Стилометрия. R, Stylo

Теория

Стилеметрия (стилометрия) -- прикладная филологическая дисциплина, занимающаяся измерением стилевых характеристик с целью систематизации и упорядочения (типологии, атрибуции, датировки, диагностики, реконструкции и т.д.) текстов и их частей.

История изучения стилометрии

Впервые исследование стиля текста с целью атрибуции (Установление авторства анонимного произведения литературы или искусства, времени и места его создания) было предпринято еще в XV веке. Итальянский филолог Лоренцо Валла опубликовал трактат «Рассуждение о подложности так называемой дарственной грамоты Константина», в котором на основе различных, в том числе стилистических критериев доказывалось, что данный текст является подделкой.

История современной статистической стилистики начинается в середине XIX в., когда английский математик Аугустус де Морган в 1851 г. высказал предположение, что различные авторы могут быть определены посредством скрытых статистических черт. Рассматривая проблемы греческой прозы, Морган утверждал, что средняя длина слов в произведении автора может быть характерной чертой авторского стиля. Однако, насколько нам известно, сам де Морган никаких вычислений не делал.

В середине XIX в. также существовала группа ученых, разрабатывающая так называемый метод «стилометрики» (Ф.Г. Фриари, Дж. К. Инграм, Ф.У. Фурнивал). Они подсчитывали количество повторений определенного слова и изменение размера в стихах. Главным результатом их работы было открытие медленного, но постоянного изменения стиля Шекспира в течение 22-х лет.

Термин «стилометрия» был изобретен германским филологом Вильгельмом Диттенбергером (1880), который сделал попытку решить проблему атрибуции и хронологии диалогов Платона. Он исследовал частоту употребления слов, особенно служебных, в текстах Платона, реализация которых не зависит от тематики текста. Позже его исследования на различных материалах продолжили Е. Зеллер (1887), Ф. Чада (1901), Ц. Риттер (1903).

В России впервые Н.А. Морозов поднял проблему отличия плагиата от оригинальных работ известных авторов и применил вероятностно-статистический метод в целях атрибуции. В 1915 г. он опубликовал статью «Лингвистические спектры». Предшествующие ему исследователи опирались, главным образом, на частоту употребления знаменательных слов. Н.А. Морозов, применяя простые вычислительные способы, рассматривал частоту употребления служебных слов и их вариаций в индивидуальных текстах.

В 20-е гг. XX в. можно назвать только несколько серьезных исследователей по стилостатистике, таких как Р.Е. Паркер (1925), З.Е. Чендлер (1928), М. Пэрри (1928), и, в особенности, А. Бусман (1925), автора так называемого соотношения глагол-прилагательное.

В 30-е гг. XX в. был сделан новый шаг в применении статистических методов в стилистике такими лингвистами, как Дж. В. Флетчер (1934), рассматривавшим развитие стиля Спенсера, Г.М. Боллинг (1937), с критическим эссе по статистическому

исследованию языка Гомера, Дж. Б. Кэрролл (1938), поднимавший проблему разнообразия словаря, и У.Г. Юл (1938), первым исследовавший дистрибуцию длины предложений как статистическую характеристику стиля.

Именно с него начинается применение современных статистических методов в стилистике. С этого периода применение статистических методов в исследовании стиля распространяется по всему миру. Резко возрос интерес к статистической лингвистике, особенно в 1960-70 гг. (Дж. Б. Кэрролл, Г. Хердан, Х.Х. Сомерс, Ч. Мюллер, Б. Келман, Л.Т. Милик, Дж. Мистрик, Л. Долежел, К.Б. Уильямс, Б.Н. Головин, Й. Краус, М.Н. Кожина и др.). Именно в этот период возникают и развиваются разнообразные идеи анализа авторского стиля.

Современным отечественным лингвистом, занимающимся статистическими методами атрибуции текста, является М.А. Марусенко. Ему принадлежит идея теории распознавания образов. Он разделяет процедуру атрибуции на три относительно самостоятельных этапа: формирование литературно-критической гипотезы, проверка литературно-критической атрибутивной гипотезы методами теории распознавания образов, интерпретация результатов проверки атрибутивной гипотезы. В данной работе статистико-вероятностные методы анализа языка и стиля произведения используются автором для проверки атрибутивной гипотезы.

Одним из значительных отечественных лингвистов, занимающихся стилометрией, является Г.Я. Мартыненко. В 1988 г. он написал монографию «Основы стилометрии» и на протяжении более чем двадцати лет занимается статистическим методами в лингвистике. Некоторые научные работы по стилометрии написаны в соавторстве с Сергеем Викторовичем Чебановым.

Объект и предмет стилометрии

Объектом стилометрии является текст, созданный конкретным автором в конкретное время в конкретной ситуации.

С точки зрения теории множеств объект стилометрии - собирательное множество, а с точки зрения теории систем текст может быть отнесен к классу внутренних систем, являющихся целостными образованиями, к которым можно применять процедуры членения, представляя их в виде некоторой структуры составляющих их частей. С точки зрения теории статистики текст может рассматриваться как реальная совокупность.

Предметом исследования являются элементы стиля, которые понимаются как особенности периферии характеристики объекта. Стиль может быть описан через факультативные, поверхностные признаки текста, которые лишь неявным образом затрагивают его сущностные, глубинные характеристики. Разные уровни стилевой организации можно соотнести с разными уровнями достоверности выводимости признаков из существенных.

Методы стилометрии

Стилометрия имеет дело с количественным классифицированием, а эта область классификационных занятий тесно соприкасается с несколькими научными направлениями: теорией группировок, теорией оценивания, распознаванием образов, теорией корреляции, количественной таксономией, методами психологического тестирования и др. Границы между этими направлениями стираются, и сегодня можно говорить о комплексе подходов и методов, занимающихся теми или иными видами количественной систематизации объектов произвольной природы.

В последние годы круг решаемых стилометрией задач и репертуар применяемых ею методов существенно расширились. Практической повседневностью стала количественная таксономия текстов, стилистическое приложение нашли дешифровочные модели, относительно самостоятельное направление образовала квантитативная типология текста, начала формироваться стилистическая диагностика, большое развитие получили методы реконструкции древних текстов.

Ведущую роль в стилометрическом исследовании играет статистический метод. Статистический метод - это комплекс приемов и принципов, согласно которым производятся сбор, систематизация, обработка и интерпретация статистических данных с целью получения научных и практических выводов. В филологии этот метод сочетается с основными методами научного познания: наблюдением и экспериментом. Но в стилометрии господствует наблюдение - слежение за теми явлениями, которые заданы только и только в тексте (или в корпусе текстов).

Статистические ряды - это единственный надежный инструмент, с помощью которого можно обнаружить правильность, регулярность, устойчивость в переменчивой стилистической картине текста, выявить характер, направление и силу стилистической связи, измерить степень стилистического сходства или различия между текстами и т.д. Это основной рабочий инструмент стилометрии, с помощью которых осуществляется свертка и обобщение стилистических данных. Г.Я. Мартыненко выделяет следующую систему типов распределений.

Вероятностная теория - лингвистическая реальность

Гауссовость - негауссовость

Типичность - нетипичность

Ранг - частота

Разнообразие - ограничение разнообразия

Строение - поведение

Элемент - совокупность элементов

Виртуальность - актуальность

Однородность - неоднородность

Устойчивость - неустойчивость

Редкость - частота

Симметрия - ассиметрия

Одновершинность - многовершинность

Классификация - важный элемент научной деятельности. Основные задачи стилометрии (атрибуция, датировка, диагностика, периодизация и др.) должны рассматриваться в

контекте форм упорядочивающей и систематизирующей работы, исследуемых современной теорией классификации.

Р. В. Манекин выделяет в современном стилометрическом исследовании 3 этапа:

анализ нескольких фрагментов текстовой действительности, в соответствии с установками “идеологии изучения феномена смысла”;

процедура экстраполяции полученных выводов, в соответствии с установками идеологии “квантитативной нарратологии”;

сопоставление полученных результатов.

Пример стилометрического анализа

В качестве примера стилистического анализа приведем лингвистический этюд Н. А. Морозова.

Стилеобразующим элементом автор считает распорядительные частицы (служебные слова): «даже и при разнородности сюжетов, есть во всех языках ряд слов, которые употребляются почти одинаково во всех родах литературы и которые по своему характеру могут быть названы, как я уже выражался ранее, служебными или распорядительными частицами человеческой речи. Это прежде всего союзы, предлоги и отчасти местоимения и наречия, а затем и некоторые вставные словечки, в роде: “т.-е.”, “например” или “и так далее”.»

Такие частицы взяты из первой тысячи слов из произведений нескольких авторов XIX века - Гоголя (“Майскую ночь”, “Страшную месть” и “Тараса Бульбу"), Пушкина (“Капитанскую дочку”, “Дубровского” и “Барышню-крестьянку”), Толстого (“Смерть Ивана Ильича”, “Корнея Васильева”, “Три смерти” и “Три старца”), Тургенева (“Малиновую воду”), Карамзина (“Бедную Лизу”) и Загоскина (“Юрия Милославского”). Исследователь составил графики для каждого автора, обозначая каждую распорядительную частицу на горизонтальной линии, а число ее повторения на вертикальной. Эти графики он и назвал «лингвистическими спектрами» и предложил «исследование по ним назвать лингвистическим анализом, соответственно спектральному анализу состава небесных светил».

В числе употребляемых авторами служебных частиц (союзов и предлогов) оказались ясные процентные различия (слоговые типы).

Автор выделил разные виды спектров: «Чтоб не давать очень сложных общих спектров при нанесении этих цифр на графики, я разделил их здесь на предложные, союзные, местоименные спектры и т. д., судя по тому, что они представляют.» Все естественные спектры автор обратил в приведенные, следуя такому правилу: «Среднее число повторений каждой служебной частицы на тысячу слов данного произведения нужно разделить на среднюю повторяемость той же частицы, вычисленную по многим авторам данной эпохи. Тогда вместо предыдущих абсолютных цифр получатся Коэффициенты индивидуальности авторов, величиною своею то более, то менее единицы.» (см. таблицу).

Таковы общие основы лингвистического анализа, предлагаемого Морозовым, необходимому «для доказательства плагиатов и апокрифов, которыми полна литература, приписываемая авторам древности и начала средних веков. Лингвистический анализ дает

нам здесь объективные основы для суждений об одноавторности и разноавторности произведений.»

R, Stylo

R — язык программирования для статистической обработки данных и работы с графикой.

R широко используется как статистическое программное обеспечение для анализа данных и фактически стал стандартом для статистических программ.

R поддерживает широкий спектр статистических и численных методов и обладает хорошей расширяемостью с помощью пакетов. Пакеты представляют собой библиотеки для работы специфических функций или специальных областей применения. В базовую поставку R включен основной набор пакетов, а всего по состоянию на 2017 год доступно более 11778 пакетов.

Ещё одна особенность R - возможность создания качественной графики, которая может включать математические символы.

Stylo – пакет для R, представляющий собой набор стилометрических инструментов в виде отдельных скриптов.

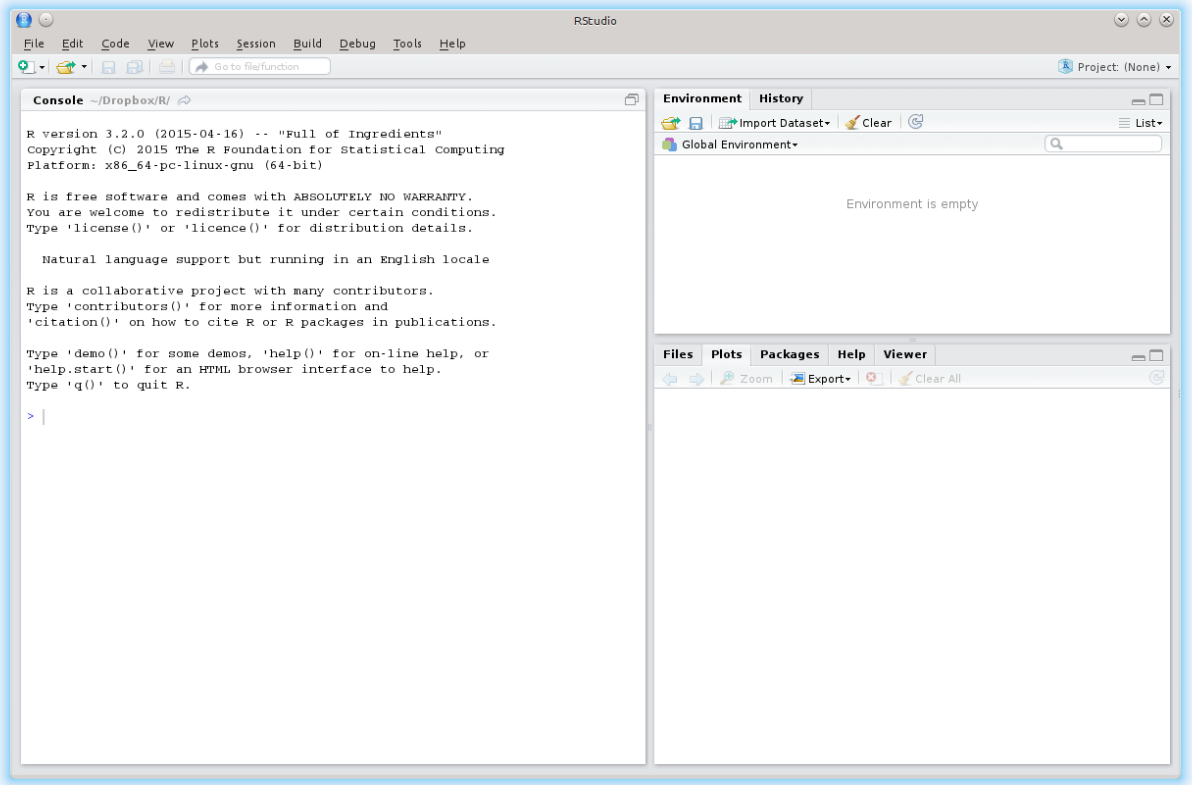
Лабораторная

Давайте научимся основам R и изучим некоторые функции Stylo.

R для Windows можно скачать отсюда. Устанавливать R рекомендуется в корневой каталог, вроде «C:\R\».

На время работы с R забудьте про мышку — практически все самые важные действия в ней выполняются с использованием командной строки. Однако для того чтобы сделать жизнь чуть легче, а саму программу чуть более приветливой, есть программа-frontend (внешний интерфейс) под названием RStudio. Скачать её можно отсюда. Устанавливается она после того, как уже установлен сам R. В RStudio много удобных инструментов и приятный интерфейс, тем не менее анализ и прогнозирование в нём всё так же осуществляются с использованием командной строки.

Интерфейс RStudio выглядит следующим образом:



В правом верхнем углу в RStudio указано имя проекта (которое пока что у нас «None» — то есть отсутствует). Если нажать на эту надпись и выбрать «New Project» (новый проект), то нам предложат создать проект. Для базовых целей прогнозирования достаточно выбрать «New Directory» (новая папка для проекта), «Empty Project» (пустой проект), а затем — ввести название проекта и выбрать директорию, в которой его сохранить.

Работая с одним проектом, вы всегда сможете обратиться к сохранённым в нём данным, командам и скриптам.

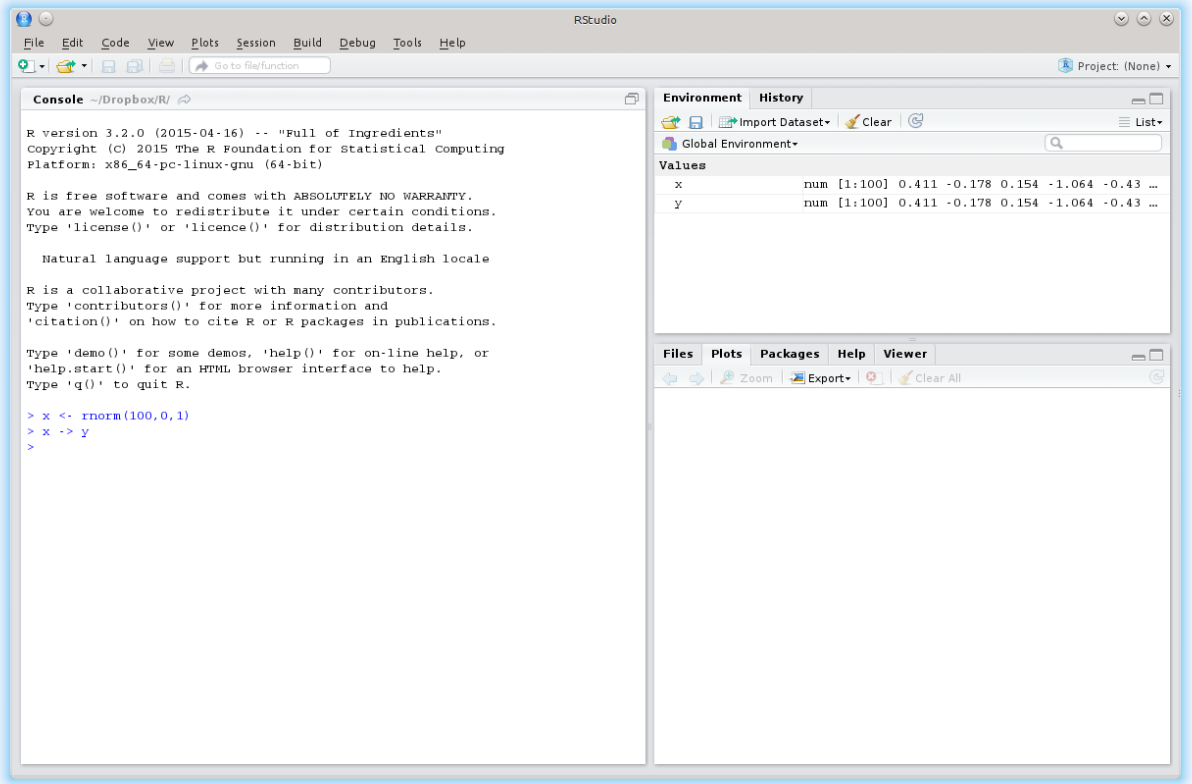
В левой части окна RStudio располагается консоль. Именно в неё мы и будем вписывать различные команды. Например, напишем следующую:

```
x <- rnorm(100,0,1)
```

Эта команда сгенерирует 100 случайных величин из нормального распределения с нулевым математическим ожиданием и единичной дисперсией, после чего создаст вектор под названием «x» и запишет полученные 100 величин в него. Символ «<-» эквивалентен символу «=» и показывает какое значение присвоить нашей переменной, стоящей слева. Иногда вместо него удобней использовать символ «->», правда наша переменная в таком случае должна стоять справа. Например, следующий код создаст объект «y» абсолютно идентичный объекту «x»:

```
x -> y
```

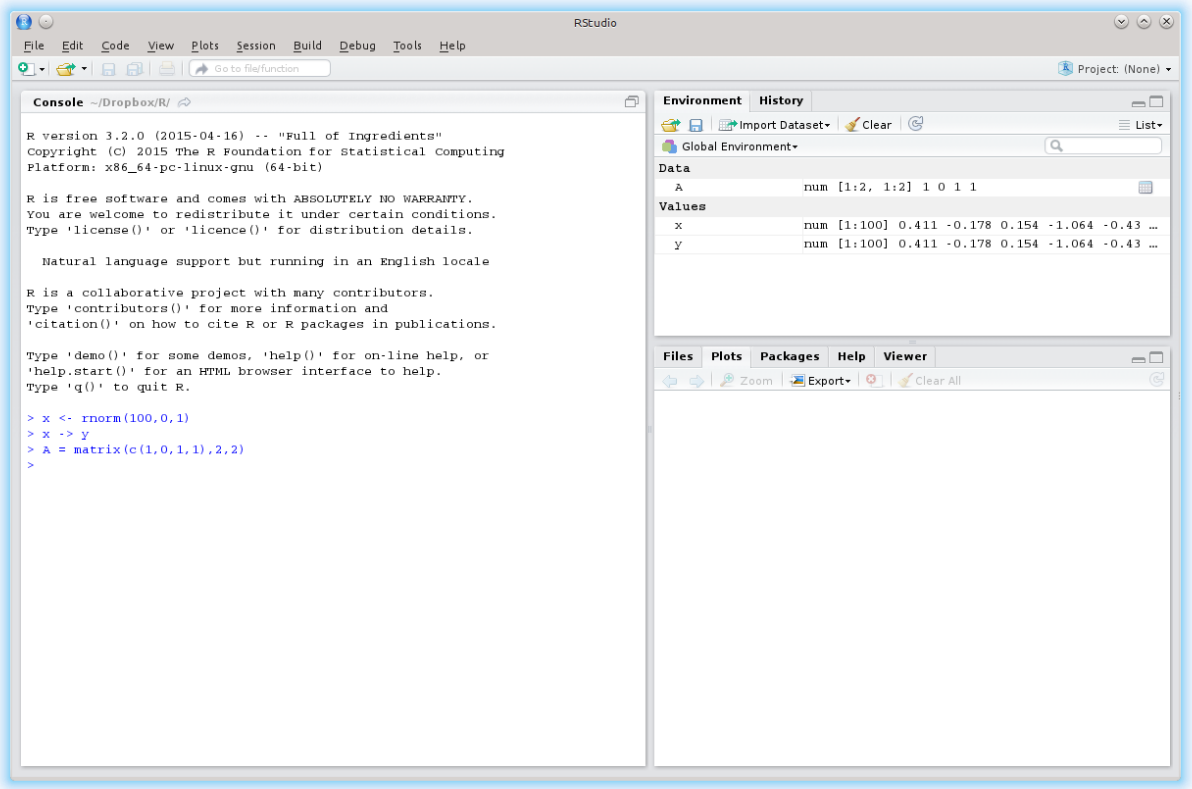
Эти векторы теперь появились в правой верхней части экрана, под закладкой, которая у меня озаглавлена «Environment»:



В этой части экрана будут отображаться все объекты, которые мы сохраняем во время сессии. Например, если мы создадим такую вот матрицу командой:

```
A <- matrix(c(1,0,1,1),2,2)
```

то она появится в закладке «Environment»:



Любая функция, которую мы используем, требует, чтобы мы задали некоторые значения определённым параметрам. В функции matrix() есть следующие параметры:

- data – вектор с данными, который должен быть записан в матрицу,
- nrow – число строк в матрице,
- ncol – число столбцов в матрице,
- byrow — логический параметр. Если «TRUE» (истина), то наполнение матрицы будет осуществляться по строкам (слева направо, строка за строкой). По умолчанию этот параметр имеет значение «FALSE» (ложь),
- dimnames — лист с именами строк и столбцов.

Некоторые из этих параметров имеют значения по умолчанию (например, byrow=FALSE), в то время как другие могут быть опущены (например, dimnames).

Одна из фишек «R» заключается в том, что к любой функции (например, к нашей matrix()) можно обратиться, задавая значения на прямую:

```
A <- matrix(data=c(1,0,1,1),nrow=2,ncol=2)
```

а можно и так, как мы сделали это ранее — соблюдая последовательность и опуская названия параметров.

Для того, чтобы увидеть содержание любого объекта, находящегося в закладке «Environment», достаточно напечатать его название в консоли:

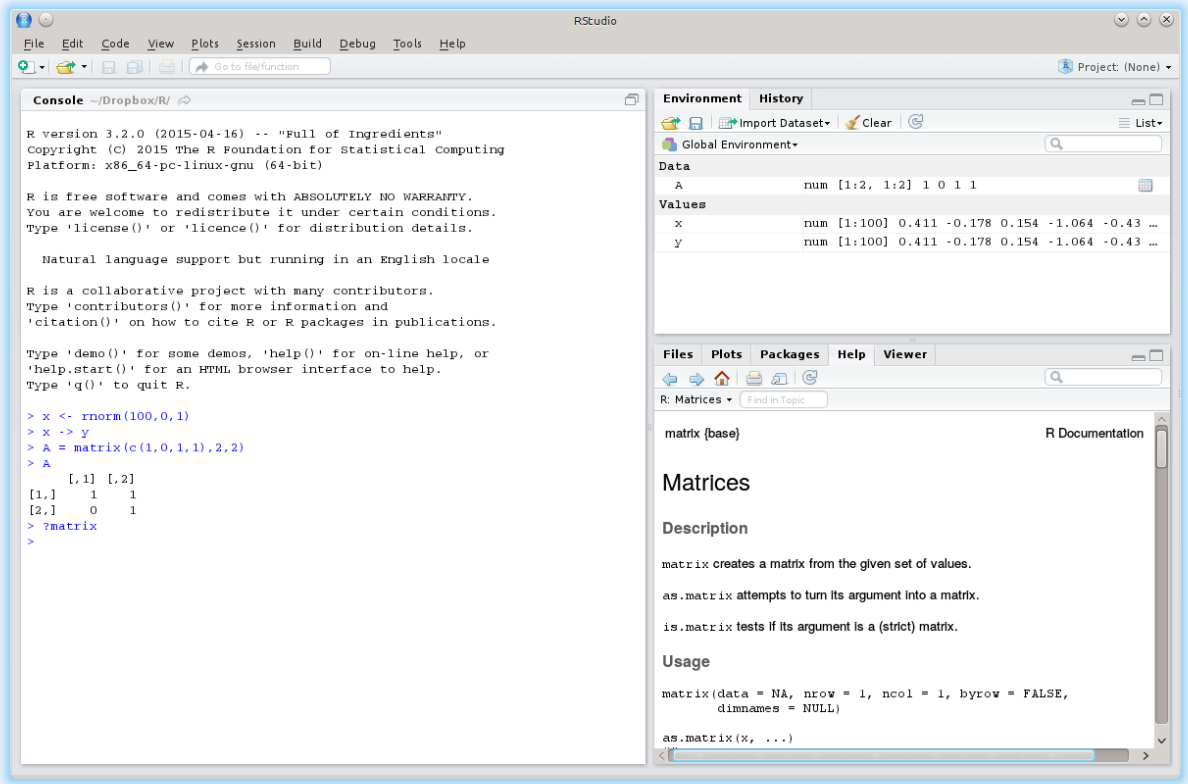
```
A
```

Другой вариант — это нажать на имя объекта в закладке «Environment».

Если вам нужно почитать подробней о какой-либо функции, можно воспользоваться следующей командой:

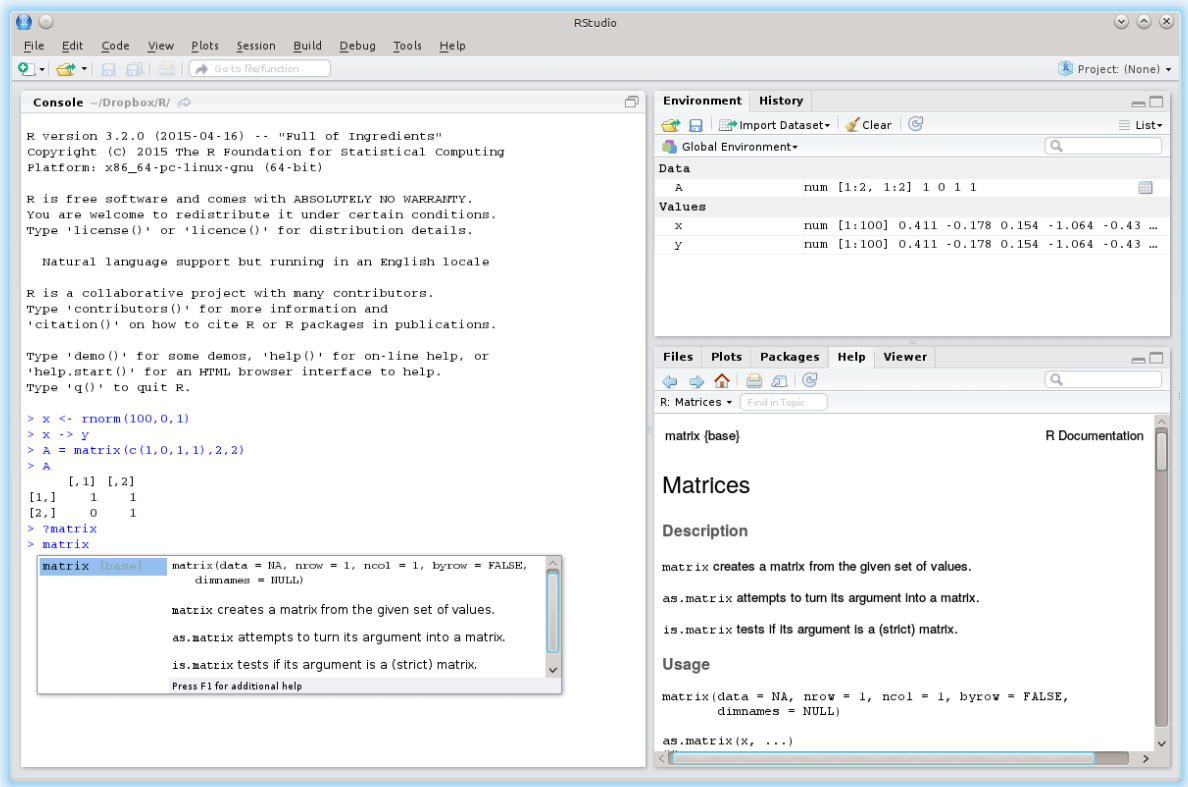
?matrix

где matrix — это название интересующей нас функции. RStudio специально для вас в таком случае откроет панель «Help» с описанием:



Найти помощь по функции можно так же, набрав название функции в окне «поиск» (иконка с линзой) в закладке «Help».

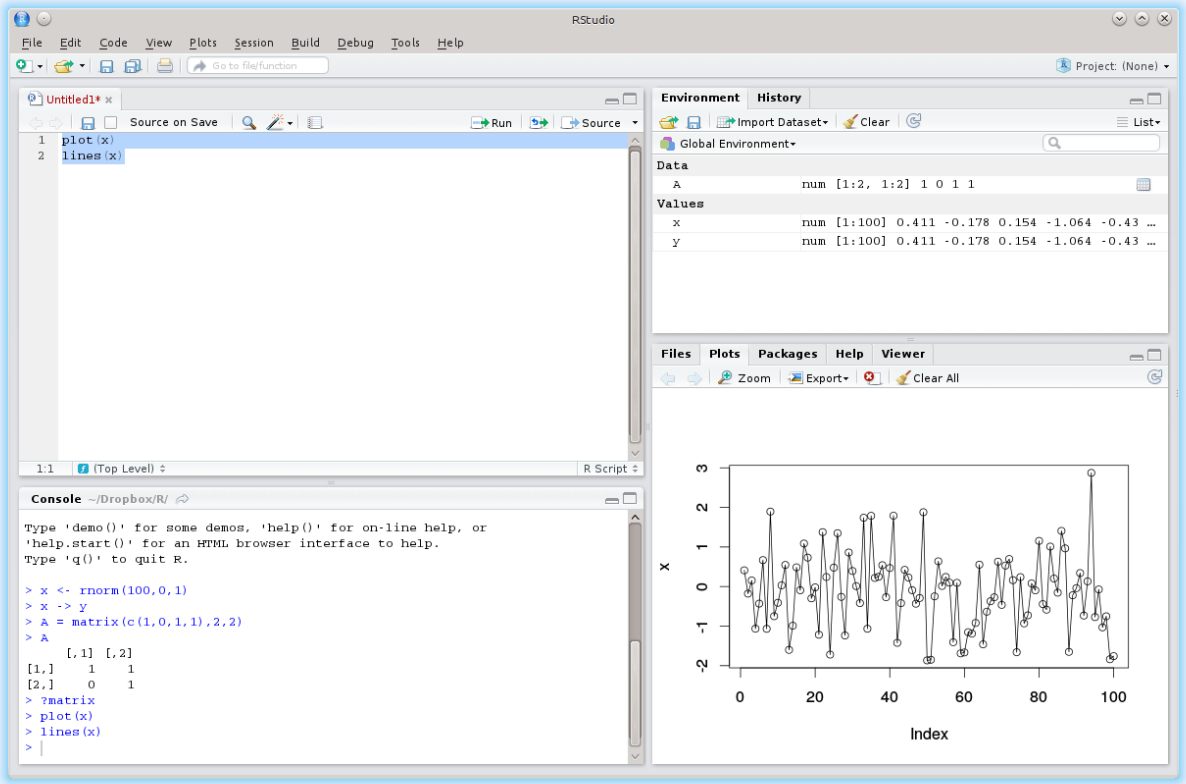
В случае, если вы не помните точно, как пишется название функции или какие в ней используются параметры, достаточно начать писать её название в консоли и нажать кнопку «Tab»:



Помимо всего этого в RStudio можно писать скрипты. Они могут понадобиться вам в том случае, если вам нужно написать программу либо вызвать последовательность функций. Создаются скрипты используя кнопку с плюсиком в верхнем левом углу (в выпадающем меню нужно выбрать «R Script»). В открывшемся после этого окне вы сможете писать любые функции и комментарии. Например, если мы хотим построить линейный график по ряду x, это можно сделать следующим образом:

```
plot(x)
lines(x)
```

Первая функция строит простейший точечный график, а вторая функция добавляет поверх точек линии, соединяющие точки последовательно. Если выделить эти две команды и нажать «Ctrl+Enter», то они будут выполнены, в результате чего RStudio откроет закладку «Plot» в правом нижнем углу и отобразит в ней построенный график.



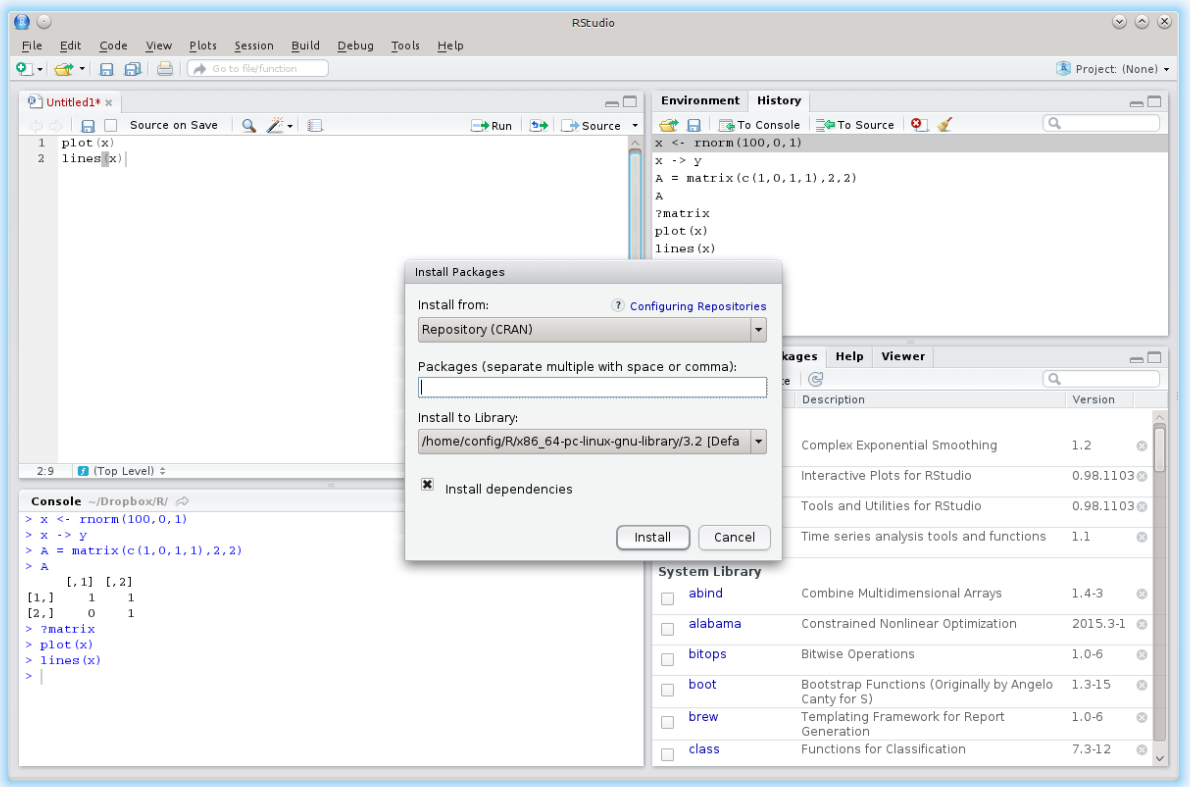
Если все набранные команды нам ещё понадобятся в будущем, то этот скрипт можно сохранить (дискетка в левом верхнем углу).

В случае, если вам нужно обратиться к команде, которую вы уже набирали когда-то в прошлом, в правой верхней части экрана есть закладка «History». В ней можно найти и выбрать любую интересующую вас команду и двойным нажатием вставить её в консоль. В самой консоли можно обращаться к предыдущим командам, используя кнопки «Up» (вверх) и «Down» (вниз) на клавиатуре. Сочетание клавиш «Ctrl+Up» позволяет в консоли показать список всех последних команд.

Вообще в RStudio много всяких полезных сочетаний клавиш, которые значительно облегчают работу с программой. Подробней о них можно почитать тут.

Список функций R можно почитать тут.

Как я уже упомянула ранее для R существует множество пакетов. Все они расположены на сервере CRAN и для установки любого из них нужно знать его название. Установка и обновление пакетов осуществляется с помощью закладки «Packages». Перейдя на неё и нажав на кнопку «Install», мы увидим следующее меню:



Наберём в открывшемся окне: stylo. Нажмём кнопку «Install» (установить), после чего пакет «stylo» будет установлен.

Как вариант мы можем установить любой пакет, зная его название, с помощью команды в консоли:

```
install.packages("stylo")
```

при условии, что он, конечно же, есть в репозитории CRAN.

Некоторые пакеты доступны только в исходных кодах на сайтах типа github.com и требуют, чтобы их перед этим собрали. Для сборки пакетов под Windows может понадобиться программа Rtools.

Чтобы использовать какой-либо из установленных пакетов, его нужно подключить. Для этого его надо найти в списке и отметить галочкой либо использовать команду в консоли:

```
library(stylo)
```

В Windows может проявиться одна неприятная проблема: некоторые пакет легко скачиваются и собираются, но ни в какую не устанавливаются. R в этом случае пишет: «Warning: unable to move temporary installation…». Всё, что нужно сделать в этом случае — добавить папку с R в исключения в антивирусе (либо выключить его на время установки пакетов).

После загрузки пакета, нам будут доступны все входящие в него функции. Например, функция stylo(), использовать которую можно так:

```
stylo(x)
```

Эта функция предназначена для того, чтобы позволить пользователям автоматически загружать и обрабатывать корпус электронных текстовых файлов из указанной папки, а также выполнять различные стилометрические анализы из многомерной статистики для оценки и визуализации стилистического сходства между входными текстами.

Самыми важными функциями помимо stylo() являются:

• classify()

• oppose()

• rolling.delta()

• rolling.classify()

Следующие функции более низкого уровня и их можно использовать для создания своих скриптов и функций:

• assign.plot.colors

• define.plot.area

• delete.markup

• delete.stop.words

• dist.argamon

• dist.cosine

• dist.delta

• dist.eder

• dist.simple

• draw.polygons

• gui.classify

• gui.oppose

• gui.stylo

• load.corpus.and.parse

• load.corpus

• make.frequency.list

• make.ngrams

• make.samples

• make.table.of.frequencies

• parse.corpus

• parse.pos.tags

• perform.culling

• perform.delta

• perform.knn

• perform.naivebayes

• perform.nsc

• perform.svm

• stylo.default.settings

• stylo.pronouns

• txt.to.features

• txt.to.words.ext

• txt.to.words

• zeta.chisquare

• zeta.craig

• zeta.eder

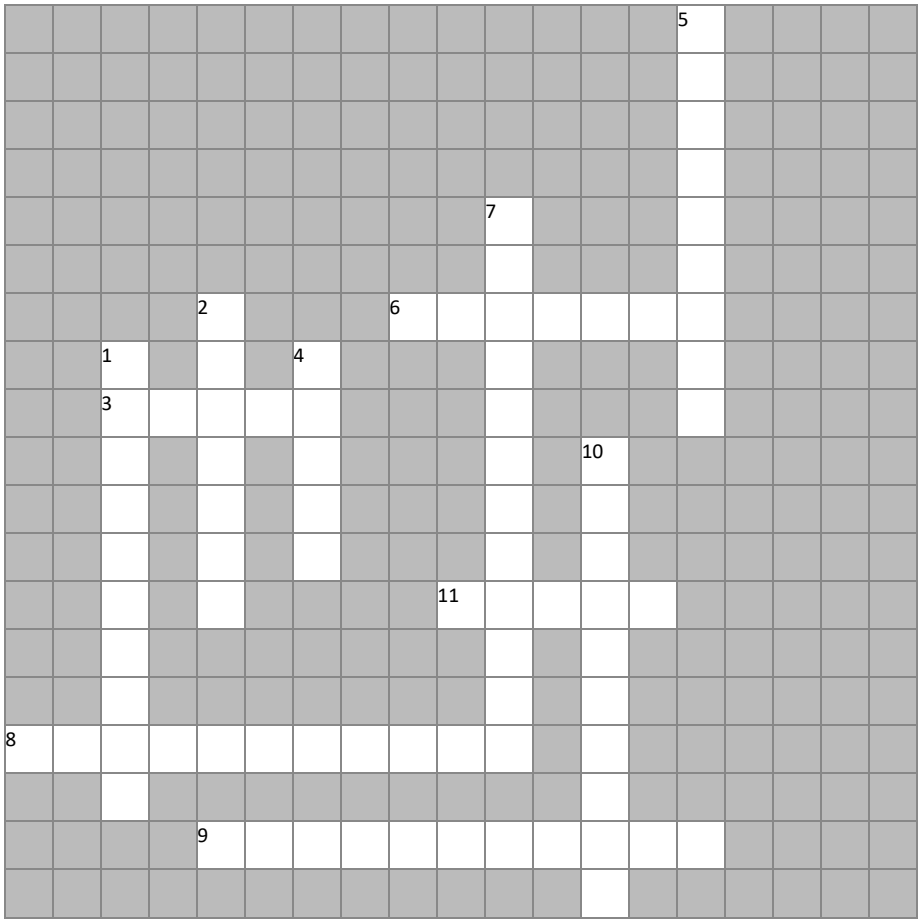
А вот вам задания для самостоятельного выполнения в R. Выполните следующие команды, посмотрите, что получится и попробуйте понять, почему так получилось:

```
(41/3 + 78/4)*2

2^3+4

1/0

0/0

max(1,min(-2,5),max(2,pi))

sqrt(3^2+4^2)

exp(2)+3i

log(1024)

log(1024, base=2)

c(1:3)

c(1:5)*2 + 4

x <- 10 + c(1:10)*0.5 +
rnorm(10,0,2)

x

mean(x)

var(x)

x <-
```

```
ts(x,start=c(2010,2),frequency=4)
```

Закрепление пройденного материала

Кроссворд



По горизонтали:

3. Объект стилометрии.

6. Российский ученый, впервые поднял проблему отличия плагиата от оригинальных работ известных авторов и применил вероятностно-статистический метод в целях атрибуции.

8. Прикладная филологическая дисциплина, занимающаяся измерением стилевых характеристик с целью систематизации и упорядочения (типологии, атрибуции, датировки, диагностики, реконструкции и т.д.) текстов и их частей.

9. Установление и изучение признаков объектов или сложных систем для характеристики их состояния; основная задача стилометрии

11. Итальянский филолог, опубликовавший трактат «Рассуждение о подложности так называемой дарственной грамоты Константина», в котором на основе различных, в том числе стилистических критериев доказывалось, что данный текст является подделкой.

По вертикали:

1. Установление авторства анонимного произведения литературы или искусства, времени и места его создания.

2. Его медленное изменение стиля открыла группа ученых в середине XIX века.

4. Метод, совокупность приёмов работы, деятельности.

5. Установление даты; основная задача стилометрии.

7. Деление на периоды; основная задача стилометрии.

10. Целенаправленное и планомерное восприятие явлений, результаты которого фиксируются наблюдателем.

соотнесите типы распределения Мартыненко

соотнесите функцию и ее определение

тест

Полезные материалы

1. Функции R
2. Скачать R
3. Скачать RStudio
4. Сочетания клавиш в RStudio
5. RTools
6. Stylo для начинающих
7. Stylo для продвинутых
8. Полная документация для Stylo