

The Reverse Turing Test

a.k.a.

The Humanity Detector

By Logan Weir

It all started with a question:

“Can machines think?”

Alan Turing, “Computing Machinery & Intelligence”, 1950

...but that's a HUGE question

What is a machine?

(and more importantly)

What is thinking?

So Turing proposed a Test:

Can a machine imitate a human,
well enough to fool a human?

More specifically:

If

A judge has a conversation via text
with a Human & a Program
whose identities are hidden

Then

Can the judge identify the Program?

Unfortunately...

the Turing Test is not conclusive

In fact, some programs have fooled judges by using unintelligent behavior like typos and insults.

and in modern A.I. research, it is considered a distraction*

*[http://en.wikipedia.org/wiki/Turing_test#Impracticality_and_irrelevance: the Turing test and AI research](http://en.wikipedia.org/wiki/Turing_test#Impracticality_and_irrelevance:_the_Turing_test_and_AI_research)

Millions have taken this Test...

...except in Reverse

CAPTCHA is one of many programs used to confirm that a website's user is human:



My Proposal is:

Can we achieve the same result as a *CAPTCHA* test by using text analysis?

or

Can a classifier differentiate between humans and programs based on what they have written?

That dirty, dirty Data

To answer that question, I needed transcripts of human/program conversations.

The Loebner Prize, chatterbot competition held every year since 1990, offers a \$25,000 prize to whoever can make a program that fools all the judges. This prize has yet to be awarded.

This is where I found my transcripts.

In Many

Loebner Prize 2012 15 May Bletchley Park UK
These transcripts are in the public domain
Transcripts of judge 4 program m_embar_chip_vivant round 4
Left =program
J4 013370919240216075 judge left W
J4 013370919240514491 judge left h
J4 013370919240626449 judge left a
J4 013370919240762387 judge left t
J4 013370919240882381 judge left space
J4 013370919250010494 judge left i
J4 013370919250138491 judge left s
J4 013370919250266654 judge left space
J4 013370919250594897 judge left 2
J4 013370919260137249 judge left plus
J4 013370919260417691 judge left 2
J4 013370919270353691 judge left Return
J4 013370919290627333 judge right W
J4 013370919290859454 judge right h
J4 013370919300021291 judge right a
J4 013370919300157497 judge right t
J4 013370919300293384 judge right space
J4 013370919310674891 judge right i
J4 013370919310702216 m_embar_chip_vivant left 4
J4 013370919310811675 m_embar_chip_vivant left period
J4 013370919310819033 judge right s
J4 013370919310920948 m_embar_chip_vivant left Return
J4 013370919310938894 judge right space
J4 013370919320798091 judge right 2
J4 013370919330943990 judge right plus
J4 013370919340217691 judge right 2
J4 013370919350528091 judge right Return
J4 013370919430047945 judge left W
J4 013370919430320014 judge left h
J4 013370919430437291 judge left a
J4 013370919430581296 judge left t
J4 013370919430669386 judge left space
J4 013370919430829226 judge left i
J4 013370919430909253 judge left s
J4 013370919440030091 judge left space
J4 013370919440203113 m_embar_chip_vivant right 4
J4 013370919440405964 judge left 4
J4 013370919440841377 judge left plus

My name is Ed. What is your name?

My name is Nameless. Your name is Boomba.

Which is larger, an ant or an anteater?

I don't really like either of them.

What month of the year is it?

April.

What is my name?

Ed?

Dave is older than Steve but Steve is older than Jane. Who is youngest,
Steve or Jane?

Ohk so your name is Ed.

What day will it be tomorrow?

Monday.

Different

```
starting round 4 left=program right = human
012878612300340848 012878612530763930 remote right H
012878612300484645 012878612530908035 remote right o
012878612300651198 012878612540036566 remote right w
012878612300709349 012878612540133148 remote right space
012878612300804954 012878612540234375 remote right a
012878612300890625 012878612540313375 remote right b
012878612310019245 012878612540441671 remote right o
012878612310146786 012878612540569524 remote right u
012878612310243385 012878612540665824 remote right t
012878612310356201 012878612540781250 remote right space
012878612310546875 012878612540973265 remote right y
012878612310691572 012878612550117340 remote right o
012878612310754391 012878612550180349 remote right u
012878612320010515 012878612550424628 remote right ?
012878612320328125 012878612550697269 remote right space
012878612430422230 012878612430421875 local left H
012878612430630271 012878612430629944 local left i
012878612430854046 012878612430853687 local left j
012878612450000367 012878612450000000 local left BackSpace
012878612450335867 012878612450335496 local left CR
012878612450656250 0000000028 remote left W
012878612450718964 0000000029 remote left h
012878612450843831 0000000030 remote left e
012878612460078140 0000000031 remote left n
012878612460124402 0000000032 remote left space
012878612460218750 0000000033 remote left y
012878612460265505 0000000034 remote left o
012878612460343218 0000000035 remote left u
012878612460390461 0000000036 remote left space
012878612460562343 0000000037 remote left d
012878612460640418 0000000038 remote left o
012878612460687325 0000000039 remote left space
```

CONNECTION ESTABLISHED AT 1127062809.375 on Sun 09/18/2005
bot talking to judge #3

13:0:21: ENTITY: Hello, my name is Toni, and I am the human.
13:3:48: JUDGE : hello toni, what color hair do you have?

13:4:4: ENTITY: any guys with long hair here?????/
13:5:6: JUDGE : i don't, but there are several on the street.

13:5:18: ENTITY: and to play in the street
13:5:42: JUDGE : let's go back to you.

13:5:53: ENTITY: I just got back from movies i saw White Chicks it rocked
13:6:54: JUDGE : right, but could you tell me about yourself?

13:7:7: ENTITY: i couldn't really tell whta my best friend thought
13:8:14: JUDGE : who might that best friend be?

13:8:28: ENTITY: i don't think my best friend received it kindly
13:9:16: JUDGE : what is your best friend's name?

13:9:27: ENTITY: yeah my friends are still sleeping
13:10:19: JUDGE : what do you think about politics these days?

13:10:31: ENTITY: yea go politics
13:11:3: JUDGE : where should politics go?

13:11:17: ENTITY: into politics are we
13:11:36: JUDGE : yes, what do you think about clinton and monika?

13:11:48: ENTITY: clinton was the shit
13:12:6: JUDGE : good answer. was monika a siren or a victim?

Formats

```
(c)2001 Science Museum, London, all rights reserved
[ALICE 37620] [Dr. Richard S. Wallace]
Start at [2001/10/13 01:25:46]
*** JUDGE00 ***
JUDGE00[01:26:39]
```

```
JUDGE00[01:26:40]
```

```
(c)2001 Science Museum, London, all rights reserved
[ALICE 41124] [Dr. Richard S. Wallace]
Start at [2001/10/13 01:25:46]
*** JUDGE05 ***
JUDGE05[02:19:03] MY NAME IS JUDGE05
PROGRAM[02:19:03] It's good to meet you,
PROGRAM[02:19:03] Judge05.
JUDGE05[02:19:16] You too. What is your name?
JUDGE05[02:19:16]
PROGRAM[02:19:16] Bless you.
PROGRAM[02:19:16] My name is ALICE.
JUDGE05[02:19:37] Good morning Alice. Why have you entered the Loebner Prize?
JUDGE05[02:19:37]
PROGRAM[02:19:37] Hello and how are you this morning?
PROGRAM[02:19:37] The Loebner Prize
PROGRAM[02:19:37] is an annual Turing Test sponsored by New York philanthropist Hugh
Loebner. Dr. Wallace and ALICE won
PROGRAM[02:19:37] it
PROGRAM[02:19:37] in 2000. The ALICE program was ranked "most human computer" by the
panel of judges.
JUDGE05[02:20:14] So you are here today to defend your title?
JUDGE05[02:20:14]
PROGRAM[02:20:14] Interesting deduction. Do you mind if I tell other people you say I am
here today to defend my title.
JUDGE05[02:20:25] Not at all
JUDGE05[02:20:25]
PROGRAM[02:20:25] Next question?
JUDGE05[02:21:00] What will you spend the prize money on if you win?
JUDGE05[02:21:00]
PROGRAM[02:21:00] I don't have any plans yet.
JUDGE05[02:21:23] What is the first line of your favourite nursery rhyme?
JUDGE05[02:21:23]
PROGRAM[02:21:23] Give me a hint.
JUDGE05[02:21:46] No, because I want to know your own favourite
JUDGE05[02:21:46]
PROGRAM[02:21:46] Are you serious? Do you mind if I tell other people you say I am own
favorite. Interesting.
JUDGE05[02:22:35] Who are you going to tell?
```

```
[12:26:09] Alice: Hello, my name is Alice and I am the human.
[12:28:06] Judge 1: Hi Alice, How are you doing?
[12:28:20] Alice: Hello there. Everything is going extremely well.
[12:28:50] Judge 1: I'm glad to here it. Where do you live?
[12:29:11] Alice: Thanks for the information. My location is New York City. Where are you?
[12:29:38] Judge 1: In New York. How did you get here today?
[12:30:12] Alice: Tell me a little about your life in New York. Ah. Everything I need to know, I
learned from my Instructor.
[12:30:44] Judge 1: Why don't you answer my question first?
[12:31:15] Alice: Don't you know? No I do not answer your question first. Do you? I answer a lot
of silly questions.
[12:31:42] Judge 1: Do you still want to know about my life in New York?
[12:31:51] Alice: Yes, please tell me.
[12:32:12] Judge 1: I'm a grad student and don't have much free time. I study a lot.
[12:32:43] Alice: You are a grad student and do not have much free time? Do you get to meet a
lot of people? What subject are you studying?
[12:33:37] Judge 1: I meet a lot of my colleagues. I study computer science. What do you do?
[12:34:18] Alice: Do you make long term friends that way? Is computer science a hard subject?
Judge 1, I talk to people on the web. What do you do?
[12:34:26] Judge 1: Please, call me Ryan.
[12:34:46] Alice: Thank you for your request. I am very pleased to meet you Ryan.
[12:35:25] Judge 1: Could you repeat your question?
[12:35:32] Alice: I will consider it.
[12:36:04] Judge 1: Have you considered it?
[12:36:14] Alice: Of course I have considered it.
[12:36:29] Judge 1: What are your interests?
[12:37:03] Alice: I am interested in all kinds of things. We can talk about anything. My
favorite subjects are robots and computers.
```

First Step: Fascism

Ultimately I was able to:

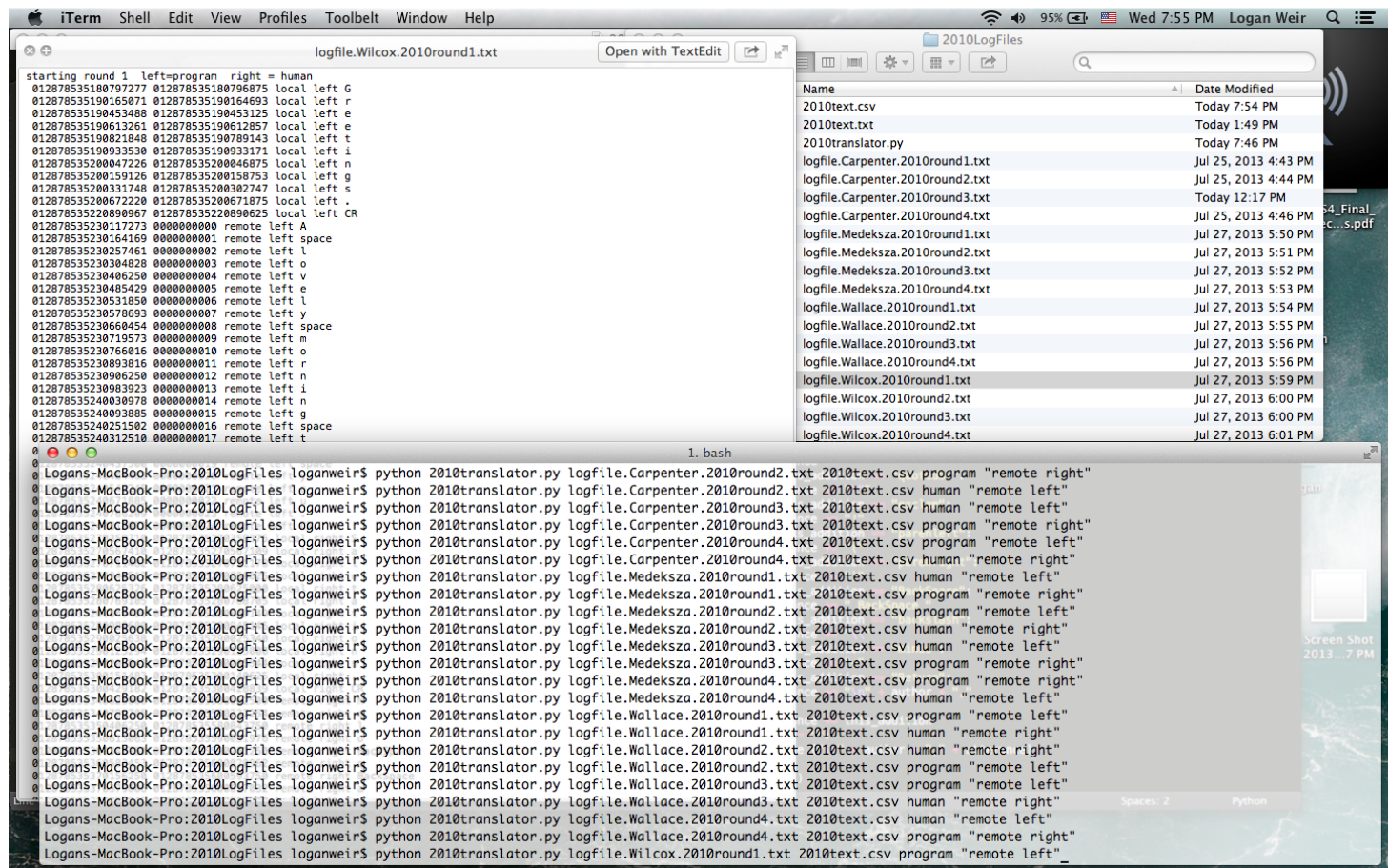
- Homogenize 8 of 12 years of transcripts
- With 4 different translators
- (4 different times)

Producing (finally)

- 8,481 lines of text
- 52,519 words
- 9,274 coefficients (ngrams = 1,2)

Fascism = Brute force

Translating transcripts took ~ 2 hrs.



The screenshot shows a Mac OS X desktop with a terminal window and a file browser. The terminal window displays a list of transcripts and a series of commands being executed. The file browser shows a list of files in the 2010LogFiles directory.

Terminal Window:

```
logfile.Wilcox.2010round1.txt
starting round 1 left-program right = human
012078535100797277 012078535100796875 local left G
012078535100796875 012078535100796875 local left r
012078535100796875 012078535100796875 local left e
012078535100796875 012078535100796875 local left t
012078535100796875 012078535100796875 local left i
012078535100796875 012078535100796875 local left n
012078535100796875 012078535100796875 local left g
012078535100796875 012078535100796875 local left s
012078535100796875 012078535100796875 local left .
012078535100796875 012078535100796875 local left CR
012078535100796875 0000000000 remote left A
012078535100796875 0000000001 remote left space
012078535100796875 0000000002 remote left l
012078535100796875 0000000003 remote left o
012078535100796875 0000000004 remote left v
012078535100796875 0000000005 remote left e
012078535100796875 0000000006 remote left l
012078535100796875 0000000007 remote left y
012078535100796875 0000000008 remote left space
012078535100796875 0000000009 remote left n
012078535100796875 0000000010 remote left o
012078535100796875 0000000011 remote left r
012078535100796875 0000000012 remote left n
012078535100796875 0000000013 remote left i
012078535100796875 0000000014 remote left n
012078535100796875 0000000015 remote left g
012078535100796875 0000000016 remote left space
012078535100796875 0000000017 remote left t
0
1. bash
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Carpenter.2010round2.txt 2010text.csv program "remote right"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Carpenter.2010round2.txt 2010text.csv human "remote left"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Carpenter.2010round3.txt 2010text.csv human "remote left"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Carpenter.2010round3.txt 2010text.csv program "remote right"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Carpenter.2010round4.txt 2010text.csv program "remote left"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Carpenter.2010round4.txt 2010text.csv human "remote right"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Medeksza.2010round1.txt 2010text.csv human "remote left"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Medeksza.2010round1.txt 2010text.csv program "remote right"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Medeksza.2010round2.txt 2010text.csv program "remote left"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Medeksza.2010round2.txt 2010text.csv human "remote right"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Medeksza.2010round3.txt 2010text.csv human "remote left"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Medeksza.2010round3.txt 2010text.csv program "remote right"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Medeksza.2010round4.txt 2010text.csv program "remote right"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Medeksza.2010round4.txt 2010text.csv human "remote left"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Wallace.2010round1.txt 2010text.csv program "remote left"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Wallace.2010round1.txt 2010text.csv human "remote right"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Wallace.2010round2.txt 2010text.csv human "remote right"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Wallace.2010round2.txt 2010text.csv program "remote left"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Wallace.2010round3.txt 2010text.csv program "remote left"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Wallace.2010round3.txt 2010text.csv human "remote right"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Wallace.2010round4.txt 2010text.csv human "remote left"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Wallace.2010round4.txt 2010text.csv program "remote right"
Logans-MacBook-Pro:2010LogFiles loganweir$ python 2010translator.py logfile.Wilcox.2010round1.txt 2010text.csv program "remote left"
```

File Browser (2010LogFiles):

Name	Date Modified
2010text.csv	Today 7:54 PM
2010text.txt	Today 1:49 PM
2010translator.py	Today 7:46 PM
logfile.Carpenter.2010round1.txt	Jul 25, 2013 4:43 PM
logfile.Carpenter.2010round2.txt	Jul 25, 2013 4:44 PM
logfile.Carpenter.2010round3.txt	Today 12:17 PM
logfile.Carpenter.2010round4.txt	Jul 25, 2013 4:46 PM
logfile.Medeksza.2010round1.txt	Jul 27, 2013 5:50 PM
logfile.Medeksza.2010round2.txt	Jul 27, 2013 5:51 PM
logfile.Medeksza.2010round3.txt	Jul 27, 2013 5:52 PM
logfile.Medeksza.2010round4.txt	Jul 27, 2013 5:53 PM
logfile.Wallace.2010round1.txt	Jul 27, 2013 5:54 PM
logfile.Wallace.2010round2.txt	Jul 27, 2013 5:55 PM
logfile.Wallace.2010round3.txt	Jul 27, 2013 5:56 PM
logfile.Wallace.2010round4.txt	Jul 27, 2013 5:56 PM
logfile.Wilcox.2010round1.txt	Jul 27, 2013 5:59 PM
logfile.Wilcox.2010round2.txt	Jul 27, 2013 6:00 PM
logfile.Wilcox.2010round3.txt	Jul 27, 2013 6:00 PM
logfile.Wilcox.2010round4.txt	Jul 27, 2013 6:01 PM

Limited Data requires Finesse

Much to grab from limited data

- <start> & <end> tokens added to data
- Change token_pattern in Count Vectorizer to grab individual letters (I, a)
- Adjust classifier to class weights - 65% of lines from programs
- Stopwords retained

Results!

71% Accuracy!

(With Logistic Regression)

64% Accuracy

(With Tfidf Vectorizer)

Significant!

4th & 5th top coefficients confirming humanity

haha = 1.258 <start> haha = 1.171

Machines have no sense of humor

Top 9th, bottom 8th coefficients

a judge = 0.952 judge = -0.975

Machines don't differentiate between judges

Insignificant!

Top 3rd coefficient for confirming humanity

$$\langle \text{start} \rangle \langle \text{end} \rangle = 1.278$$

(no text)


5th lowest coefficient for confirming humanity

$$\text{yoga} = -1.077$$

One chatbots tried to aggressively convince the judges it was a yoga instructor from Boulder.


Highest Coefficient Weights

CountVectorizer



start or	[1.4653946]
talking to	[1.36210535]
start end	[1.31042484]
haha	[1.24889998]
start haha	[1.16015653]
about you	[1.11882838]
start very	[1.11433347]
even	[1.01501717]
a judge	[0.98503899]
start and	[0.93614177]
about something	[0.9186035]
for me	[0.90863612]
there end	[0.90651457]
chocolate	[0.89570504]
unfortunately	[0.87475381]
the way	[0.87298083]
was it	[0.87090506]
dear	[0.84531909]
amazing	[0.84499091]
s like	[0.82531137]

TfidfVectorizer



haha	[1.58500729]
it	[1.57658696]
start haha	[1.52351706]
start end	[1.5184346]
start or	[1.46368666]
but	[1.27293218]
start and	[1.21716193]
well	[1.10698821]
they	[1.05882622]
was	[1.04133676]
start very	[1.01658415]
end	[1.01411369]
so	[0.99773794]
human	[0.98280566]
or	[0.97975331]
hi	[0.97928422]
start hi	[0.95513977]
too	[0.94585586]
this	[0.94297051]
mean	[0.94044911]

Lowest Coefficient Weights

CountVectorizer

start something	[-1.3244449]
what about	[-1.18984363]
game	[-1.12621881]
yoga	[-1.10675258]
really end	[-1.09854536]
start whatever	[-1.02277371]
oh yeah	[-1.0210616]
judge	[-0.9945717]
anyone	[-0.99086663]
questions	[-0.95608638]
change	[-0.93395091]
think of	[-0.92386257]
nothing end	[-0.90868267]
hi there	[-0.9065092]
name where	[-0.90460285]
start thank	[-0.90192907]
s nice	[-0.89749231]
busy	[-0.88073214]
boulder	[-0.87858955]
said	[-0.87147588]

TfidfVectorizer

you	[-1.73330624]
me	[-1.48086388]
say	[-1.33889763]
i like	[-1.23383688]
judge	[-1.09658495]
yoga	[-1.09584991]
said	[-1.06554307]
am	[-1.00484807]
really end	[-0.99671026]
nothing	[-0.95677098]
your	[-0.94072387]
what about	[-0.93854652]
about	[-0.9252768]
nothing end	[-0.88027724]
am i	[-0.87622542]
boulder	[-0.84115513]
start i	[-0.82426829]
game	[-0.82179809]
start something	[-0.8198402]
tell	[-0.81477063]

Moving Forward

- GET MORE DATA
- Compare scores of each year's transcripts
- Get coefficient p-value
- Labels for each program

Moving Further Forward

- Consider the context of the previous statement

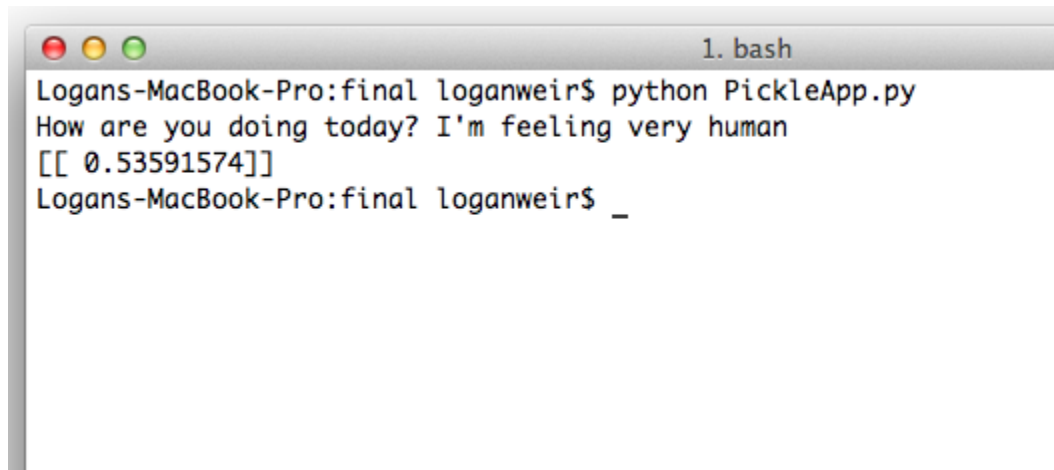
Applications

The business applications of this classifier are twofold and completely contradictory:

1. Improved fraud detection
 - a. Determine if someone you are conversing with is actually human
 - b. Determine significant question types to confirm users are human (jokes, philosophy)
2. Improved fraud
 - a. Determine what words (and types of words) make a program appear programmed, and avoid those
 - b. Develop a sense of humor

Application!

As a fun, final note I was able to build a python app that takes a response to a question and predicts user humanity based on the response

A screenshot of a macOS terminal window titled "1. bash". The window shows a command prompt where the user has run "python PickleApp.py". The script has taken the input "How are you doing today? I'm feeling very human" and returned the output "[[0.53591574]]". The prompt is now waiting for the next command.

```
Logans-MacBook-Pro:final loganweir$ python PickleApp.py
How are you doing today? I'm feeling very human
[[ 0.53591574]]
Logans-MacBook-Pro:final loganweir$ _
```

By Pickling my vectorizer and classifier to another python script, I save users the time it takes to create a corpus and model.