Computational methods for text analysis

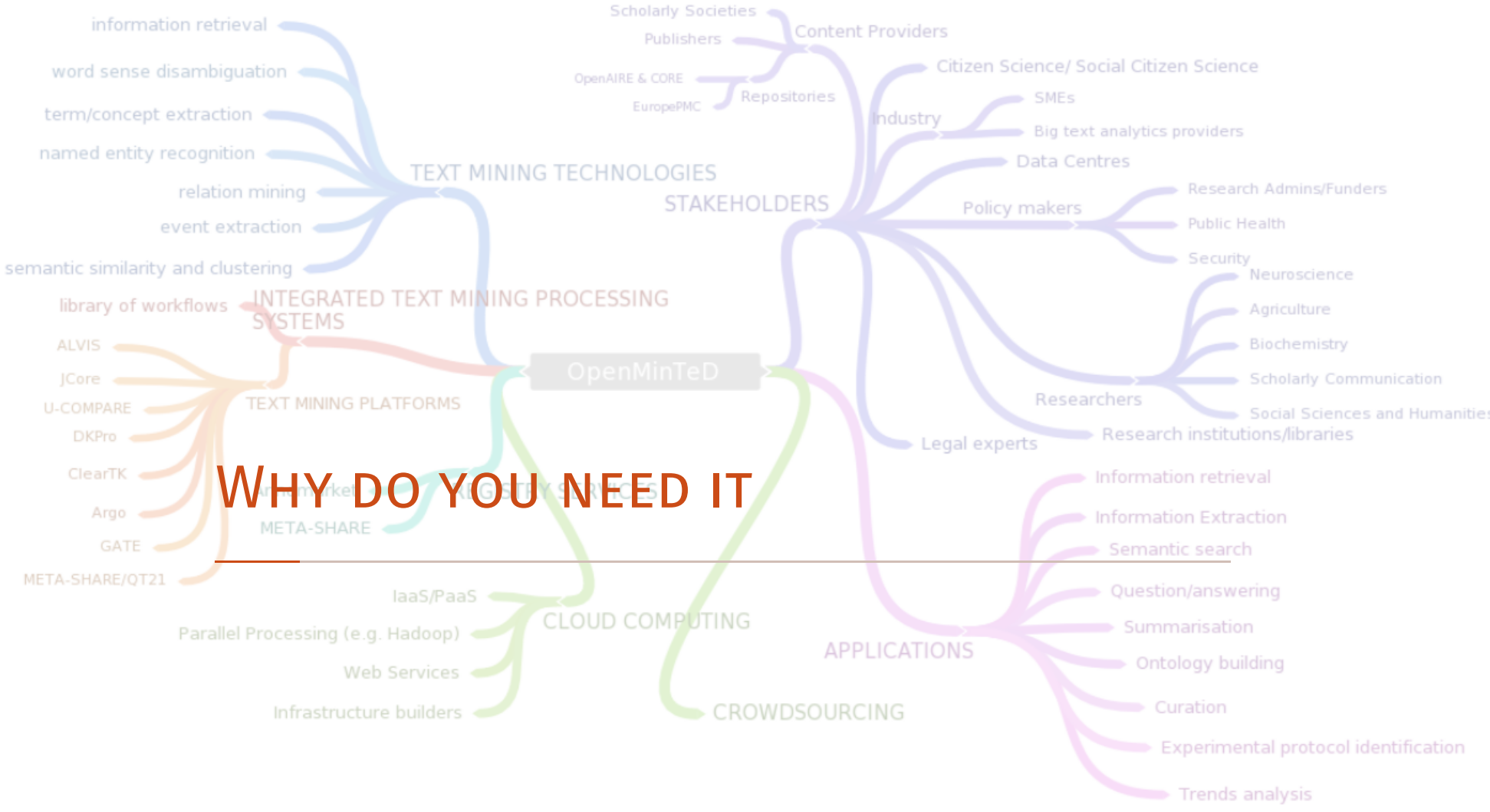
BA program "Sociology and Social Informatics"

Kirill Maslinsky

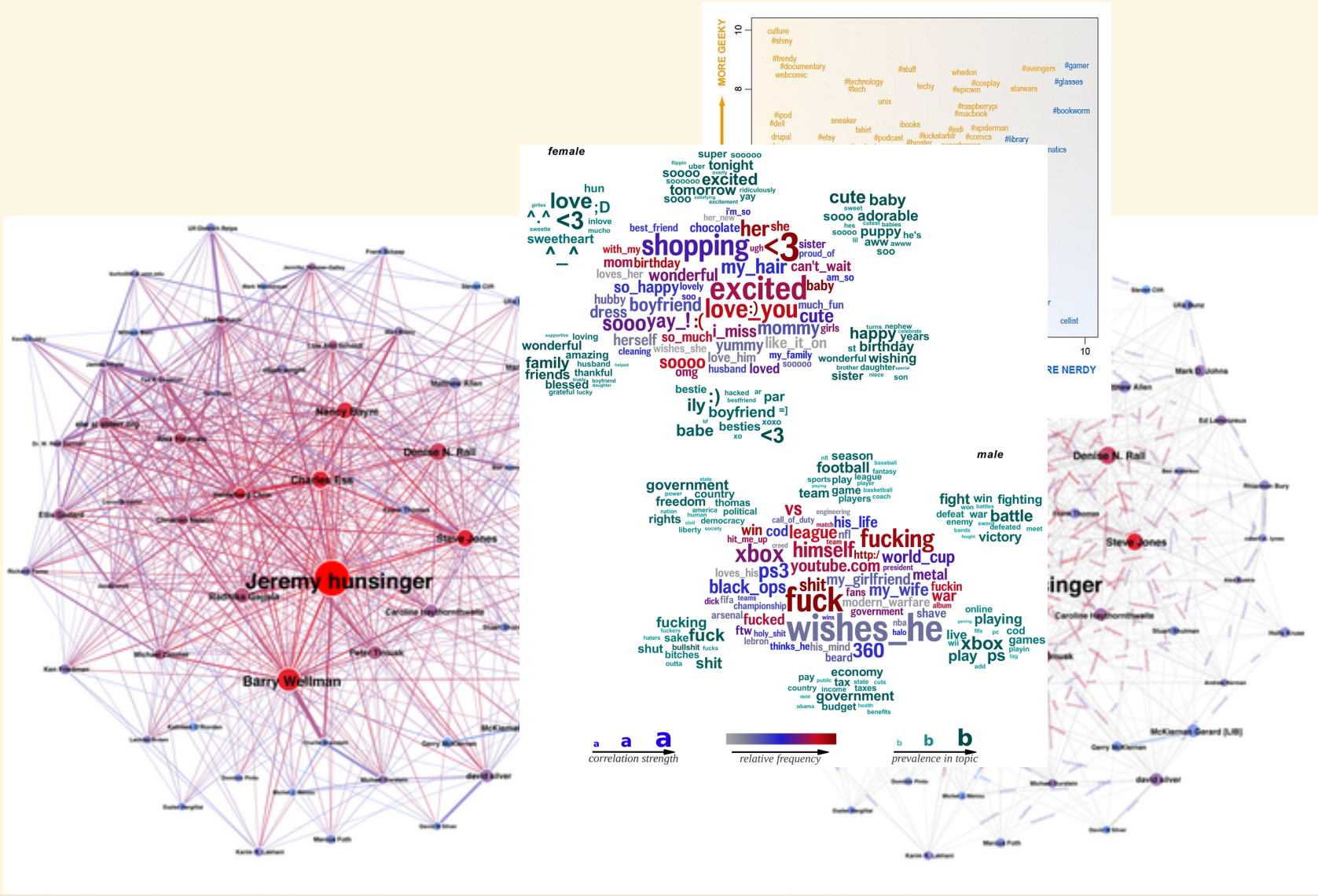
2018

Higher School of Economics — Saint Petersburg

Why do you need it



Just to learn to make those pictures



Scale up

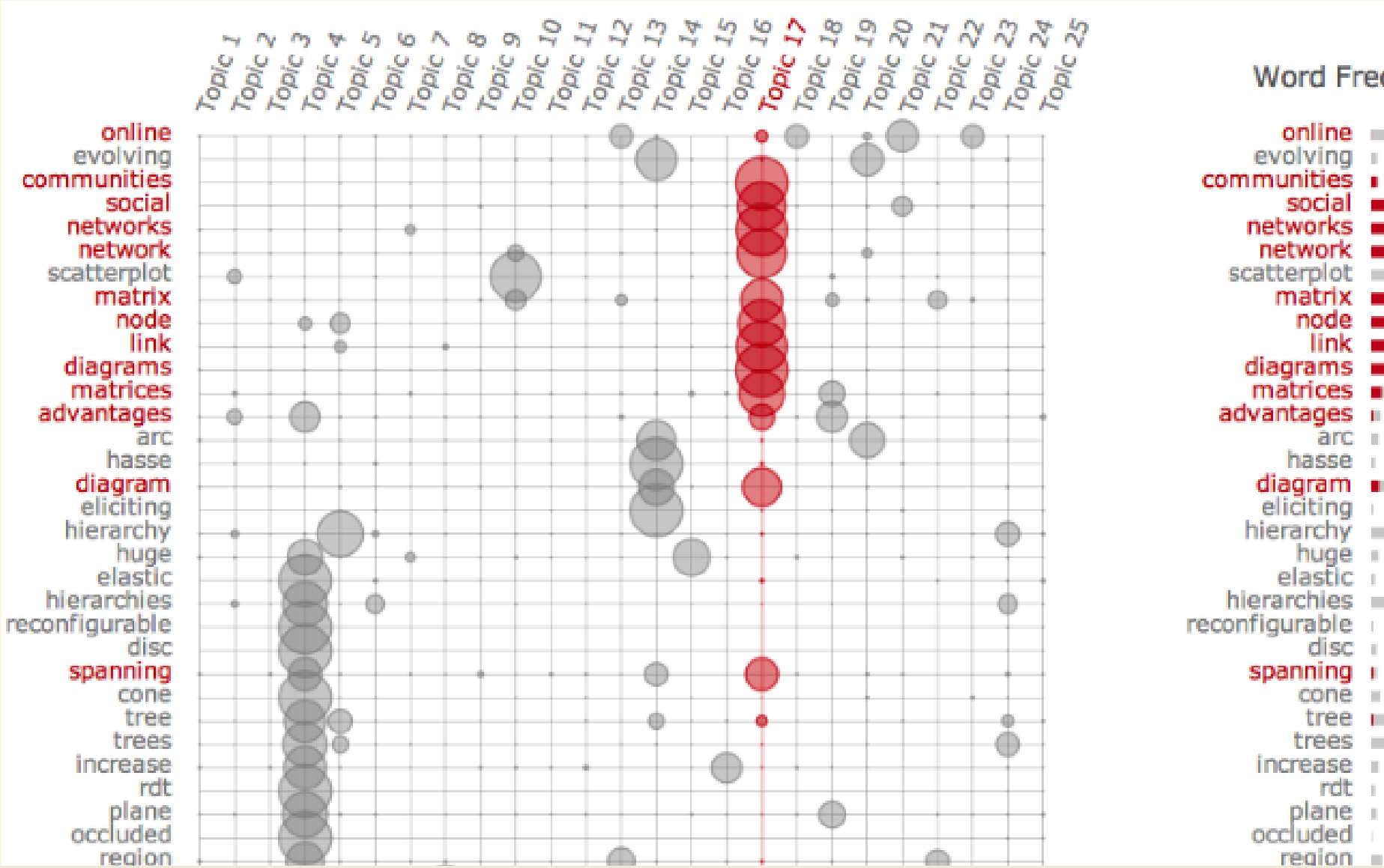
population studied	“all social media users of a town”
time spans	“all of the Post-Soviet history”
geographical scope	“all educational migration in Russia”

Course goals

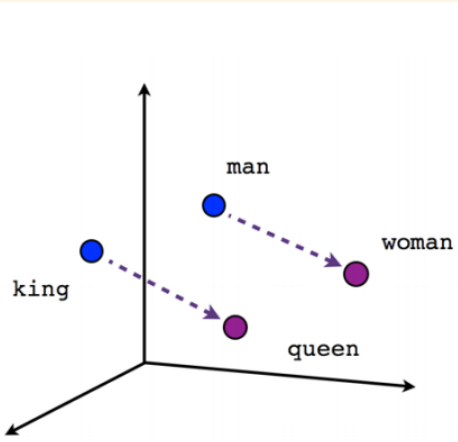
- provide basic understanding of how to properly use collections of texts as quantitative evidence,
- and to make this knowledge practical

Course content

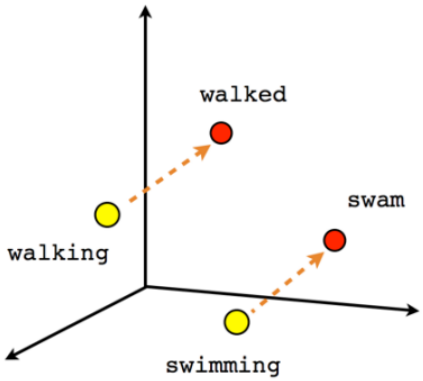
Bread and butter: Topic modeling



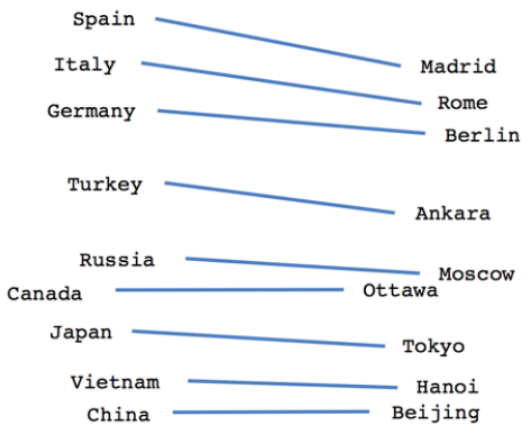
Killer feature: Word embeddings



Male-Female

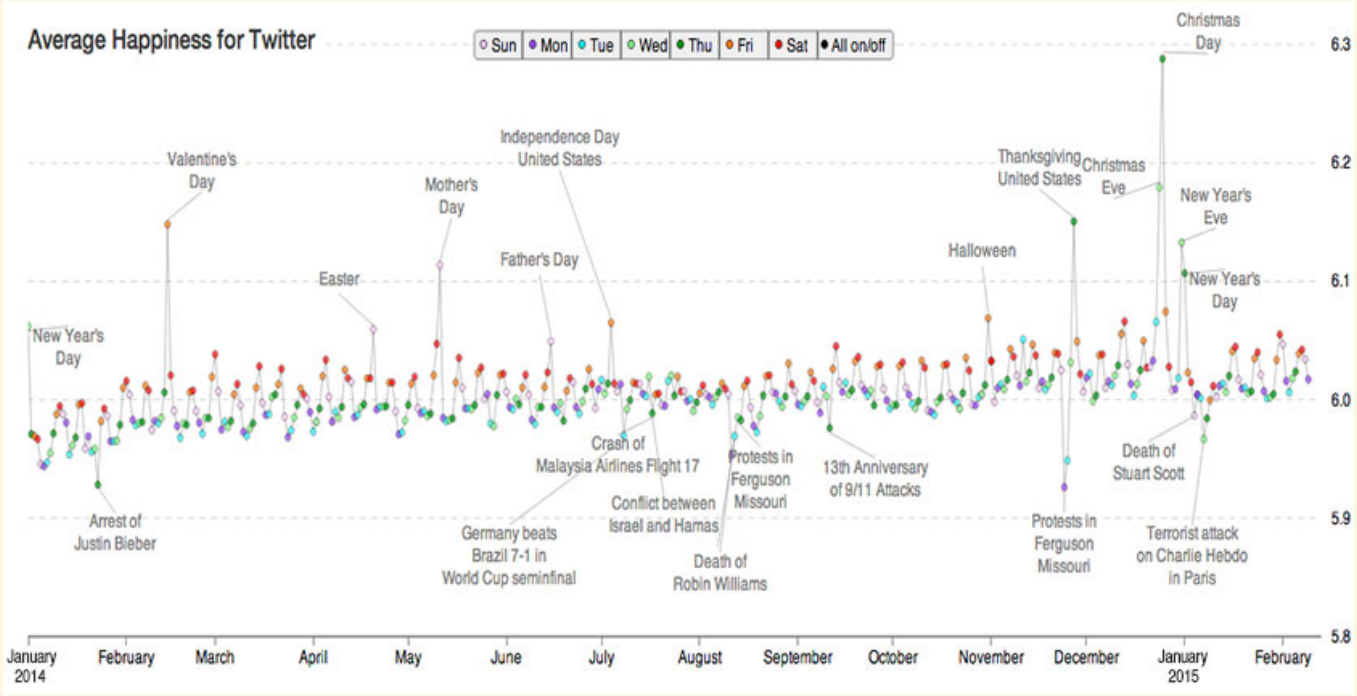


Verb tense



Country-Capital

The icing on the cake: Sentiment analysis



The icing on the cake: Sentiment analysis

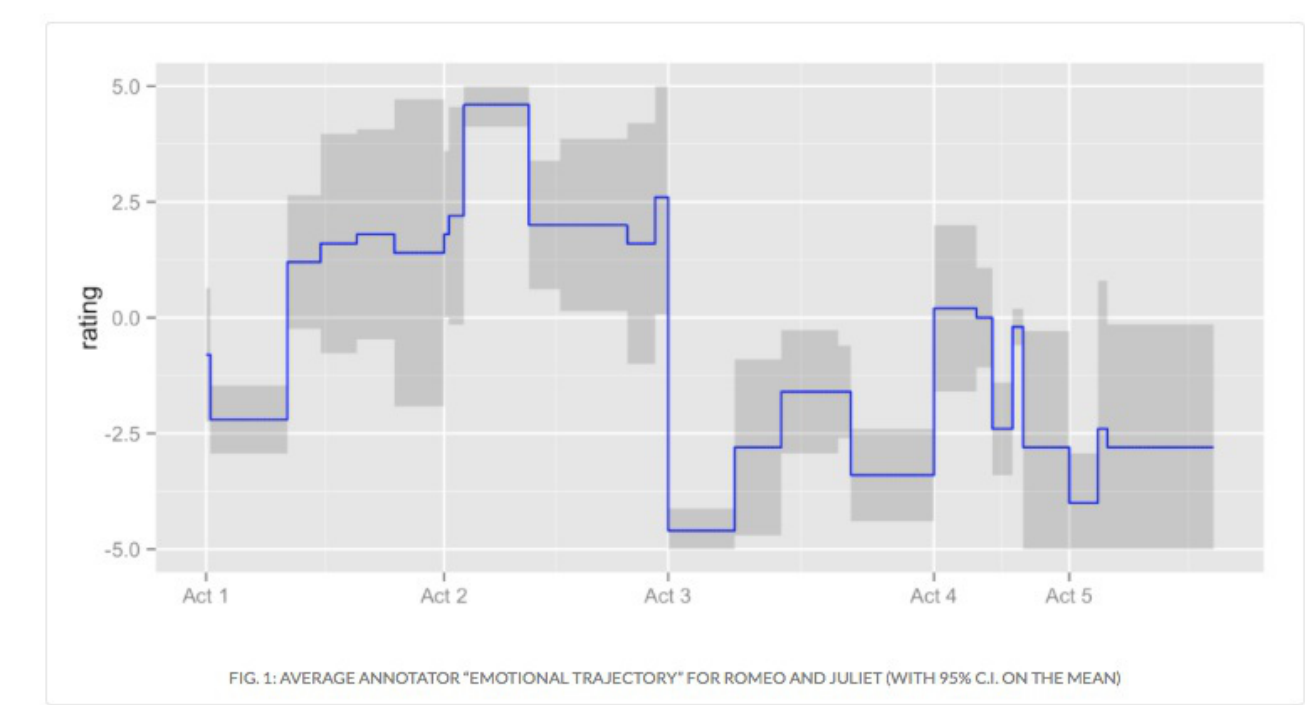
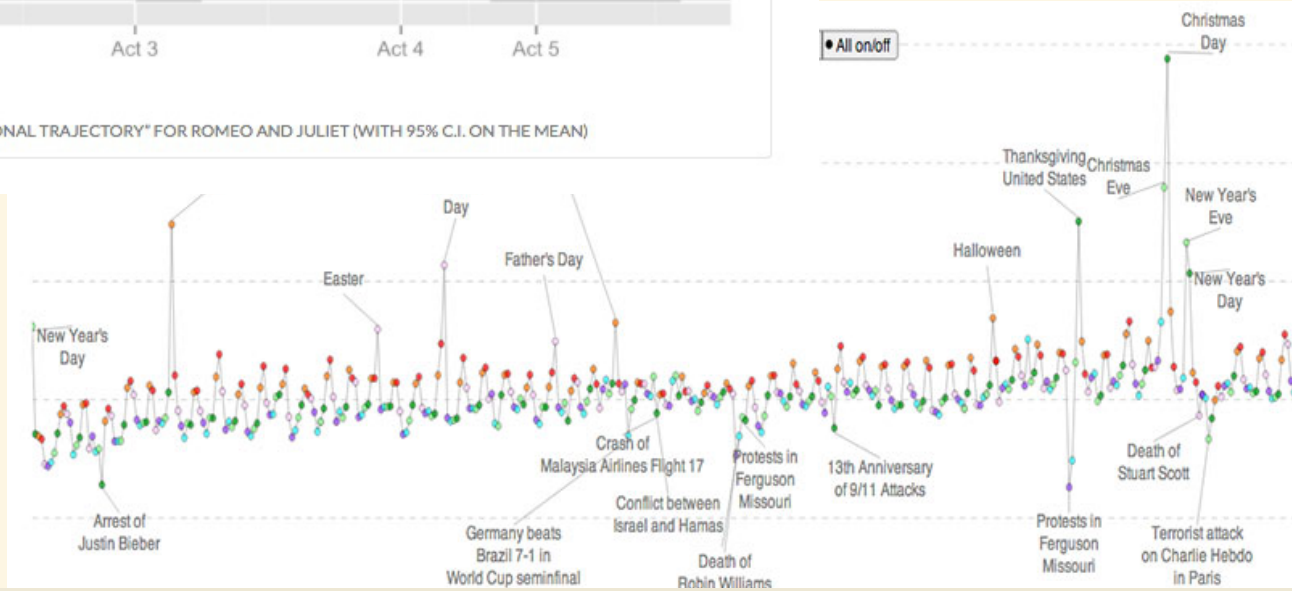
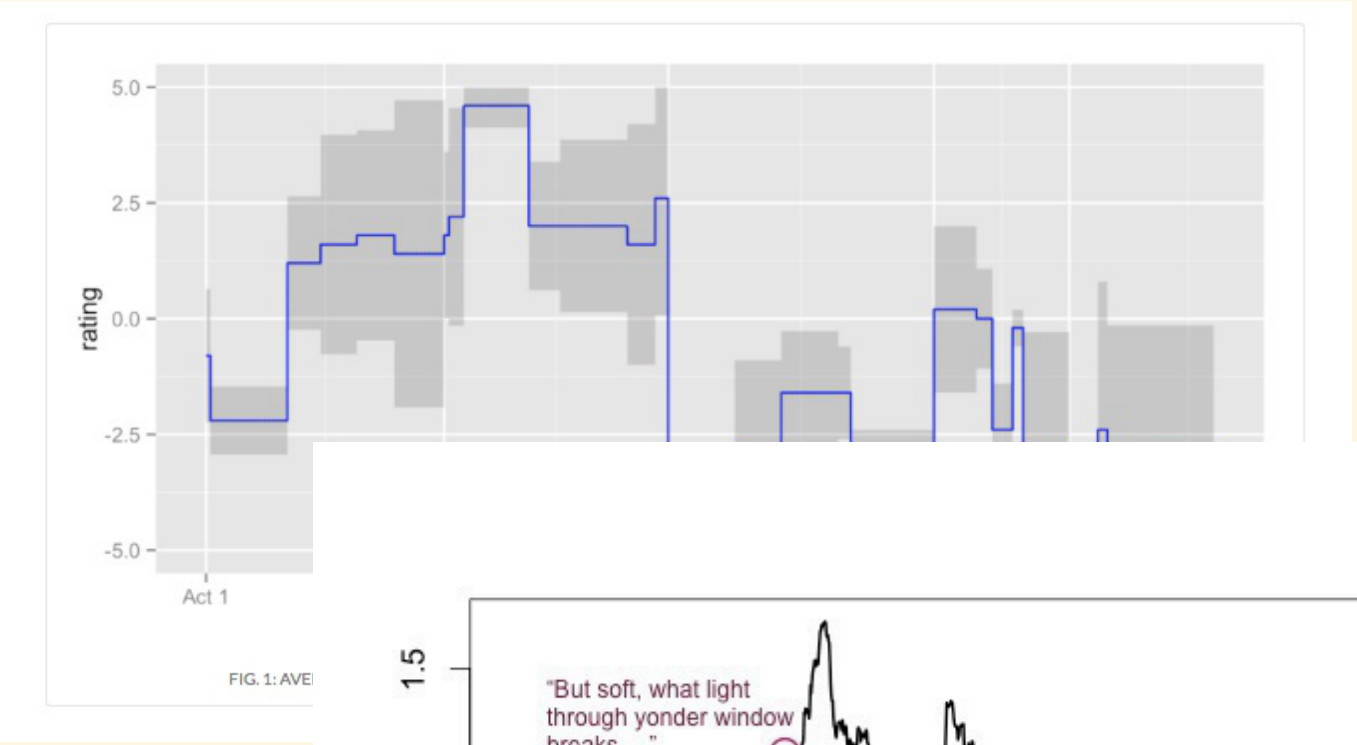
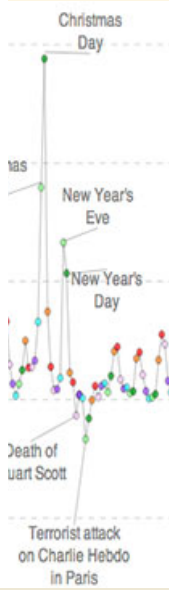


FIG. 1: AVERAGE ANNOTATOR "EMOTIONAL TRAJECTORY" FOR ROMEO AND JULIET (WITH 95% C.I. ON THE MEAN)



The icing on the cake: Sentiment analysis





Course topics

- Basic word statistics:
  - lexical statistics (word frequency distributions),
  - distributive semantics (word co-occurrence patterns),
  - vector representation of text.
- Methods for supervised and unsupervised modeling:
  - dictionary methods,
  - document classification and clusterization,
  - topic modeling,
  - word embeddings,
  - sequence modeling.
- Applied tasks:
  - automating content analisys (extracting theme and topic),
  - sentiment analysis,
  - information extraction from unstructured text.

What to expect

How coursework will be organized

- An interesting recent article
- with an explanation of the necessary concepts and methods during lecture
- followed by detailed analysis of the method in class
- concluded by the task to reproduce the method with your own data

Expectations

Practical work with real texts in class and at home.

- command line
- mining your own text collection
- R scripts
- bugs in scripts, googling, bugs in scripts again
- seeking and getting help from your peers and course instructor
- happy end

Work in groups

What you can learn

- State-of-the-art of natural language processing:
  - solved problems
  - topical issues and unsolved problems
- Terms:
  - a minimal vocabulary of necessary linguistic terms (with meanings! :))
  - appropriate keywords to search for current research and tools
- Tools:
  - Where to apply methods for computational text analysis and how to interpret their results
  - Existing software for text analysis (for Russian and English)
  - Existing linguistic resources — dictionaries, corpora, pre-trained models (for Russian and English)