conclusion

- Problem of document sets visualization solved using topic models.
- Web service implemented.
- Global goal: generic all-in-one system: preprocessing + topic modeling + visualization.