Adding Structure and Extracting Features

flickr.com/photos/neekon59/

What is a text?

```
 DRIED BEAN SOUP (FROM STOCK).
  2 quarts dried  beans .  1 gallon  stock.   2 pounds  bacon.   1
gallon  boiling water.    Salt  and  pepper  to taste.   Wash the
beans  and soak them over night. In the morning drain the  water
off, and cover them again with the  boiling water;  add the  bacon
and boil gently two hours or more; now add the  stock.  Press the
beans  through a sieve, return them to the soup kettle, and bring
to a boil; add  salt  and  pepper,  and serve with  toasted bread.

  MACARONI  SOUP (FROM  STOCK ).
  4 ounces  macaroni.   1 gallon  stock.    Salt  and  pepper  to
taste.   Break the  macaroni  into pieces about two inches long;
put it into a stewing pan and cover it with one quart of  boiling
water;  boil it twenty minutes, drain, and cut each piece in two.
Melt the  stock,  bring it to a boiling point; add the  macaroni,
let it simmer five minutes; add  salt  and  pepper  and serve.

A plate of  cheese  may be served with this if liked.

  SAGO  SOUP (FROM  STOCK ).
  4 ounces  sago.   1 gallon  stock.    Salt  and  pepper  to
taste.   Wash the  sago  through several waters, then cover it with
warm water  and let it soak one hour. Melt the  stock  and bring it
to the boiling point; drain the  sago,  and add it to the  stock.
Let it boil slowly half an hour, stirring very often to prevent
scorching; add  salt  and  pepper  and serve.
```

```
<purpose align="center" rend="bold" placement="heading">DRIED BEAN SOUP (FROM
    STOCK).</purpose>
<list><item align="center">2 quarts dried <ingredient>beans</ingredient>.</item><
    item align="center">1 gallon <ingredient>stock.</ingredient></item><item align=
    "center">2 pounds <ingredient>bacon.</ingredient></item><item align="center">1
    gallon <ingredient>boiling water.</ingredient></item><item align="center"><
    ingredient>Salt</ingredient> and <ingredient>pepper</ingredient> to taste.</
    item></list> Wash the <ingredient>beans</ingredient> and soak them over night.
    In the morning drain the <ingredient>water</ingredient> off, and cover them
    again with the <ingredient>boiling water;</ingredient> add the
    <ingredient>bacon</ingredient> and boil gently two hours or more; now add the <
    ingredient>stock.</ingredient> Press the <ingredient>beans</ingredient> through
    a sieve, return them to the soup kettle, and bring to a boil; add
    <ingredient>salt</ingredient> and <ingredient>pepper,</ingredient> and serve
    with <ingredient>toasted bread.</ingredient>
</p>
</recipe>
<recipe class1="soups">
<p>
<purpose align="center" rend="bold" placement="heading"><ingredient>MACARONI</
    ingredient> SOUP (FROM <ingredient>STOCK</ingredient>).</purpose>
<list><item align="center">4 ounces <ingredient>macaroni.</ingredient></item><item
    align="center">1 gallon <ingredient>stock.</ingredient></item><item align=
    "center"><ingredient>Salt</ingredient> and <ingredient>pepper</ingredient> to
    taste.</item></list> Break the <ingredient>macaroni</ingredient> into pieces
    about two inches long; put it into a stewing pan and cover it with one quart of
    <ingredient>boiling water;</ingredient> boil it twenty minutes, drain, and cut
    each piece in two. Melt the <ingredient>stock,</ingredient> bring it to a
    boiling point; add the <ingredient>macaroni,</ingredient> let it simmer five
```

unstructured **structured**

Feeding America: The Historic American Cookbook Dataset. East Lansing: Michigan State University Libraries Special Collections. https://www.lib.msu.edu/feedingamericadata/

Optical Character Recognition (OCR)

GDSC001790001300.txt

```
Mo% ff1 aL~ le~z~ PON - - DOUARS CENTS SALANCE BROT FO, ......... ... r_2 AMOUNT DEPOSITED-n TOTAL..-.. __ ! TOR DOUARS CENTS BALANCE ,BRT Poo ....... ........ 7 AMOUNT DEPOSITED..-.......--- N T. A . .............. / 7 2/ AMOUNT OSTE en/r   AMOUNT THIS CHCK... .......... . / . 7 .0 / AILANCE CAfDO
/0V I m 4d t

_,. ',..5_ -/]' r F ?, l I , u. re,,' . C S G.
 ' (r e 1*1 - - DOLARS CENTS BALANCE Id Rolt ....... 141 2I AMOUNT DEPOSITED......... ___ AMOUNT THIS CECK............ BALANCE , FOROD ............. ,I .l / ALANCE AOT ............. R I DOUARS CENTS AMOUNT DEPOSITEO ........ TOTAL............ AMOUNT THIS CHECK ........ ...... ALACE CAD ......... ...

dt~ql .~ "ome s cour s*AlsNeLLr... I 5'N AMOUNT OPOIITIO........_.. - A.0UNT THIS CECK.......... .l o v --- ALANC CAODe 402rro .31 , " 2-.. 1, 4, t- ,, + Zj -. " -rorh''11 Bee Wptj4 DOUARS CENTS ALANC..e ..mdr ......... . AMOUNT DEPOSITED......-.- TOTAL..-- i AMOUNT THIS CHECK.......... SAL.ANC CAD FOWD .........D SALANCE C ...D..... Z"S 1' AMQUN vum uear

DOLLARS Colr NALANCKE sdt OM -. ---V 4- AMOUNT ODPOSITED-------- AMOUNT THIS CnHCK............ OOU.ARS CENTS *ALANCE sRmT FORb ... .......... !1.9 AMOUNT DEPOSITED ...... TOTAL. AMOUNT THIS CHECK... . . - ,ALANC CA ............. 2 --zs- iagS oouARS CINTS SALAUNC eOTdoT ....... ..... AMOUNT DEPOSITED. ...... AMOUNT THIS CHECK ... 9... SALAMCE CAI^D Pon...... · C

N.- t J3> -i .A..LAc *T i ............ ANCE tdT NT AMOUNT DEPOSITeD ...... . TOTAL-..-- . 12 AMOUNT THIS CHCK......... I ..f .ALA.C CAIo r o......... ... l / .. WT , 1/7, 5 (- n- DOLLARS CENTS AMOUNT DEPOSITED ....... TOTA ..... AMOUNT THIS CHECK ........... 71 I Jc BALANCE CA~D P0 b............. DOLLARS CENTS MALANCE RdT of . TOTAL......-- AMOUNT THIS CHECK... . L..... .. SALANC CAb PO f0 .........

--- (g-- r, --- 'C^I POR DOLLARS CENTS BALANCE RT FOd......... AMOUNT DEPOSITD .... TOTAL.......... AMOUNT THS CCK......... VALANCE CA#(* D.......... .. ^ /~~~uc url^ - ^^ ^ ^^- No.-i -- ->s ja __ o 19 _ ron DUAS CENTS SALANCE andT ro ............. -7 b AMOUNT EPOITED...... TOTAL.......... . AMOUNT THIS CHECK .. . ........... BALANCE CAD POo 0............ . NO. $ 19- W1' POR DOLLARS CENTS BALANCE BRT FOD........... AMOUNT DEPOSITED...... TOTAL .......... . AMOUNT THIS CHECK. BALANCE CAfD POW ............

a~/.L popn sALANCE SNdT Parn .......... {,/,/#*P 7'fdas-9 CrNTs rr r-r I ' AM TOTAOUNT E __ - - AMOUNT THIS CHKCK..... . - • ·AAAACE CAAFD DOLLARS CENTS BALANCE OdT ORb.............. AMOUNT DEPOSITED ...... - TOTAL-..-7 - AMOUNT THIS CHC.......... AANC CAD OD............... 4 7 S ZSt O2d DOLLARS CENTS BALANCE aRdT FOR D .......... 4 47 AMOUNT DEPOSITED.-.. AMOUNT THIS CHECK ....... - SALANCE CAoD POWD........... LANIUms,

-Ig3z-- C c 4~s -^AA*- FOR DOLLARS CENTS SMANCE RndT Fon.............. AMOUNT DEPOSITED ..--.... -- AMOUNT THIS CHEC........... ? 7 BALANCE CAb o.D -......... ,Z3n ' R1· 7 ~Lh/rc// i OGUARS CINTS BALANCE smdT FORD...... Z AMOUNT OEPOSIT.O....... oas. o iIi c TAL .......... A~. /7 AMOUNT mTHIS Ce .......... .. / BALANCE CA.O POWD ............ C j/ 7 NO.-"__-$____ _____ LLZe~ , 3S, ,. DOS CENTS ALANCE RdOT FO.TO n ..... .... - /o AMOUNT DePOSTE ___________ or TOTAL -r -- AMOUNT THIS CHICK .. ......... BALANCE CA FO ro _......... 0 . /Ai L Ij/ / <^Z,7 ^^^^ S<:7

FOR - DOMARS CENTS BAANCE a d F ................ Z A3 69- AMOUNT DEPOSITeD. L.-. .,o 3 . ", Z AMOUNT THIS CHICK..-... . . ..1 ALANCE CA rD mFO D... . ..'. 5 To 49 1 49k;: DOLLArS CENTS BALANCE ,otoT roW........ - - - 31 AMOUNT DEPOSITE ....... TOTAL -- 3- 9 Ap AMOUNT THIS CHEsCK.......t BALANCE CAf , oo........ ...... NO.-$ DOLLARS CENTS *ALANCE *dT Frot...... ...... .. 9 n / AMOUNT DOPOSITED..... TOTAL_... . .4I, .- AMOUNT THIS CHECK.__ ... BALANCE CARD VOWO-.............. nV2' 7 o9 ,.le22= ·ce
```

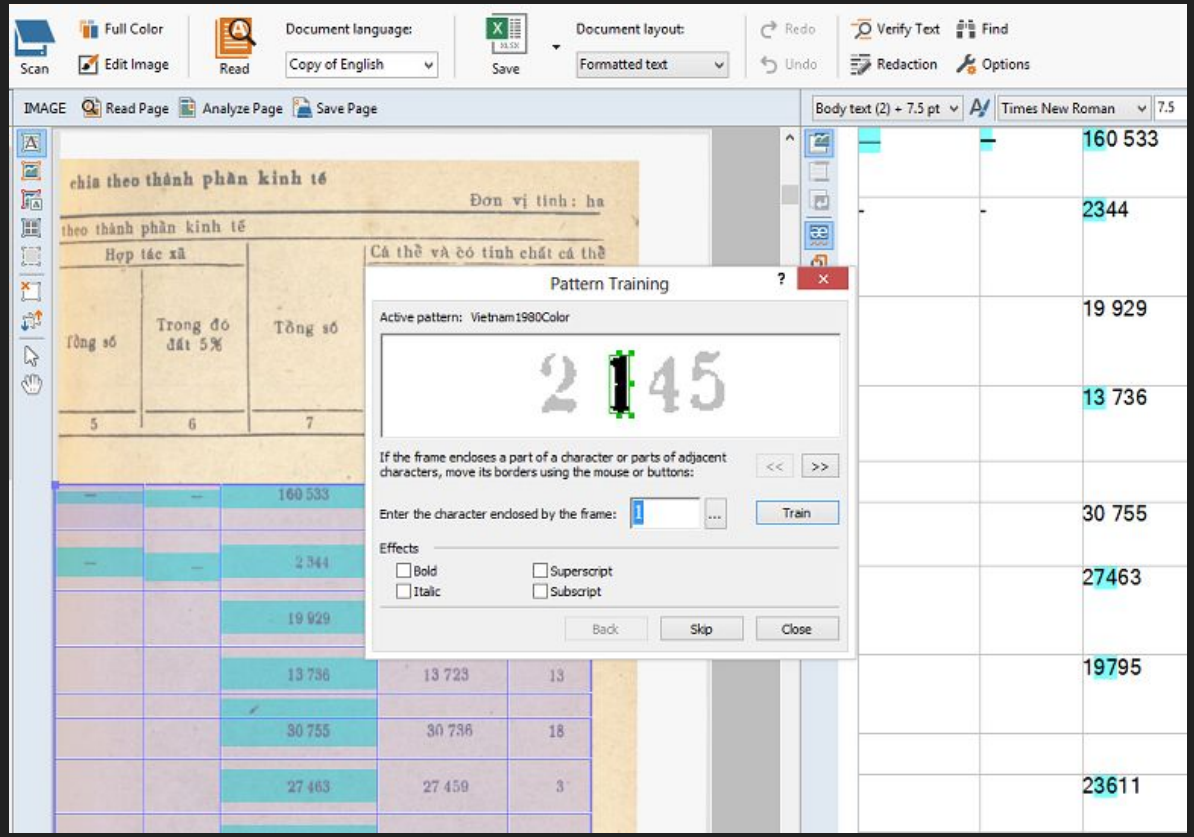
GDSC001790000200_ner-results.txt

```
PERSON

Nora L. Campbell
Nora L. Capbell
IPmmei Lou
Fannie Lou Haer
Shirley Edmrds
Fann Lou
T.L. Hill
Fannie Lou Haar
Nora L. Campbell
Nora L. Capbell
Reid
Idell Walker
Famie Lou Haer
WH PICA

ORGANIZATION
FREEDOM FARM CORP.
United Gas Company Eleven Dollars
N lr N EG
IB
Machine Shop
Five
FREEDOM FARM CORP.
FREEDOM FARM CORP.
Drew Separate School District
Sunflom
Hum Freedm Farm Corp.
RULEVILLE
FREEDOM FARM CORP.
Tax Collectors Office
G
C EG
OB
FREEDOM FARM CORP.
ULEVILLE
Town of Rulewvlle
FREEDOM FARM CORP.
OB
ORGANIZATION Corp.

LOCATION
RULEVILLE, RULEVILLE
MISSISSIPPI
RULEVILLE, RULEVILLE
```

Fannie Lou Haer
Fann Lou
Fannie Lou Haar
Famie Lou Haer
IPmmei Lou (?)

78% OCR confidence

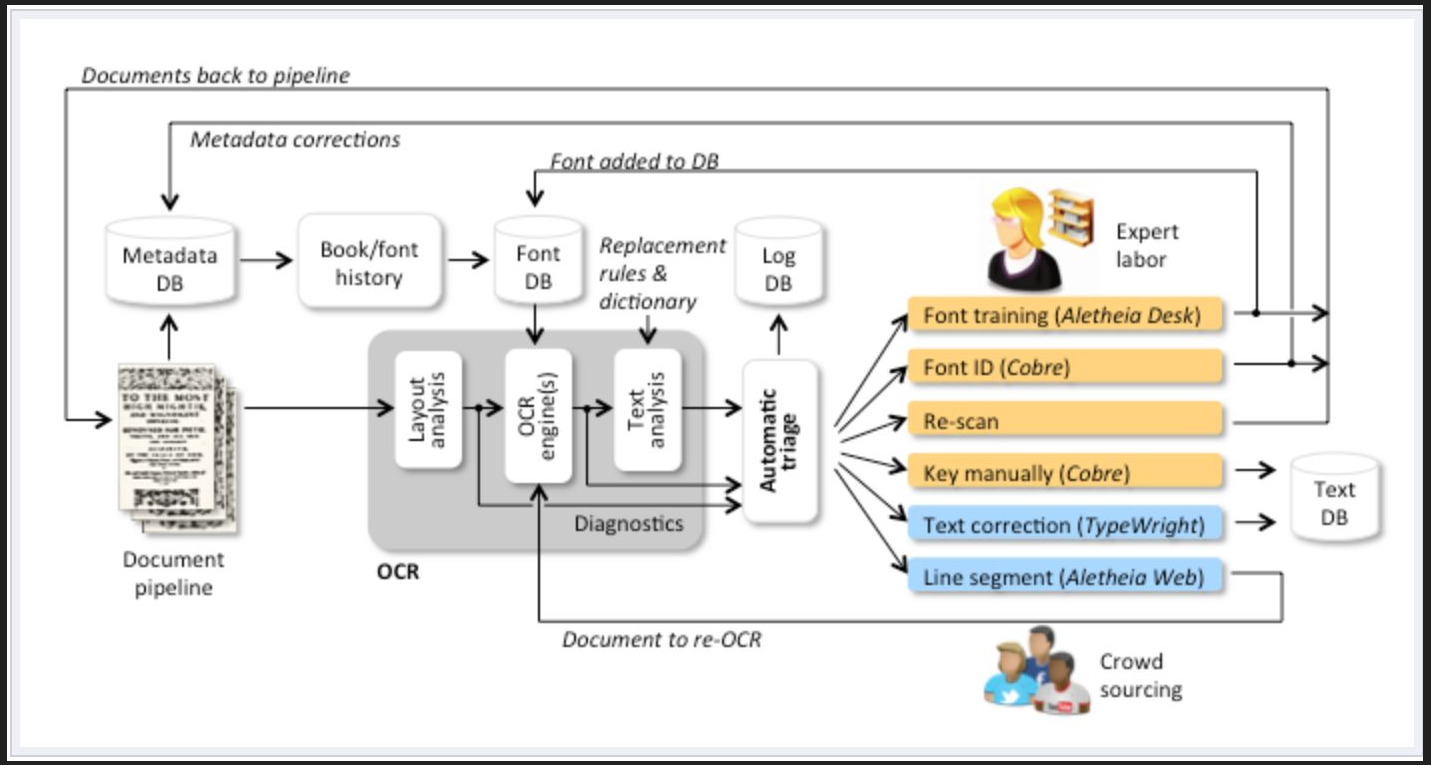
Abbyy Finereader



- Proprietary ($200-600)
- Users can create a pattern by manually correcting OCR, running improved pattern over entire corpus
- Time intensive, still not perfect

photo from: nathanlane.info/blog.html

Early Modern OCR Project (eMOP)

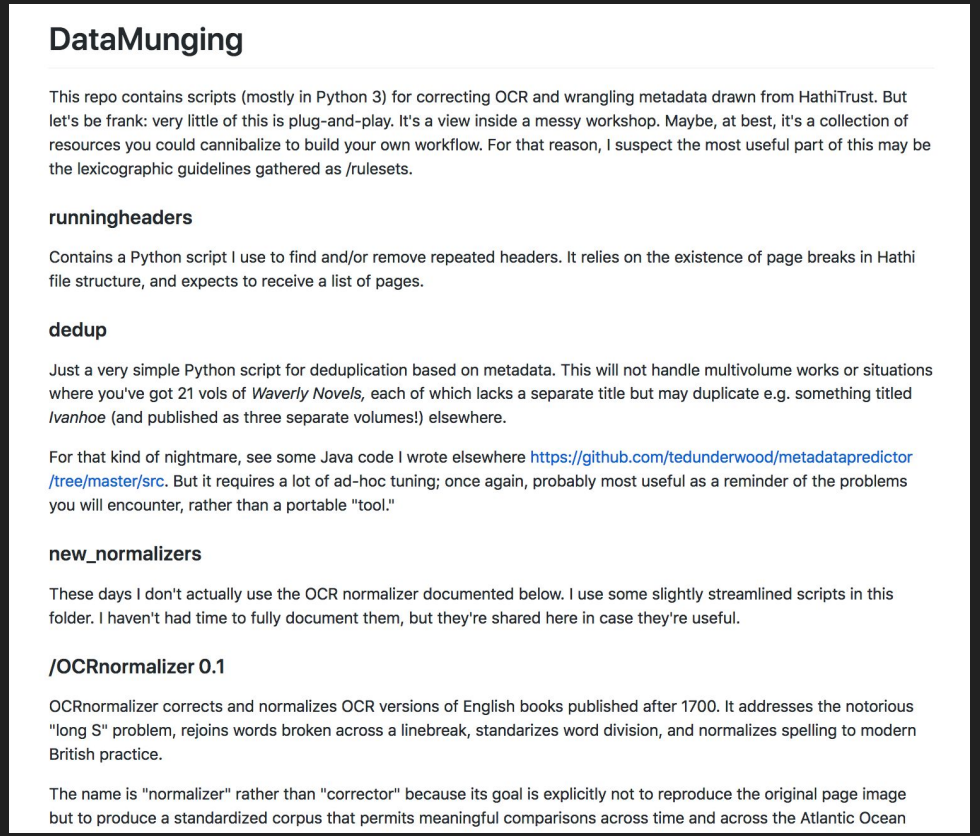


emop.tamu.edu | github.com/idhmc-tamu/eMOP

Post-OCR Algorithmic Editing

- Using hard rules to correct common mistakes
- Using dictionaries to determine which words are likely incorrect
- Quality/reuse varies widely based on time period, print methods, scan quality, etc
- Will not solve all errors and will likely introduce some

Screenshot from github.com/tedunderwood/DataMunging



DataMunging

This repo contains scripts (mostly in Python 3) for correcting OCR and wrangling metadata drawn from HathiTrust. But let's be frank: very little of this is plug-and-play. It's a view inside a messy workshop. Maybe, at best, it's a collection of resources you could cannibalize to build your own workflow. For that reason, I suspect the most useful part of this may be the lexicographic guidelines gathered as /rulesets.

runningheaders

Contains a Python script I use to find and/or remove repeated headers. It relies on the existence of page breaks in Hathi file structure, and expects to receive a list of pages.

dedup

Just a very simple Python script for deduplication based on metadata. This will not handle multivolume works or situations where you've got 21 vols of *Waverly Novels,* each of which lacks a separate title but may duplicate e.g. something titled *Ivanhoe* (and published as three separate volumes!) elsewhere.

For that kind of nightmare, see some Java code I wrote elsewhere https://github.com/tedunderwood/metadatapredictor/tree/master/src. But it requires a lot of ad-hoc tuning; once again, probably most useful as a reminder of the problems you will encounter, rather than a portable "tool."

new_normalizers

These days I don't actually use the OCR normalizer documented below. I use some slightly streamlined scripts in this folder. I haven't had time to fully document them, but they're shared here in case they're useful.

/OCRnormalizer 0.1

OCRnormalizer corrects and normalizes OCR versions of English books published after 1700. It addresses the notorious "long S" problem, rejoins words broken across a linebreak, standarizes word division, and normalizes spelling to modern British practice.

The name is "normalizer" rather than "corrector" because its goal is explicitly not to reproduce the original page image but to produce a standardized corpus that permits meaningful comparisons across time and across the Atlantic Ocean

Regular Expressions

- A method for searching for structure, words, phrases that match a pattern
- Widely used in programming languages and software, but syntax can vary slightly between different languages
- regex101.com can help tremendously

right: comic from xkcd.com/208/

RegEx

INT. APARTMENT

DANIEL KALUUYA is hanging out with his girlfriend KEIRA KNIGHTLEY.

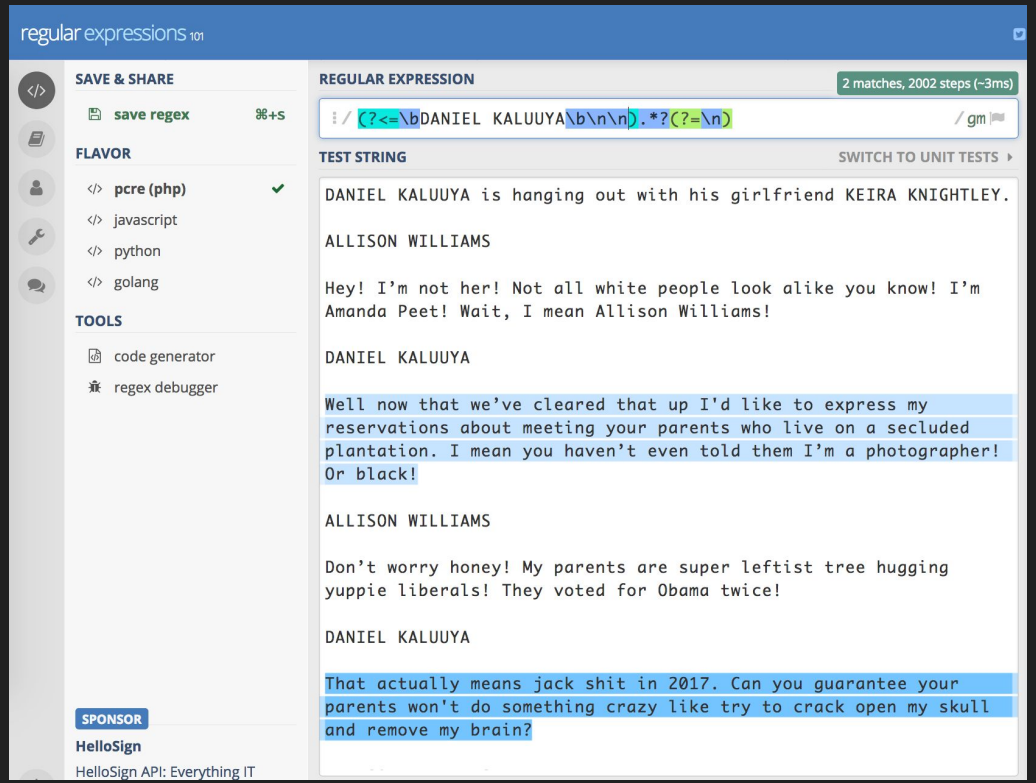
ALLISON WILLIAMS
Hey! I'm not her! Not all white people look alike you know! I'm Amanda Peet! Wait, I mean Allison Williams!

DANIEL KALUUYA
Well now that we've cleared that up I'd like to express my reservations about meeting your parents who live on a secluded plantation. I mean you haven't even told them I'm a photographer! Or black!

ALLISON WILLIAMS
Don't worry honey! My parents are super leftist tree hugging yuppie liberals! They voted for Obama twice!

DANIEL KALUUYA
That actually means jack shit in 2017. Can you guarantee your parents won't do something crazy like try to crack open my skull and remove my brain?

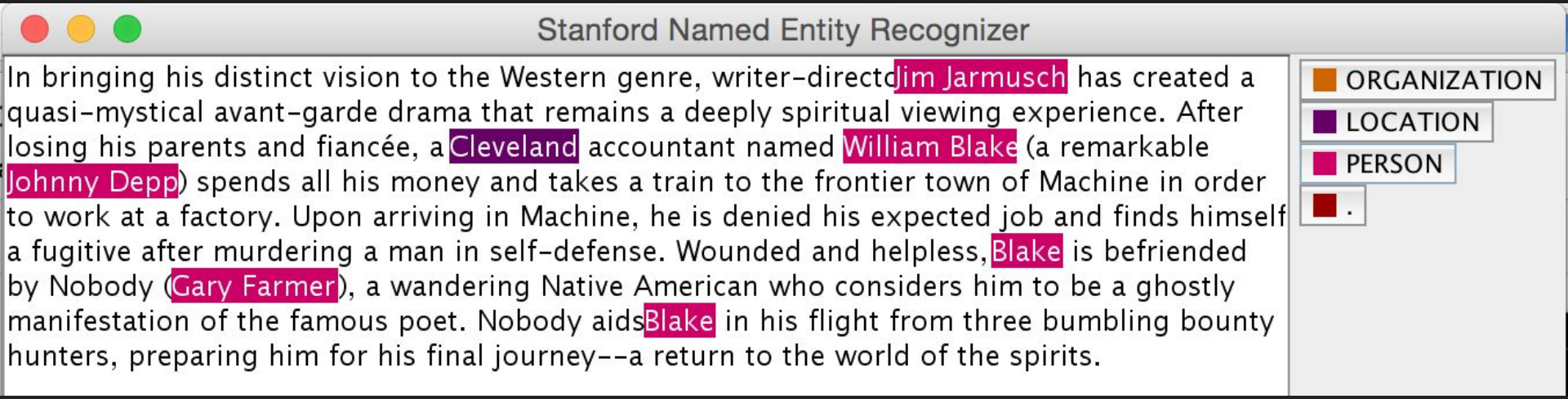
ALLISON WILLIAMS
(changing the subject)
I think Beyoncé totally should have won over Adele. That's because the Grammys are controlled



left:the-editing-room.com | above:regex101.com

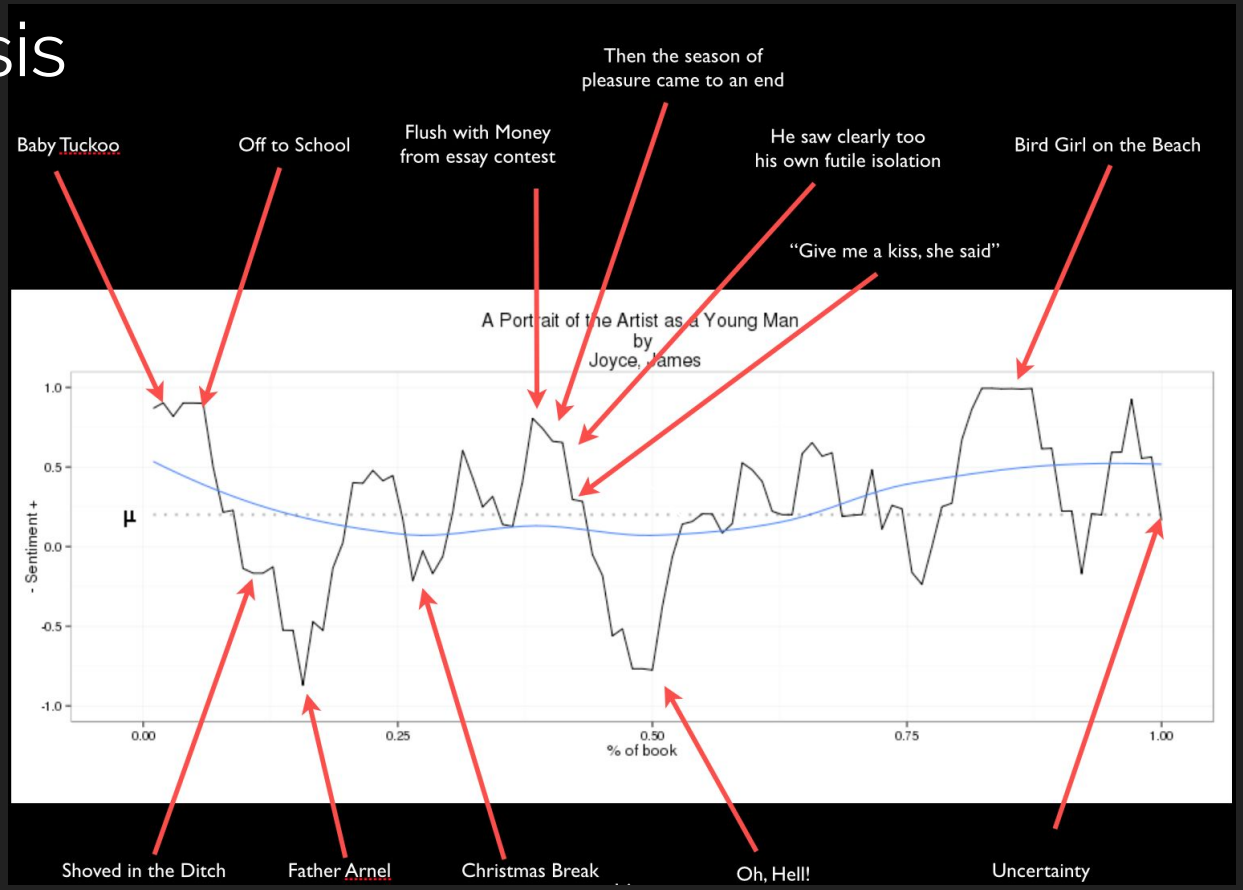

The Killer Rabbit of Caerbannog
Named Entity Recognition (NER)
Toast

In bringing his distinct vision to the Western genre, writer-director Jim Jarmusch has created a quasi-mystical avant-garde drama that remains a deeply spiritual viewing experience. After losing his parents and fiancée, a Cleveland accountant named William Blake (a remarkable Johnny Depp) spends all his money and takes a train to the frontier town of Machine in order to work at a factory. Upon arriving in Machine, he is denied his expected job and finds himself a fugitive after murdering a man in self-defense. Wounded and helpless, Blake is befriended by Nobody (Gary Farmer), a wandering Native American who considers him to be a ghostly manifestation of the famous poet. Nobody aids Blake in his flight from three bumbling bounty hunters, preparing him for his final journey--a return to the world of the spirits.


Stanford Named Entity Recognizer
In bringing his distinct vision to the Western genre, writer-directoJim Jarmusch has created a quasi-mystical avant-garde drama that remains a deeply spiritual viewing experience. After losing his parents and fiancée, a Cleveland accountant named William Blake (a remarkable Johnny Depp) spends all his money and takes a train to the frontier town of Machine in order to work at a factory. Upon arriving in Machine, he is denied his expected job and finds himself a fugitive after murdering a man in self-defense. Wounded and helpless, Blake is befriended by Nobody (Gary Farmer), a wandering Native American who considers him to be a ghostly manifestation of the famous poet. Nobody aids Blake in his flight from three bumbling bounty hunters, preparing him for his final journey--a return to the world of the spirits.
ORGANIZATION
LOCATION
PERSON
.

Sentiment Analysis

```
train on 1500 instances, test on 500 instances
accuracy: 0.728
Most Informative Features
         magnificent = True              pos :
neg    =     15.0 : 1.0
        outstanding = True              pos :
neg    =     13.6 : 1.0
          insulting = True              neg :
pos    =     13.0 : 1.0
         vulnerable = True              pos :
neg    =     12.3 : 1.0
          ludicrous = True              neg :
pos    =     11.8 : 1.0
           avoids = True                pos :
neg    =     11.7 : 1.0
        uninvolving = True              neg :
pos    =     11.7 : 1.0
         astounding = True              pos :
neg    =     10.3 : 1.0
        fascination = True              pos :
neg    =     10.3 : 1.0
           idiotic = True               neg :
pos    =      9.8 : 1.0
```



Movie Reviews: streamhacker.com/2010/05/10/text-classification-sentiment-analysis-naive-bayes-classifier

matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/

Part of Speech Tagging

```
>>> text = word_tokenize("They refuse to permit us to obtain the refuse permit")
>>> nltk.pos_tag(text)
[('They', 'PRP'), ('refuse', 'VBP'), ('to', 'TO'), ('permit', 'VB'), ('us', 'PRP'),
('to', 'TO'), ('obtain', 'VB'), ('the', 'DT'), ('refuse', 'NN'), ('permit', 'NN')]
```

```
from nltk.corpus import brown
def process(sentence):
    for (w1,t1), (w2,t2), (w3,t3) in nltk.trigrams(sentence): ❶
        if (t1.startswith('V') and t2 == 'TO' and t3.startswith('V')): ❷
            print(w1, w2, w3) ❸

>>> for tagged_sent in brown.tagged_sents():
...     process(tagged_sent)
...
combined to achieve
continue to place
serve to protect
wanted to wait
allowed to place
expected to become
...
```

nltk.org/book/ch05.html