An approach to visual thesaurus exploration: a case study for Russian language

Слайды на русском во второй половине файла

cand. sc., assistant professor at machine learning and knowledge representation lab, Innopolis University

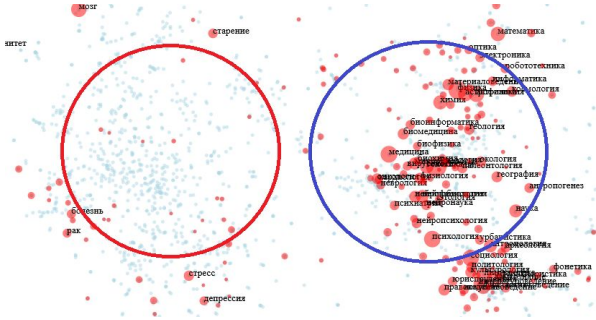
Stanislav Protasov

Agenda

- Big text dataset visualization problem
- Overview of approaches to visualization
- Solution schema
- Results and discussion

Problem statement

- 6000+ popular science items (text)
- Questions to the content:
  - What is the distribution of items in knowledge areas?
  - Which knowledge areas are missing or poorly covered?
  - Are there gaps within one knowledge area?
  - What can be an approach to building a recommender system?
  - Any interesting insights?



Approaches to visualization

- Semantic modelling
  - Graph-based vocabulary representation (WordNet)
  - Vector space model (word2vec, transformers, GloVe, fastText)
    - (ANN) models
    - **Precomputed embeddings (rusvectores, natasha/navec)**

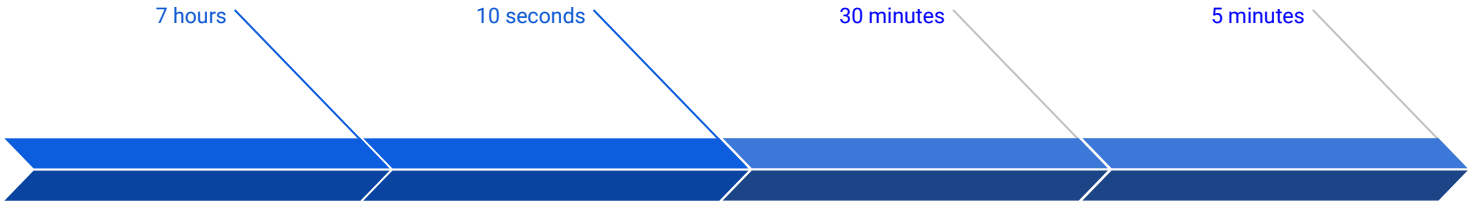
Approaches to visualization

- Dimensionality reduction
  - Global methods
    - PCA minimizes data variance loss (globally)
    - Random Projections preserves pairwise Euclidean distances (globally)
    - Pivot-Mapping allows false positive matches (for search)
  - Local methods
    - Self-organizing maps and elastic maps tunes metric locally
    - **t-SNE uses perplexity to manage “neighbourhood” size**

Approaches to visualization

- Background data:
  - National corpora of languages (НКРЯ, COCA)
  - Domain-specific corpora (Wikipedia dumps)

Solution schema



Thesaurus extraction

Use Kullback–Leibler divergence to search for anomalies in language model

Vector representation

Use *RusVectores* fastText pre-trained embeddings

Dimensionality reduction

Background vocabulary + Extracted thesaurus

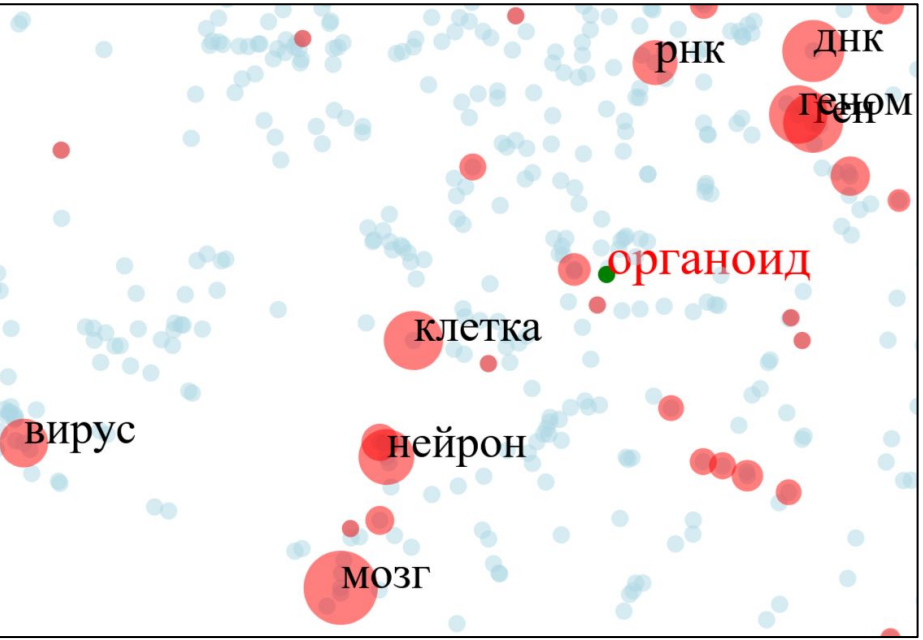
t-SNE: 3000 iterations, P=40

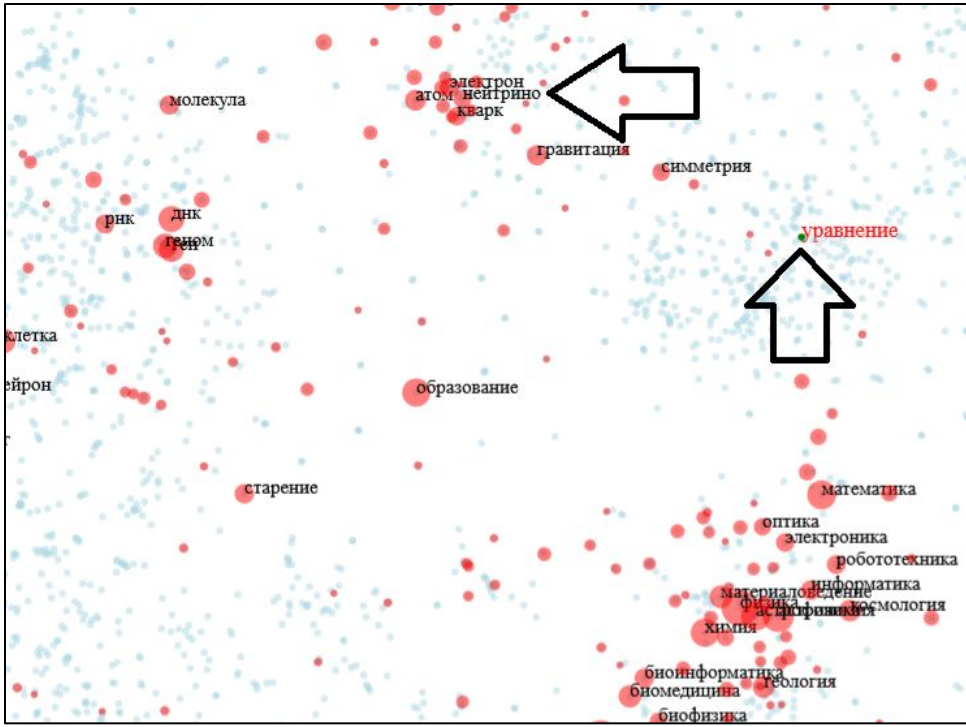
Interactive visualization

HTML + SVG + css

$$D(term) = P_{dataset}(term) * \log\left(\frac{P_{NCRL}(term)}{P_{dataset}(term)}\right)$$

Results


рнк
днк
геном
ген
органоид
клетка
вирус
нейрон
мозг


молекула
электрон
атом
нейтрино
кварк
гравитация
симметрия
рнк
днк
геном
ген
уравнение
клетка
нейрон
образование
старение
математика
оптика
электроника
робототехника
информатика
материаловедение
физика
астрофизика
космология
химия
биоинформатика
биомедицина
биология
биофизика

Discussion

- “Good” topic clusters, but clustering principles are different (“surnames”, “F1 racers”)
- Using raw wikipedia dumps with no filtering tends to create “garbage” clusters (“surnames”)

Contact me

Stanislav Protasov

s.protasov@innopolis.ru

@sprotasov

Innopolis University

Подход к визуализации тезауруса для его исследования на примере русского языка

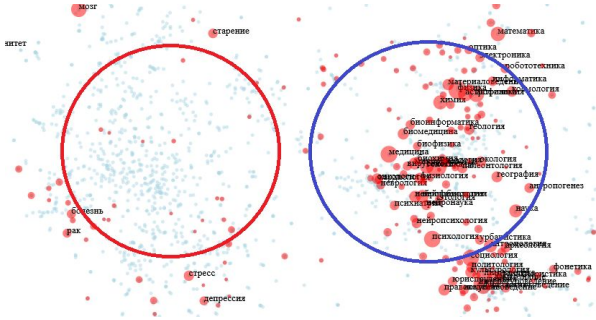
к.ф-м.н, доцент лаборатории машинного обучения и представления данных Университета Иннополис
Станислав Протасов

План презентации

- Постановка задачи визуализации
- Обзор подходов
- Схема решения
- Результаты

Постановка задачи

- Коллекция 6000+ научно-популярных единиц контента (текст)
- Вопросы, касающиеся контента в целом:
  - Как распределены материалы по областям знаний?
  - Какие области знаний отсутствуют в покрытии или плохо представлены?
  - Есть ли в рамках одной области знаний пробелы в покрытии?
  - Как построить рекомендательную систему на базе?
  - Что там ещё интересного?



Подходы к визуализации

- Моделирование семантики
  - Графовое представление словаря (WordNet)
  - Векторное представление (word2vec, transformers, GloVe, fastText)
    - Модели
    - **Предпосчитанные словари (rusvectores, natasha/navec)**

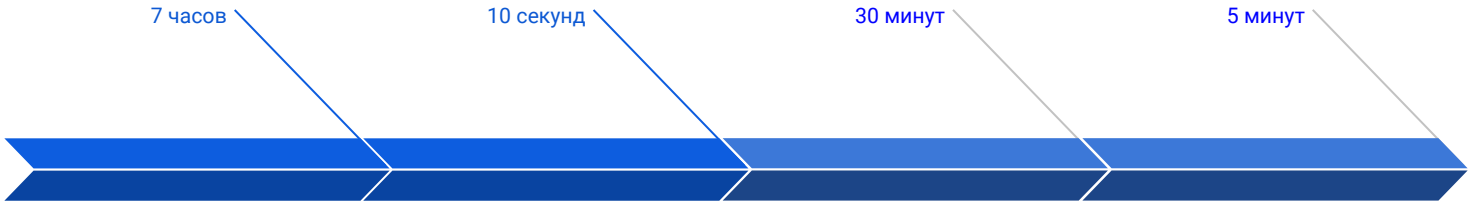
Подходы к визуализации

- Понижение размерности
  - Глобальные методы
    - PCA минимизирует потерю дисперсии в данных
    - Random Projections предохраняет попарное Евклидово расстояние (в целом)
    - Pivot-Mapping допускает ложноположительные срабатывания
  - Локальные методы
    - Карты Кохонена и упругие карты адаптируют метрику локально
    - **t-SNE использует перплексию для того, чтобы регулировать размер окрестности**

Подходы к визуализации

- Фон:
  - Национальные корпусы языков (НКРЯ, COCA)
  - Доменные корпусы (Wikipedia dumps)

Схема решения



Извлечение тезауруса

Используем дивергенцию Кульбака-Лейблера для поиска аномалий относительно языковой модели НКРЯ

Векторное представление

Используем RusVectores fastText предпосчитанные векторные представления

Понижение размерности

Фоновый словарь + целевой тезаурус

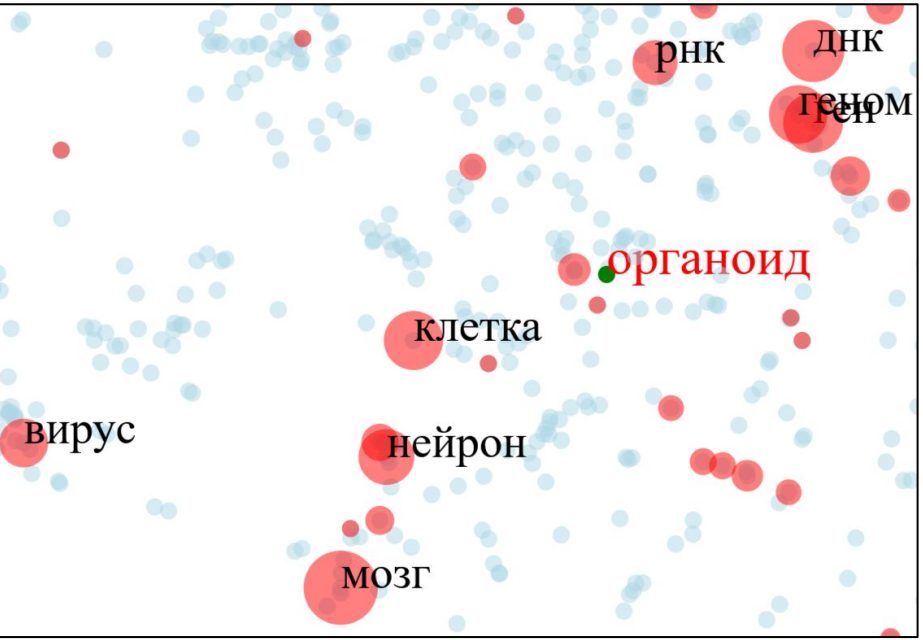
tSNE, 3000 итераций, P=40

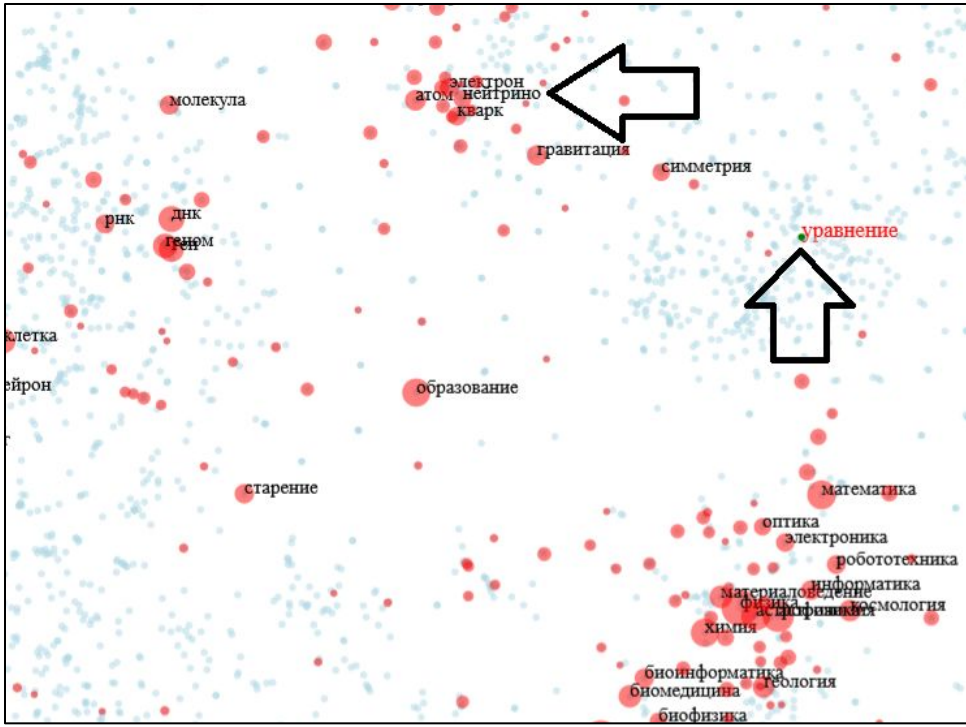
Отображение с поддержкой интерактивности

HTML + SVG + css

$$D(term) = P_{dataset}(term) * \log\left(\frac{P_{NCRL}(term)}{P_{dataset}(term)}\right)$$

Результаты


рнк
днк
геном
ген
органоид
клетка
вирус
нейрон
мозг


молекула
электрон
атом
нейтрино
кварк
гравитация
симметрия
рнк
днк
ген
уравнение
образование
старение
математика
оптика
электроника
робототехника
информатика
материаловедение
космология
химия
биоинформатика
биомедицина
биофизика

Обсуждение

- “Хорошие” тематические кластера, но с разным признаком кластеризации
- Использование wikipedia dumps в чистом виде без фильтрации приводит к образованию “мусорных” кластеров (“фамилии”)

Контакты

Станислав Игоревич Протасов

s.protasov@innopolis.ru

@sprotasov

Университет Иннополис