Олимпиада для студентов и выпускников – 2020 г.

Профиль: “Цифровые методы в гуманитарных науках”

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

Вопрос 1.

Прочтите отрывок статьи Теда Андервуда “New methods need a new kind of conversation”
Дайте развернутые ответы на следующие вопросы:

1. Какие изменения предлагает внедрить автор и зачем?
2. На какие сложности такого внедрения указывает автор?
3. Согласны ли вы с предложением автора? Почему? Опишите плюсы и минусы.
4. Объясните название третьего раздела статьи («Reproducibility is great, but replication is the real point»). Что имеет в виду автор?

New methods need a new kind of conversation

Over the last decade, the (small) fraction of articles in the humanities that use numbers has slowly grown. This is happening partly because computational methods are becoming flexible enough to represent a wider range of humanistic evidence. We can model concepts and social practices, for instance, instead of just counting people and things.
That's exciting, but flexibility also makes arguments complex and hard to review. Journal editors in the humanities may not have a long list of reviewers who can evaluate statistical models. So while quantitative articles certainly encounter some resistance, they don't always get the kind of detailed resistance they need. I thought it might be useful to stir up conversation on this topic with a few suggestions, aimed less at the DH community than at the broader community of editors and reviewers in the humanities. I'll start with proposals where I think there's consensus, and get more opinionated as I go along.

1. Ask to see code and data.
Getting an informed reviewer is a great first step. But to be honest, there's not a lot of consensus yet about many methodological questions in the humanities. What we need is less strict gatekeeping than transparent debate.

As computational methods spread in the sciences, scientists have realized that it's impossible to discuss this work fruitfully if you can't see how the work was done. Journals like Cultural

Analytics reflect this emerging consensus with policies that require authors to share code and data. But mainstream humanities journals don't usually have a policy in place yet.

Three or four years ago, confusion on this topic was understandable. But in 2018, journals that accept quantitative evidence at all need a policy that requires authors to share code and data when they submit an article for review, and to make it public when the article is published.

I don't think the details of that policy matter deeply. There are lots of different ways to archive code and data; they are all okay. Special cases and quibbles can be accommodated. For instance, texts covered by copyright (or other forms of IP) need not be shared in their original form. Derived data can be shared instead; that's usually fine. (Ideally one might also share the code used to derive it.)

2. … especially code.

Humanists are usually skeptical enough about the data underpinning an argument, because decades of debate about canons have trained us to pose questions about the works an author chooses to discuss.

But we haven't been trained to pose questions about the magnitude of a pattern, or the degree of uncertainty surrounding it. These aspects of a mathematical argument often deserve more discussion than an author initially provides, and to discuss them, we're going to need to see the code.

I don't think we should expect code to be polished, or to run easily on any machine. Writing an article doesn't commit the author to produce an elegant software tool. (In fact, to be blunt, "it's okay for academic software to suck.") The author just needs to document what they did, and the best way to do that is to share the code and data they actually used, warts and all.

3. Reproducibility is great, but replication is the real point.

Ideally, the code and data supporting an article should permit a reader to reproduce all the stages of analysis the author(s) originally performed. When this is true, we say the research is "reproducible."

But there are often rough spots in reproducibility. Stochastic processes may not run exactly the same way each time, for instance.

At this point, people who study reproducibility professionally will crowd forward and offer an eleven-point plan for addressing all rough spots. ("You just set the random number seed so it's predictable ...")

That’s wonderful, if we really want to polish a system that allows a reader to push a button and get the same result as the original researcher, to the seventh decimal place. But in the humanities, we’re not always at the “polishing” stage of inquiry yet. Often, our question is more like “could this conceivably work? and if so, would it matter?”

In short, I think we shouldn’t let the imperative to share code foster a premature perfectionism. Our ultimate goal is not to prove that you get exactly the same result as the author if you use exactly the same assumptions and the same books. It’s to decide whether the experiment is revealing anything meaningful about the human past. And to decide that, we probably want to repeat the author’s question using different assumptions and a different sample of books.

When we do that, we are not reproducing the argument but replicating it. (See Language Log for a fuller discussion of the difference.) Replication is the real prize in most cases; that’s how knowledge advances. So the point of sharing code and data is often less to stabilize the results of your own work to the seventh decimal place, and more to guide investigators who may want to undertake parallel inquiries.

<…>

Ted Underwood

Вопрос 2.
Решите задачу.

Буквы Р, П, О, О, Т написаны на отдельных карточках. Федор Михайлович берет карточки в случайном порядке и прикладывает одну к другой. Какова вероятность, что Федор Михайлович с первой попытки соберет из них слово «ТОПОР»?

Вопрос 3. (автор задания — Б.В. Орехов).
Решите задачу.

Для решения задачи не требуется знакомства с старотатарским языком, все необходимые лингвистические представления можно перенести из русского языка. Буквы ɳ, ǯ и ɣ означают согласные звуки.

В тюркском варианте восточной системы стихосложения аруз размер, в котором созданы такие поэтические строки на старотатарском языке, называется хазадж-и мусаддас-и махзуф (написание несколько упрощено в угоду метрике):

Anyɳ kem al äɳindä mäɳ jaratty,
Bujy berlä sačyny täɳ jaratty
Xäkimnärdän qalan süz dörr wä ǯäwhär;
Гaqylly kemsä alyr any ezbar.

Babaxan digänul ber šäh barirde

Размер этих строк — рамал-и мусаддас-и махзуф:

Šähidirür, ike küze jäširür.
Jad itärlär — kem belä? - räxmät ilä

Размер этих строк — рамал-и мусамман-и махзуф:

Käšt itep gäzdem bu tatar ileneꞑ jaxšylaryn.
Kürde küzem, döšde küꞑlem ul bäder suräteꞑä
Näq Qazan aertynda bardyr ber awyl — Qyrlaj dilär

Размер этих строк — хазадж-и мусамман-и салим:

Güzäl šuridä bylbyl da fävan äjlär, šahym Tahir;
Гäziz ʒanyn fida äjlär, kürep ul Zöhrä dildati
Xodāɣa küp xämid itkän xämidämdin ʒöda buldym
Menä kič. Zur awyl östendä čyqty nurly aj qalqyp

Задание: определите размер этого двустишия и объясните свой ответ:

Ässälam wä ässälam wä ässälam;
I ʒanyj, bäɣdä sälam bulyr qälam

Вопрос 4.

Представьте, что вас заинтересовал феномен русскоязычной любительской литературы, т.е. художественные тексты, написанные непрофессиональными авторами, не публикуемые издательствами и не приносящие их создателям никаких денег. Каким образом можно исследовать литературу такого рода количественными методами?

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.