Digital Humanities: подходы, методы, перспективы

Борис Орехов
НИУ ВШЭ, ИРЛИ РАН, МГУ

План разговора

- Специфика Digital Humanities
- История и современное состояние
- Методы анализа текста
- Русскоязычное сообщество
- Проекты и инфраструктура

Digital Humanities — зонтичное название, научный поиск на стыке гуманитарных (иногда социальных) наук и компьютерного анализа данных.

Специфика дисциплины

Перевод на русский язык

- Термину Digital Humanities не повезло с переводом на русский язык: нет эквивалента для слова humanities.
- Наиболее адекватным является длинный перевод из двух слов «гуманитарные науки».
- Сокращение находится в неудачном слове «гуманитаристика».
- Стилистически оно не нейтрально, создает сниженный эффект: -истика («шагистика», «ерундистика»).
- Так не воспринимаются давно заимствованные слова «журналистика», «логистика».
- Отдельно от «цифровой» это слово в русском языке почти не употреблялось.

Цифровые гуманитарные исследования: монография. Красноярск: СФУ, 2023. С. 5—6

Запрос на точность и объективность

Естественные науки узурпировали понятие научности.

К гуманитариям предъявляются требования, которым они никогда не отвечали.

Точность и объективность — мифы, которые невозможны и в естественных науках.

Digital Humanities тоже не могут их обеспечить.

Субъективность настигает нас на этапах постановки задачи, отбора материала и метода, но особенно — интерпретации материала.

Data и capta

Для цифрового исследования нужны данные

В отличие от представителей естественных наук, гуманитарии никогда не имеют полных данных.

Их наборы всегда ограничены (обстоятельствами, интересами исследователя).

Поэтому уместно говорить не о «данных», а «взятых».

Володин А. Ю. Между data и capta: проблемы датафикации исторических исследований // Вестн. Перм. ун-та. Сер. История. 2019. №3 (46). URL: https://cyberleninka.ru/article/n/mezhdu-data-i-capta-problemy-datafikatsii-istoricheskih-issledovaniy

Зачем же тогда DH?

DH позволяет измерить, калибровать то, что уже было известно.

Наиболее осмысленно на больших объемах (см. «великое непрочтенное»), где наиболее выражены надличностные факторы и параметры

История и современность

Оцифровка

До 2000-х годов цифровые гуманитарные науки почти целиком состояли из оцифровки (разработка стандартов, практика электронного пр

TEI


```
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="rus">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title type="main">Ревизор</title>
        <title type="main" xml:lang="eng">The Government Inspector</title>
        <title type="sub">Комедия в пяти действиях</title>
        <title type="sub" xml:lang="eng">A Comedy in Five Acts</title>
        <author key="wikidata:Q43718">Гоголь, Николай Васильевич</author>
      </titleStmt>
      <publicationStmt> [10 lines]
      <sourceDesc>
        <bibl type="digitalSource">
          <name>Интернет-библиотека Алексея Комарова</name>
          <idno type="URL">http://ilibrary.ru/text/473/index.html</idno>
          <availability status="free">
            <p>In the public domain.</p>
          </availability>
          <bibl type="originalSource">
            <title>Н. В. Гоголь. Собрание сочинений в 9 т. Т. 4. М.: Русская книга, 1994.</title>
            <date type="print" when="1836">"Дата первой публикации: 1836" (Wikipedia)</date>
            <date type="premiere" when="1836">"Первые представления шли в первой редакции 1836
              года." (Wikipedia)</date>
            <date type="written" when="1835">"1835 г." (ilibrary)</date>
          </bibl>
        </bibl>
      </sourceDesc>
    </fileDesc>
    <profileDesc>
      <particDesc>
        <listPerson>
          <person xml:id="gorodnichij" sex="MALE">
            <persName>Городничий</persName>
            <persName xml:lang="ger">Polizeimeister</persName>
          </person>
          <person xml:id="ammos_fedorovich_ljapkin_tjapkin" sex="MALE">
            <persName>Аммос Федорович</persName>
            <persName xml:lang="ger">Richter</persName>
          </person>
```

Анализ данных

Со второй половины 2000-х к оцифрованным данным начинают применяться методы автоматической обработки и статистического анализа.

История дисциплины:

Володин А. Ю. Цифровые гуманитарии: от академических племен к эпистемическому сообществу // Цифровые гуманитарные исследования. 2025. № 2. С. 84–116 DOI: 10.31860/cgi-2025-2-84-116

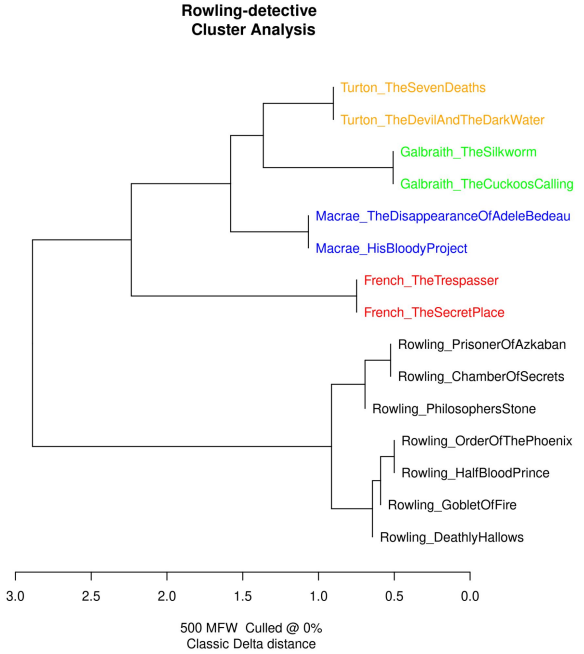
компьютерный анализ

текста

Стилеметрия и атрибуция

Методы автоматического

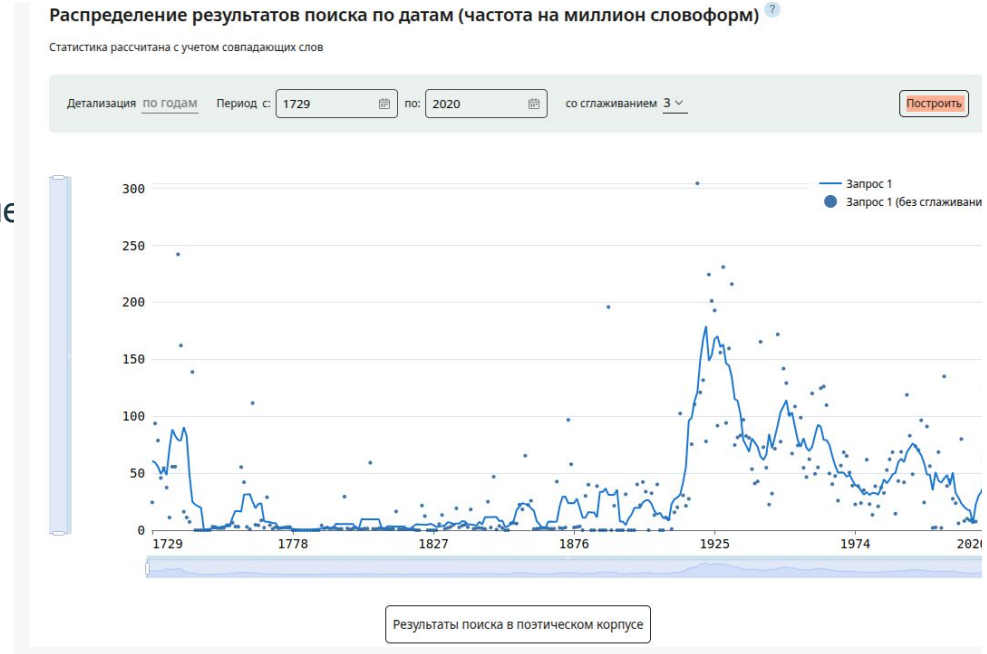
определения авторства



Культуромика

Частотность слов в связи с эволюцие

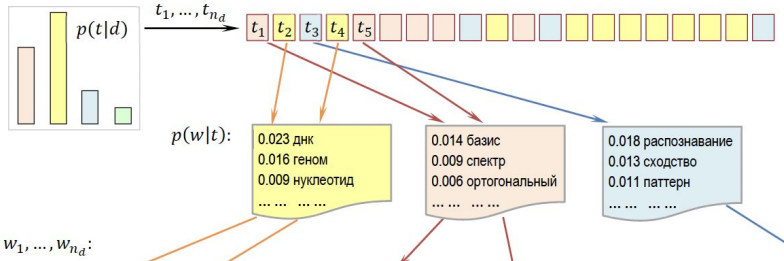
общественного сознания



Тематическое моделирование

Способ узнать, о чем текст,

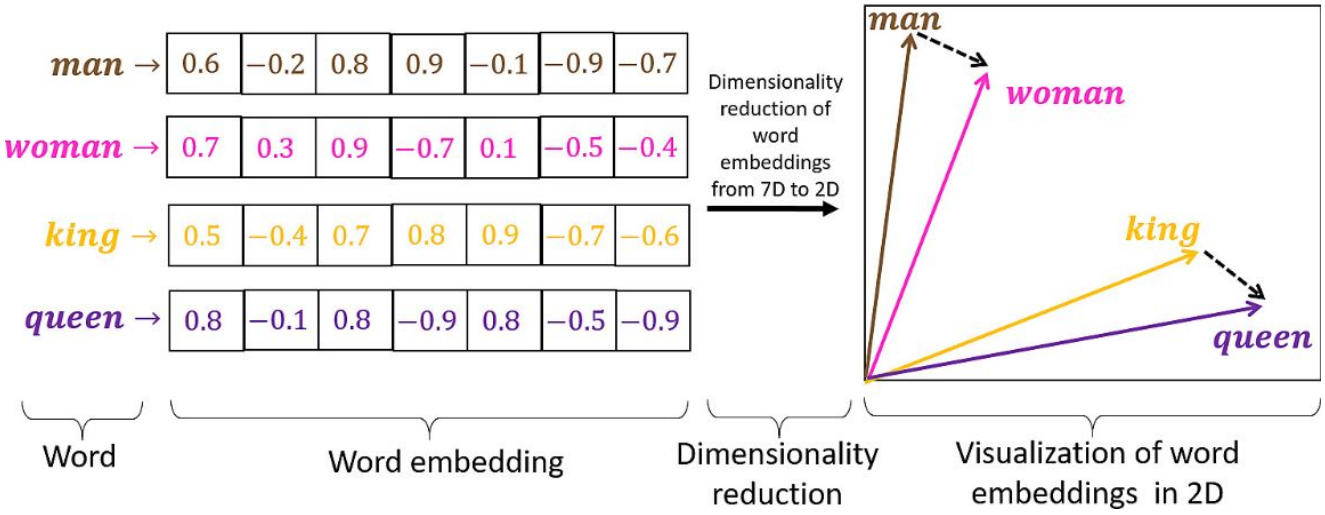
не читая его



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Векторные модели

Способ вычислить семантику слова



Глава в монографии

Цифровые гуманитарные исследования: монография /

А. Б. Антопольский, А. А. Бонч-Осмоловская,

Л. И. Бородкин [и др.]. — Красноярск: Сиб. федер. ун-т,

2023. — 272 с.

глава 6. Компьютерный анализ текста, С. 120–157

Русскоязычное сообщество

Центры

НИУ ВШЭ: магистерская программа, https://github.com/nevmenandr/awesome-dh-hse

ИТМО: магистерская программа, https://t.me/dhcenter

УрФУ, СФУ, МГУ (историческая информатика), ИРЛИ РАН, https://t.me/tozhe_nauka

DH CLOUD, https://t.me/dhcloud

Издания

Журнал «Цифровые гуманитарные исследования» (с 2024 г. 2 раза в год)

Монографии:

[Библиотека]

Конференции

«СПИЛ» (Смоленск), «Информационные технологии в гуманитарных исследованиях» (Красноярск), ИТМО (Санкт-Петербург), конференция Ассоциации «История и компьютер» (АИК, Москва).

Регулярный семинар «Цифровая среда» под руководством А. Ю. Володина: dhri.ru/projects/sreda

Памятка для наших студентов: nevmenandr.github.io/portfolio/assets/pdf/memo_public_hse.pdf

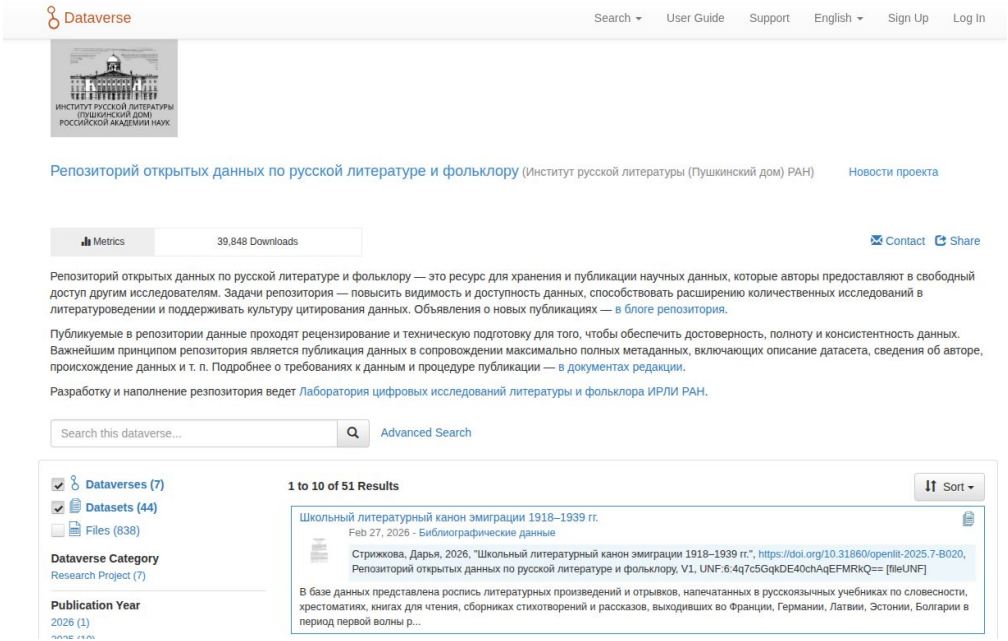
Об истории русскоязычной науки, использующей точные методы в гуманитарной сфере:

Володин А. Ю., Орехов Б. В. Digital Humanities в России и конец истории // Цифровые гуманитарные исследования. 2024. № 1. С. 63—85. DOI: 10.31860/cgi-2024-1-63-85

Проекты и инфраструктура

Репозиторий открытых данных

https://dataverse.pushdom.ru/



Пакеты для языка Python

- Модуль для транслитерации старой орфографии в новую https://pypi.org/project/prereform2modern/
- Модуль для акцентуации русского поэтического текста https://pypi.org/project/ru-accent-poet/
- Модуль для вычленения прямой речи персонажей в художественном тексте https://pypi.org/project/direct-speech-extractor-ru/
- Модуль для оценки формульности фольклорного текста https://pypi.org/project/formularity-rfs/
- Модуль для преобразования текстов в формате TEI https://pypi.org/project/TEItransformer/

Модель для задач OCR при работе с текстами в старой орфографии

- OCR-модель для распознавания текстов в старой орфографии: https://huggingface.co/Serovvans/trocr-prereform-orthography

https://dhcloud.org/python/

Орехов Б. В. Открытые компьютерные инструменты для решения задач оцифровки и анализа русскоязычного текста в области Digital Humanities // Цифровые гуманитарные исследования. 2025. № 2. С. 71–83.

Мои проекты

Параллельный корпус переводов «Слова о полку Игореве»: http://nevmenandr.net/slovo/

91-й том: указатель к ПСС Л. Н. Толстого: http://index.tolstoy.ru/

Сеть персоналий в Словаре русских писателей XVIII века: https://nevmenandr.github.io/rus-dict18-persons/

Русский стилеметрический датасет: https://github.com/nevmenandr/RSD/

Векторные романы: https://nevmenandr.github.io/novel2vec/

@SCHONENREDE

https://nevmenandr.github.io/portfolio/
https://t.me/schonenrede