Олимпиада студентов и выпускников «Высшая лига»– 2020 г.

Направление «Фундаментальная и прикладная лингвистика»

Профиль: «Цифровые методы в гуманитарных науках»

Время выполнения задания – 180 мин., язык - русский.

Максимальное количество баллов - 100.

Вопрос 1. КОД 313

Прочтите отрывок статьи Кристофа Шёха «Big? Smart? Clean? Messy? Data in the Humanities»

Дайте развернутые ответы на следующие вопросы:

1. Чем, по мнению автора, отличаются данные в гуманитарных науках от прочих данных? В чем сложность работы с ними? Приведите свои примеры
2. Какие две разновидности «гуманитарных данных» предлагает автор? В чем различия между ними? Приведите свои примеры обеих разновидностей. В чем преимущества и недостатки каждой разновидности?
3. В чем проблема с самим словом *data* и его этимологией? Какое свойство данных плохо отражается через исходное латинское значение этого слова?
4. Объясните утверждение: «*Trends in literature can be observed across the entire literary production of a given time and given genre*». Что имеет в виду автор? Что позволяет ему делать такое утверждение?
5. Объясните утверждение: «*Data (as well as the tools with which we manipulate them) add complexity to the relation between researchers and their objects of study*». Что имеет в виду автор?
6. Как можно понимать термин «репрезентативность» в контексте цифровых гуманитарных исследований и проблематики big data in the humanities?

Big? Smart? Clean? Messy? Data in the Humanities

<…> *Data in the humanities is a bit special: one could in fact argue that text in a book or a manuscript, or the visual elements making up a painting, are data already. First, however, this is analog, non-discrete data, which cannot be analyzed or transformed computationally; and second, language, texts, paintings, and music are semiotic systems that have dimensions beyond the physically measurable, dimensions which depend on semantics and pragmatics, that is on meaning in context. For this latter reason particularly, speaking of "data" in the humanities is problematic and has been challenged. Criticism has come from mainstream scholars who see "data" and quantitative methods of analyzing them with suspicion, because the apparent empiricism of data-driven research in the humanities seems at odds with principles of humanistic inquiry, such as context-dependent interpretation and the inevitable "situated-ness" of the researchers and their aims.*

Some practitioners of digital humanities, notably Joanna Drucker, have argued that the term "data" is actually inadequate. And indeed, the term's etymology seems problematic in the

context of the humanities: it comes from the Latin datum, which means "that which is given." This means it carries with it the meaning of an observer-independent fact which cannot be challenged in itself. Johanna Drucker prefers to speak of "capta" instead of data, literally "that which has been captured or gathered", underlining the idea that even the very act of capturing data in the first place is oriented by certain goals, done with specific instruments, and driven by a specific attention to a small part of what could have been captured given different goals and instruments. In other words, capturing data is not passively accepting what is given, but actively constructing what one is interested in.

Similarly, Digital Archivist Trevor Owens has argued that data is not a given, but is always manufactured and created. Moreover, he shows, we can approach data from different perspectives and treat it as an artifact (something actively and purposefully created by people), as text (subject to interpretation, for example by scholars), and as computer-processable information (to be analysed with quantitative methods). According to Owens, this means that data is not a given and not some unquestionable evidence; rather, it is "a multifaceted object which can be mobilized as evidence in support of an argument."

Even without using a new term, we can now redefine what we mean by data in the humanities. Data in the humanities could be considered a digital, selectively constructed, machine-actionable abstraction representing some aspects of a given object of humanistic inquiry. Whether we are historians using texts or other cultural artifacts as windows into another time or another culture, or whether we are literary scholars using knowledge of other times and cultures in order to construct the meaning of texts, digital data add another layer of mediation into the equation. Data (as well as the tools with which we manipulate them) add complexity to the relation between researchers and their objects of study.

Basically, I would like to argue that there are two core types of data in the humanities: big data and smart data. These two types of data can be described in two dimensions: the first dimension describes how structured, clean, and explicit the data is; the second dimension describes how voluminous and how varied the data is. I suggest to view big data, in a first approximation, as relatively unstructured, messy and implicit, relatively large in volume, and varied in form. Conversely, I suggest to view smart data to be semi-structured or structured, clean and explicit, as well as relatively small in volume and of limited heterogeneity. Although you could say that these are really just differences of degree, there are more fundamental differences between them when it comes to looking at how each of them are created or captured, modeled, enriched, and analyzed.

2. Smart data (in the humanities)

<...>

First of all, I should mention that "smart data" is not an established or well-defined term. It is not very widespread and does not have a stable meaning. Smart data is data that is structured or semi-structured; it is explicit and enriched, because in addition to the raw data, it contains markup, annotations and metadata. And smart data is "clean", in the sense that imperfections of the process of capture or creation have been reduced as much as possible, within the limits of the specific aspect of the original object being represented. This also means that smart data tends to be "small" in volume, because its creation involves human agency and demands time. The process of modeling the data is essential to small/smart data; its abstract structure can be defined with elaborate schemas or as predefined database structures.

A prototypical example of smart data are scholarly digital editions produced using the Guidelines of the Text Encoding Initiative. Technically, TEI documents are usually considered semi-structured; usually, they follow a data model expressed in a schema, but such schemas allow for considerable flexibility. In addition to a very clean transcription of the text, digital editions using TEI can make a lot of information explicit: first of all, TEI files contain not just the full text, but also metadata associated with the text (in the teiHeader section); also, the data is structured and explicit: there is markup making the structure of the text explicit, identifying parts, chapters, headings, paragraphs, as well as page and line breaks, for example. Finally, many more types of information can be specified: for example person names in a novel or play, place names in a letters or documents, and many more things; and links to other parts of the documents and to external documents. Making all of these things explicit allows to visualize them in specific ways and to index, count and analyze them computationally.

3. Big data (in the humanities)

<...> big data in the humanities is not the same as big data in the natural sciences or in economics. In most cases, velocity does not play a key role in big humanities data right now. Also, the large "volume" is less usefully defined in the humanities by a shift from databases to distributed computing. Variety of formats, complexity or lack of structure does come into play, however. In fact, the distinctive mark of big data in the humanities seems to be a methodological shift rather than a primarily technological one. And it is a huge methodological shift. Paradoxically, the shift from small smart data to big data is much more radical, I would argue, than the shift from print to smart digital data was. Indeed, moving from smart data to big data implies a shift from "close reading" to "distant reading" (in the words of Franco

Moretti) or to "macroanalysis" (to use Matthew Jockers' term). In this paradigm, instead of reading a few selected texts, we analyze an entire collection of relevant textual data.

The first consequence of the macroanalytic paradigm in the humanities, where hundreds or even thousands of texts are analyzed at a time, is that instead of operating on the level of literary forms and conventions, of semantics and context, we operate with quantitative measures of low-level features, on the basis of statistics and probabilities. The second consequence is that instead of so-called "representative" texts or paintings, we can now study the entire set of texts or images relevant to a specific research question. Trends in literature can be observed across the entire literary production of a given time and given genre. Questions of representativeness, of canonization, of literary quality play a much smaller, or at least a different, role in this context.

Вопрос 2.
Решите задачу.
Буквы Т, Л, А, А, Х, Ъ написаны на отдельных карточках. Илья Ильич берет карточки в случайном порядке и прикладывает одну к другой. Какова вероятность, что Илья Ильич с первой попытки соберет из них слово «ХАЛАТЪ»?

Вопрос 3.
Решите задачу.
У царя Кащея было 24 мешка с золотом одинакового веса. Известно, что баба Яга залезла в один из мешков и подменила часть золота грецкими орехами. На вид мешки не отличаются, но мешок с грецкими орехами легче. У Кащея есть только простые равноплечные весы без делений, с их помощью можно понять, что легче, а что тяжелее. Вместительность весов неограниченна.

Какое минимальное количество взвешиваний понадобится, чтобы гарантированно найти мешок с орехами?

Вопрос 4.
Вы с друзьями решили сделать цифровой корпус текстовых подписей к любительским фотографиям. Каким фотографиям? Каковы критерии вхождения в корпус? Все это решать вам. Опишите:

1. где и каким образом вы бы предложили собрать материал для вашего исследования; какие компьютерные технологии могут быть при этом применены?
2. какие типы разметки можно было бы предложить для собранного вами материала? проявите максимум фантазии и предложите как можно больше уровней разметки
3. какого рода метаинформация о ваших объектах исследования вам может понадобиться?

4. предложите не менее трех сценариев количественных исследований на базе вашего материала. В каждом случае пропишите цель исследования и необходимые шаги.