DeepDive Quick Start

DeepDive helps you extract structured knowledge from less-structured data with statistical inference without having to write any sophisticated machine learning code. Here we show how you can quickly install and run your first DeepDive application.

Launching or Installing DeepDive

Quick launching

First, you can quickly launch DeepDive with minimal installation using Docker by running the following command:.

```
bash <(curl -fsSL git.io/getdeepdive)
```

批注 [h1]: 会要求安装 curl 和 git package

Then selecting the `deepdive_docker_sandbox` option:

```
### DeepDive installer for Mac
1) deepdive                  5) jupyter_notebook
2) deepdive_docker_sandbox    6) postgres
3) deepdive_example_notebook  7) run_deepdive_tests
4) deepdive_from_release      8) spouse_example
# Install what (enter to repeat options, a to see all, q to qui
t, or a number)? 2
```

Now, point your web browser to a terminal with shell access to an environment where DeepDive is installed. You will find our examples included there as well.

```
cd deepdive-examples/spouse
```

You can also see a notebook version of the spouse example tutorial if you point your browser to the tutorial notebook.

Quick installation

If you cannot or do not want to use Docker for any reason, you can quickly install DeepDive by selecting the `deepdive` option:

```
### DeepDive installer for Mac
```

```
1) deepdive                    5) jupyter_notebook
2) deepdive_docker_sandbox    6) postgres
3) deepdive_example_notebook  7) run_deepdive_tests
4) deepdive_from_release      8) spouse_example
# Install what (enter to repeat options, a to see all, q to qui
t, or a number)? 1
```

While the sandbox provides you with a database, you are on your own with this option. You need to have a database instance to run any DeepDive application. You can select `postgres` from DeepDive's installer to install it and spin up an instance on you machine, or just run the following command:

```
bash <(curl -fsSL git.io/getdeepdive) postgres
```

Alternatively, if you have access to a database server, you can configure how to access it as a URL in the application's `db.url` file.

批注 [h2]: 默认为 postgresql://localhost/deepdive_spouse_$USER，在调用 deepdive do articles 命令时，会先尝试删除该数据库根据 postgresql 的配置，数据库 data 保存在 /var/lib/postgresql/9.5/main 中，可以用命令 du –sh 来查看文件夹大小

Running your first DeepDive app

Now, let's see what DeepDive can do for us. We grab a copy of the spouse example app explained in the tutorial. This app extracts mentions of spouses from a corpus of news articles.

(If you launched DeepDive's Docker image, then you can skip this downloading step as it's already included under `deepdive-examples/spouse/`.)

```
bash <(curl -fsSL git.io/getdeepdive) spouse_example
```

This will download a copy of the example app's code and data from GitHub to a folder whose name begins with `spouse_example-`. So, let's move into it:

批注 [h3]: 使用该版本，bug 较少

```
cd spouse_example-*
```

Then, check if we have everything there:

```
ls -F
app.ddlog  db.url  deepdive.conf  input/  labeling/  mindbender
/  README.md  udf/
```

1. Load input

First, you have to compile the DeepDive application using the following command:

批注 [h4]: 每次修改 app.ddlog 文件后，都需要重新编译

```
deepdive compile
```

Once it has compiled with no error, you can run the following `deepdive` commands.

You can find some of our sampled datasets under `input/`. You can also download the full corpus, but let's proceed with the one that has 1000 sampled articles. Run the following command to load the sampled articles into DeepDive: `bash` `deepdive load articles input/articles-1000.tsv.bz2` *Note that everytime you use the* `deepdive do` *command, it opens a list of commands to be run in your text editor. You have to confirm it by saving and quiting the editor.*

Here are a few lines from an example article in the input corpus that has been loaded.

```
deepdive query '?- articles("5beb863f-26b1-4c2f-ba64-0c3e93e721
62", content).' format=csv | grep -v '^$' | tail -n +16 | head

8:30 a.m.

Raeann Meier and Mary Darnell are among the lucky ones to land t
ickets for Thursday's papal mass at the Basilica of the Nationa
l Shrine of the Immaculate Conception.

Meier, who's from Round Hill, Virginia, won a pair of tickets i
n her church lottery and is bringing fellow parishioner Darnel
l.

Meier says of Francis: ""There is just no pope like this one.""
 She says ""Jesus hung out with the dregs — the tax collectors,
the prostitutes"" and ""that's the way this pope is.""

---

7:50 a.m.

An elaborate welcoming ceremony full of American pomp and pagea
ntry awaits Pope Francis when he goes to the White House.

The pope is scheduled to arrive by motorcade at about 9 a.m., hi
s car pulling slowly up the South Lawn driveway to a red carpet,
 where President Barack Obama and his wife, Michelle, will be w
aiting to greet him.

In front of an estimated 15,000 people who were invited by the W
hite House to witness the historic moment, Obama will then lead
 Francis to a dais decked out with even more red carpet and red,
 white and blue bunting, and ringed by military color guards. T
he Vatican and American national anthems will play. Obama will
deliver a welcome address to the pope, followed by the pope's a
ddress.

Francis will also receive a thunderous 21-gun salute.
```

2. Process input

批注 [h5]: 直接向数据库中导入数据，需要先建立好对应的数据库以及表格。多次导入时，直接在数据库中添加新的 row。
运行该命令后，则不需运行 do articles 命令，但是数据库和表格的初始化是在 do articles 命令中，需要手动初始化数据库和表格

批注 [h6]: 在使用 deepdive do articles 命令时，仍然会默认调用 input/articles.tsv.sh，从而搜索 input/signalmedia/signalmedia-1m.jsonl，然后报错找不到文件。需要在 http://research.signalmedia.co/newsir16/signal-dataset.html 中下载该文件(因为用到了 google 表单，需要翻墙)，输入邮箱后，下载地址会发送至邮箱，可用工具下载，一个大小约 1G 的 gz 压缩文件

批注 [h7]: 也可以直接使用 psql 命令查看数据库
psql deepdive_spouse_hill103
select id from articles limit 10
或者使用 pgAdmin III 来查看数据库

This app adds some useful NLP markups to the English text using Stanford CoreNLP. Based on the marked up *named entity recognition*(NER) tags, it can tell which parts of the text mention people's names. All pairs of names appearing in the same sentence are considered as *candidates* for correct mentions of married couples' names.

```
deepdive do sentences
```

After running the NLP markup process, we can see the tokens and NER tags for the example article we saw earlier.

```
deepdive query '?- sentences("5beb863f-26b1-4c2f-ba64-0c3e93e72
162", _, _, tokens, _, _, ner_tags, _, _, _).' format=csv | grep
 PERSON | tail
```

批注 [h8]: 同理，可以直接查看数据库中的 sentences 表格

```
"{An,elaborate,welcoming,ceremony,full,of,American,pomp,and,pa
geantry,awaits,Pope,Francis,when,he,goes,to,the,White,House,.}
","{O,O,O,O,O,O,MISC,O,O,O,O,PERSON,PERSON,O,O,O,O,O,ORGANIZAT
ION,ORGANIZATION,O}"

"{The,pope,is,scheduled,to,arrive,by,motorcade,at,about,9,a.
m.,"","",his,car,pulling,slowly,up,the,South,Lawn,driveway,to,
a,red,carpet,"","",where,President,Barack,Obama,and,his,wife,"
","",Michelle,"","",will,be,waiting,to,greet,him,.}","{O,O,O,
O,O,O,O,O,O,TIME,TIME,TIME,O,O,O,O,O,O,O,O,LOCATION,LOCATION,O,
O,O,O,O,O,O,O,PERSON,PERSON,O,O,O,O,PERSON,O,O,O,O,O,O,O,O}"

"{In,front,of,an,estimated,""15,000"",people,who,were,invited,
by,the,White,House,to,witness,the,historic,moment,"","",Obama,
will,then,lead,Francis,to,a,dais,decked,out,with,even,more,re
d,carpet,and,red,"","",white,and,blue,bunting,"","",and,ringe
d,by,military,color,guards,.}","{O,O,O,O,O,NUMBER,O,O,O,O,O,O,
ORGANIZATION,ORGANIZATION,O,O,O,O,O,O,O,PERSON,O,O,O,PERSON,O,O,
O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O}"

"{Obama,will,deliver,a,welcome,address,to,the,pope,"","",follo
wed,by,the,pope,'s,address,.}","{PERSON,O,O,O,O,O,O,O,O,O,O,O,
O,O,O,O,O}"

"{Francis,will,also,receive,a,thunderous,21-gun,salute,.}","{P
ERSON,O,O,O,O,O,O,O,O}"

"{People,hoping,to,catch,a,glimpse,of,Pope,Francis,during,a,la
te,morning,parade,are,lining,up,for,a,coveted,spot,along,the,r
oute,.}","{O,O,O,O,O,O,O,O,PERSON,O,O,TIME,TIME,O,O,O,O,O,O,O,
O,O,O,O,O}"

"{As,a,head,of,state,"","",Pope,Francis,officially,is,in,the,
U.S.,on,what,'s,known,as,a,``,state,visit,.,''}","{O,O,O,O,O,
O,O,PERSON,O,O,O,O,LOCATION,O,O,O,O,O,O,O,O,O,O,O,O}"
```

```
"{For,one,thing,"","",President,Barack,Obama,and,Francis,will,not,review,the,troops,"","",as,presidents,do,with,other,visiting,leaders,.}","{O,NUMBER,O,O,O,PERSON,PERSON,O,PERSON,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O}"

"{Nor,will,Francis,return,to,the,White,House,in,the,evening,as,the,guest,at,a,lavish,state,dinner,"","",one,of,the,highlights,of,most,state,visits,.}","{O,O,PERSON,O,O,O,LOCATION,LOCATION,O,O,TIME,O,O,O,O,O,O,O,O,O,O,NUMBER,O,O,O,O,O,O,O,O}"

"{That,'s,largely,because,of,Francis,',busy,schedule,.}","{O,O,O,O,O,PERSON,O,O,O,O}"
```

We can continue running the processes until all candidates of spousal mentions are mapped, and see the pairs of names from the example article.

```
deepdive do spouse_candidate
deepdive query 'name1, name2 ?-
    spouse_candidate(p1, name1, p2, name2),
    person_mention(p1, _, "5beb863f-26b1-4c2f-ba64-0c3e93e72162", _, _, _).
'
    name1      |    name2
---------------+--------------
 Raeann Meier | Mary Darnell
 Meier        | Darnell
 Meier        | Francis
 Barack Obama | Francis
 Francis      | Obama
 Barack Obama | Michelle
 Obama        | Francis
 Barack Obama | Francis
(8 rows)
```

批注 [h9]: 同理可以直接查询 spouse_candidate 和 person_mention 表格

For supervised machine learning, the app continues with extracting *features* from the context of those candidates and creating a training set programmatically by finding promising positive and negative examples using *distant supervision*.

3. Run the model

Using the processed data, the app constructs a statistical inference model to predict whether a mention is a correct mention of spouses or not, estimates the parameters (i.e., learns the weights) of the model, and computes their *marginal probabilities*.

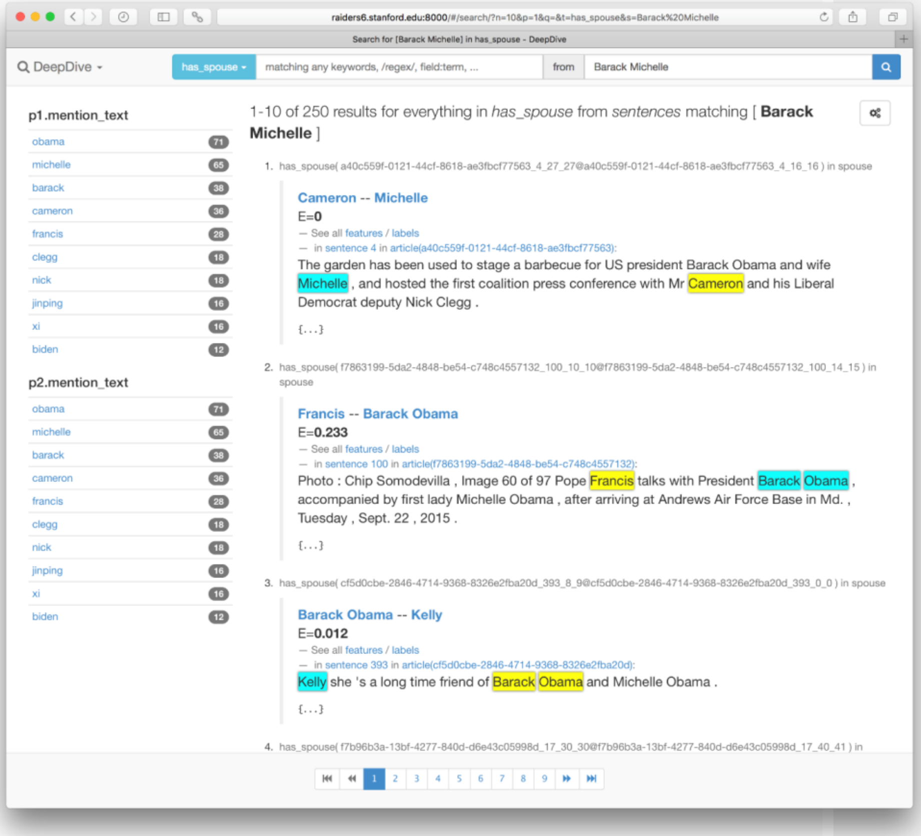
```
deepdive do probabilities
```

As a result, DeepDive gives the expectation (probability) of every variable being true. Here are the probabilities computed for the pairs of names from the example article we saw earlier:

```
deepdive sql "
    SELECT p1.mention_text, p2.mention_text, expectation
    FROM has_spouse_label_inference i, person_mention p1, person_mention p2
    WHERE p1_id LIKE '5beb863f-26b1-4c2f-ba64-0c3e93e72162%'
      AND p1_id = p1.mention_id AND p2_id = p2.mention_id
  "
 mention_text | mention_text | expectation
--------------+--------------+-------------
 Raeann Meier | Mary Darnell |       0.129
 Meier        | Darnell      |           0
 Meier        | Darnell      |           0
 Meier        | Francis      |       0.009
 Barack Obama | Francis      |       0.002
 Francis      | Obama        |       0.011
 Barack Obama | Michelle     |       0.648
 Barack Obama | Michelle     |       0.598
 Obama        | Francis      |       0.014
 Barack Obama | Francis      |       0.017
(10 rows)
```

DeepDive provides a suite of tools and guidelines to work with the data produced by the application. For instance, below is a screenshot of an automatic interactive search interface DeepDive provides for browsing the processed data with predicted results.

批注 [h10]: 同理，可以直接查看 has_spouse_label_inference 表格



raiders6.stanford.edu:8000/#/search/?n=10&p=1&q=&t=has_spouse&s=Barack%20Michelle

Search for [Barack Michelle] in has_spouse - DeepDive

matching any keywords, /regex/, field:term, ...

Barack Michelle

p1.mention_text	
obama	71
michelle	65
barack	38
cameron	36
francis	28
clegg	18
nick	18
jinping	16
xi	16
biden	12

p2.mention_text	
obama	71
michelle	65
barack	38
cameron	36
francis	28
clegg	18
nick	18
jinping	16
xi	16
biden	12

1-10 of 250 results for everything in *has_spouse* from *sentences* matching [**Barack Michelle**]

1. has_spouse(a40c559f-0121-44cf-8618-ae3fbcf77563_4_27_27@a40c559f-0121-44cf-8618-ae3fbcf77563_4_16_16) in spouse

   Cameron -- Michelle
   E=**0**
   — See all features / labels
   — in sentence 4 in article(a40c559f-0121-44cf-8618-ae3fbcf77563):
   The garden has been used to stage a barbecue for US president Barack Obama and wife Michelle , and hosted the first coalition press conference with Mr Cameron and his Liberal Democrat deputy Nick Clegg .

   {...}

2. has_spouse(f7863199-5da2-4848-be54-c748c4557132_100_10_10@f7863199-5da2-4848-be54-c748c4557132_100_14_15) in spouse

   Francis -- Barack Obama
   E=**0.233**
   — See all features / labels
   — in sentence 100 in article(f7863199-5da2-4848-be54-c748c4557132):
   Photo : Chip Somodevilla , Image 60 of 97 Pope Francis talks with President Barack Obama , accompanied by first lady Michelle Obama , after arriving at Andrews Air Force Base in Md. , Tuesday , Sept. 22 , 2015 .

   {...}

3. has_spouse(cf5d0cbe-2846-4714-9368-8326e2fba20d_393_8_9@cf5d0cbe-2846-4714-9368-8326e2fba20d_393_0_0) in spouse

   Barack Obama -- Kelly
   E=**0.012**
   — See all features / labels
   — in sentence 393 in article(cf5d0cbe-2846-4714-9368-8326e2fba20d):
   Kelly she 's a long time friend of Barack Obama and Michelle Obama .

   {...}

4. has_spouse(f7b96b3a-13bf-4277-840d-d6e43c05998d_17_30_30@f7b96b3a-13bf-4277-840d-d6e43c05998d_17_40_41) in

Next steps

- For more details about the spouse example we just ran here, continue reading the tutorial.
- Other parts of the documentation will help you pick up more background knowledge and learn more about how DeepDive applications are developed.

Reading them will prepare you to write your own DeepDive application that can shed light on some dark data and unlock knowledge from it!