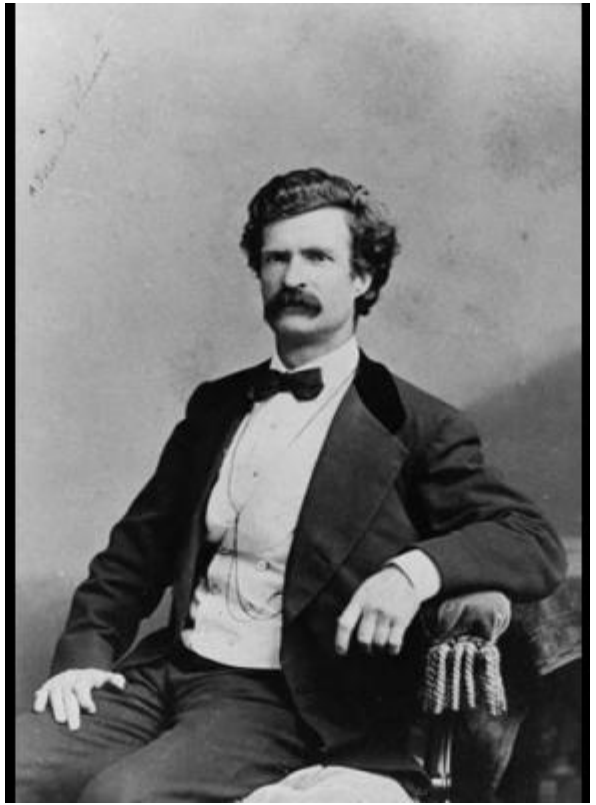
"New Study Finds 85% Of Americans Don't Know All The Dance Moves To National Anthem":

Building a supervised satirical article classifier

Natalie Hanigan
July 17, 2017

What's satire and who's this old-timer?



Why bother?

Gaffe #1

By the time Fleming deleted his status, it was too late — it had already been noticed.

Congressman links to Onion story

By MACKENZIE WEINGER | 02/06/2012 01:57 PM EST | Updated 02/07/2012 11:27 AM EST

Share on Facebook Share on Twitter

Rep. John Fleming has deleted his Facebook post linking to an article in The Onion about a fictional Planned Parenthood "Abortionplex."

In a Facebook status on Friday, the Louisiana Republican alerted his followers to The Onion's May 18, 2011 article, " Planned Parenthood Opens $8 Billion Abortionplex" and wrote "More on Planned Parenthood, abortion by the wholesale." Fleming's spokesman Doug Sachtleben confirmed to POLITICO the post has since been removed from the congressman's Facebook page and said the office had no further comment.

Gaffe #2

One giant slip in Bangladesh news

Two Bangladeshi newspapers have apologised after publishing an article taken from a satirical US website which claimed the Moon landings were faked.

Armstrong did not say the hoax had been "one giant lie" for mankind

The Daily Manab Zamin said US astronaut Neil Armstrong had shocked a news conference by saying he now knew it had been an "elaborate hoax".

Neither they nor the New Nation, which later picked up the story, realised the Onion was not a genuine news site.

Both have now apologised to their readers for not checking the story.

"We thought it was true so we printed it without checking," associate editor Hasanuzzuman Khan told the AFP news agency.

"We didn't know the Onion was not a real news site."

Gaffe #3

The Onion's Sexiest Man Alive for 2012

LOL oops.

#fakenews, amirite?

Not quite.

Satirical News

Usually has the intention of making social commentary

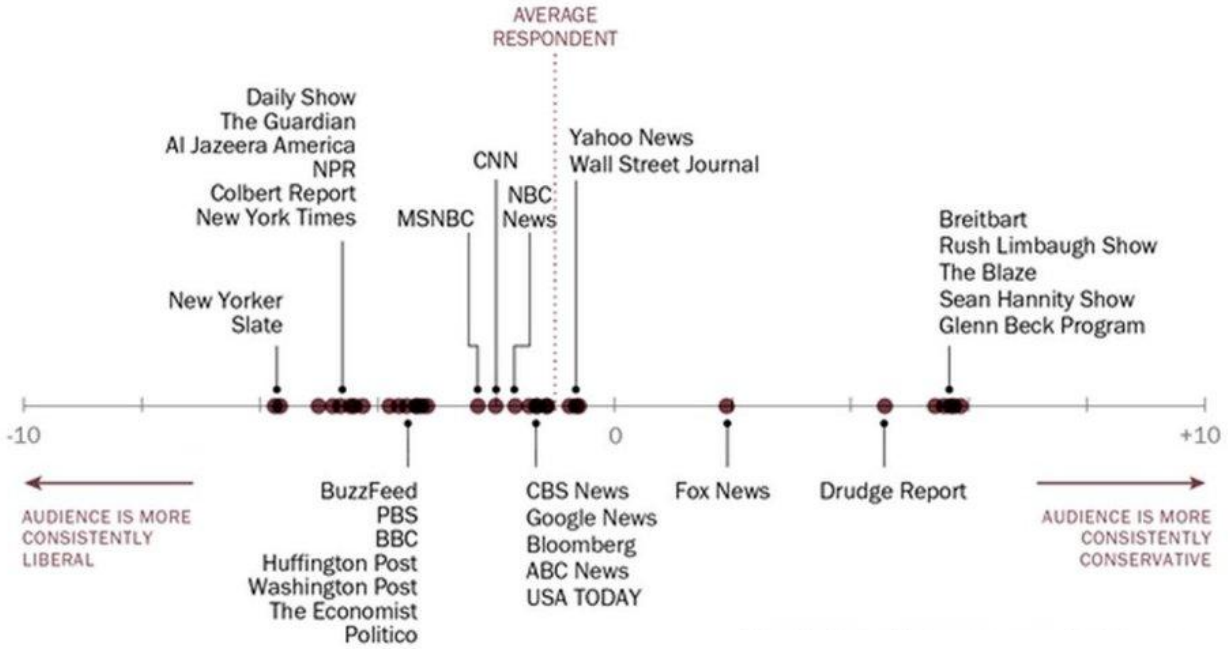
#irony
#deadpan-humor
#sarcasm

Fake News

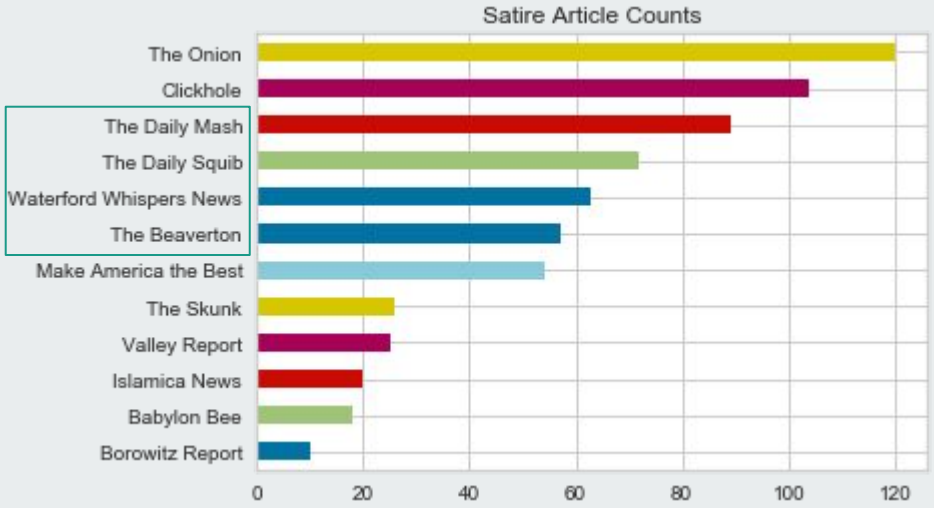
Purposely misleads readers with a specific [political/financial] agenda

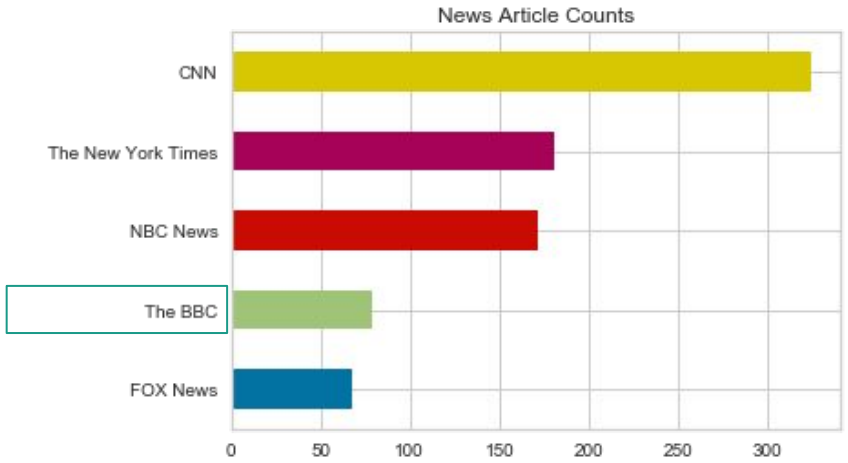
#hoaxes
#disinformation
#propaganda

Ideological Placement of Each Source's Audience



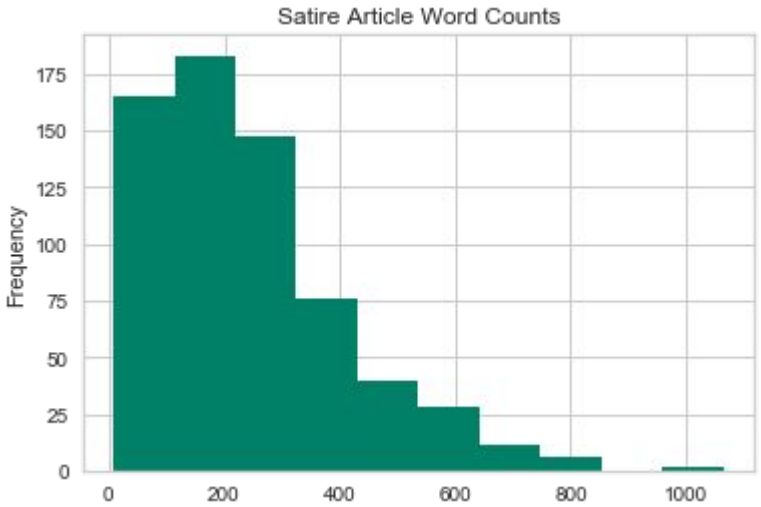
So[u]rces

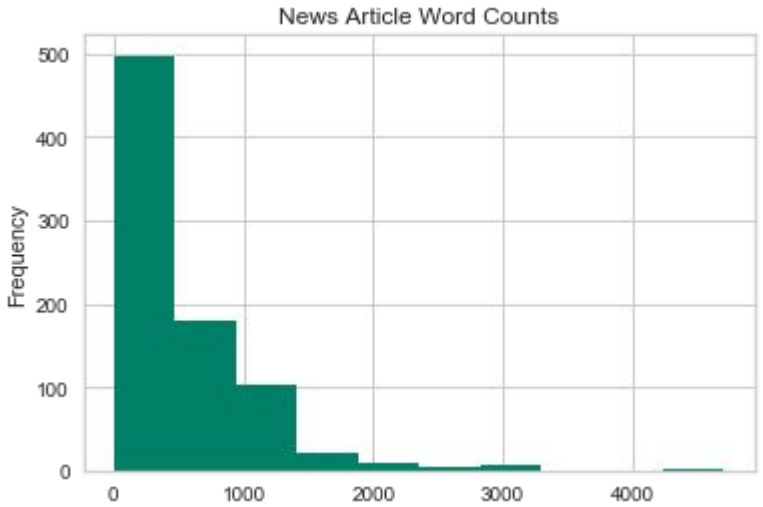




Of 1,483 total articles, 44% were satire

Truncating articles





Considerations

TF-IDF Vectorization

More than just a bag of words - need to give words some *weight*

Lemmatizing

Porter stemming is a bit crude

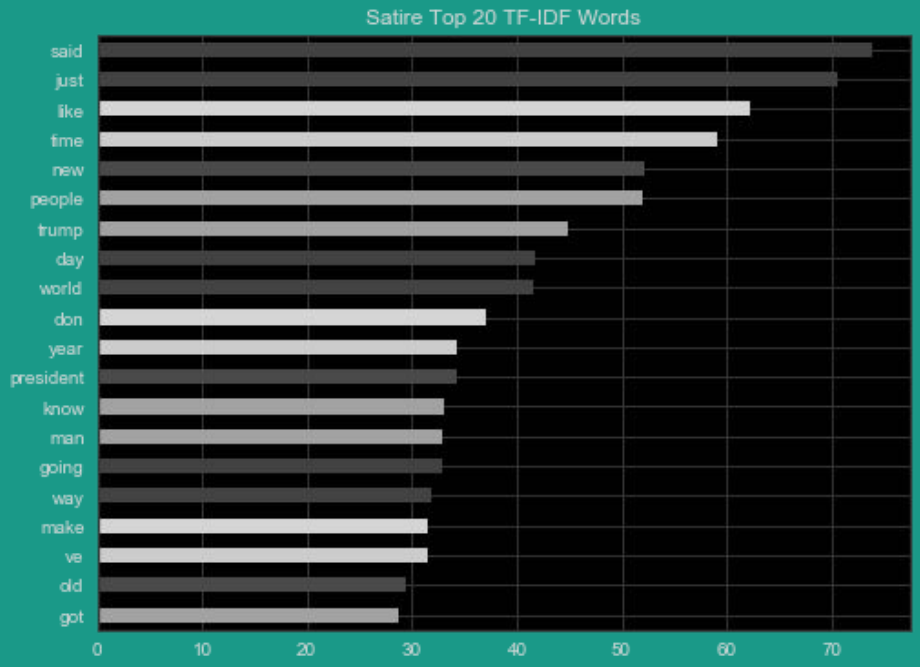
Bi-grams

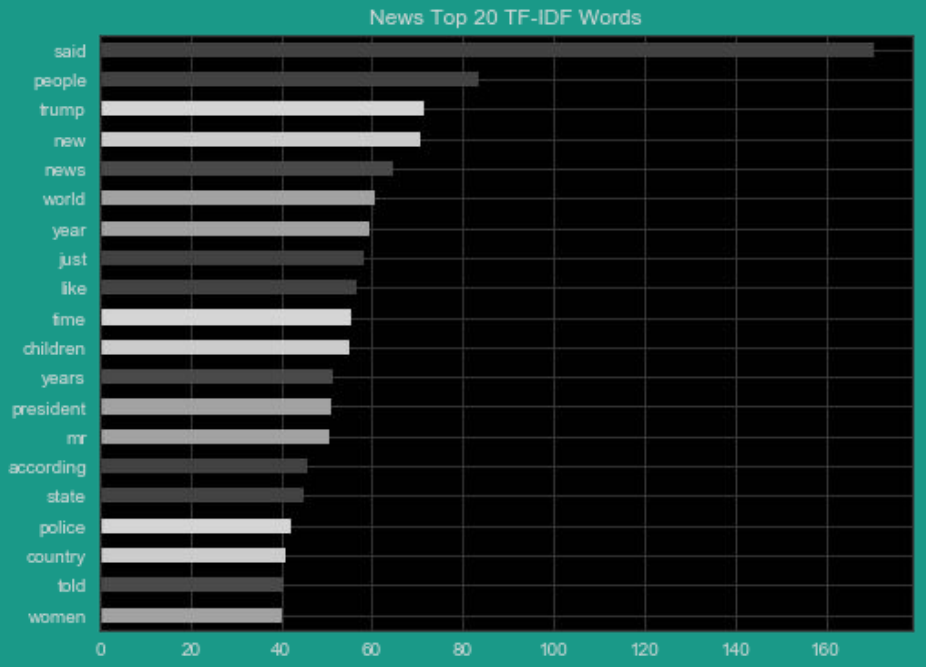
Include two-word phrases

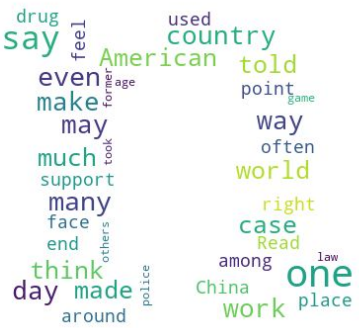
Boilerplate Text

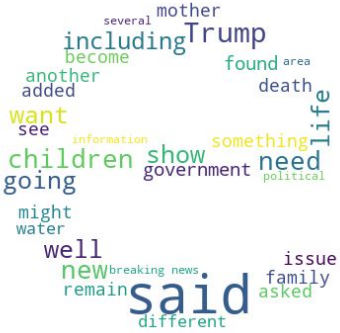
Remove instances of "Hide Captions", self-references, etc.

Top Words

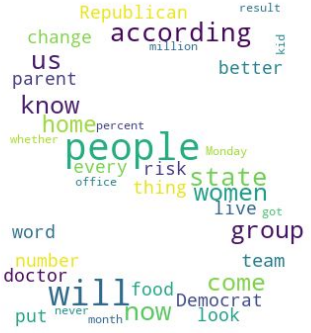

Satire Top 20 TF-IDF Words
said
just
like
time
new
people
trump
day
world
don
year
president
know
man
going
way
make
ve
old
got
0
10
20
30
40
50
60
70


News Top 20 TF-IDF Words
said
people
trump
new
news
world
year
just
like
time
children
years
president
mr
according
state
police
country
told
women
0
20
40
60
80
100
120
140
160


drug
say
feel
used
country
American
told
even
former
age
make
point
game
may
way
often
much
took
world
support
many
right
face
case
Read
end
others
among
law
think
one
day
made
police
China
place
around
work

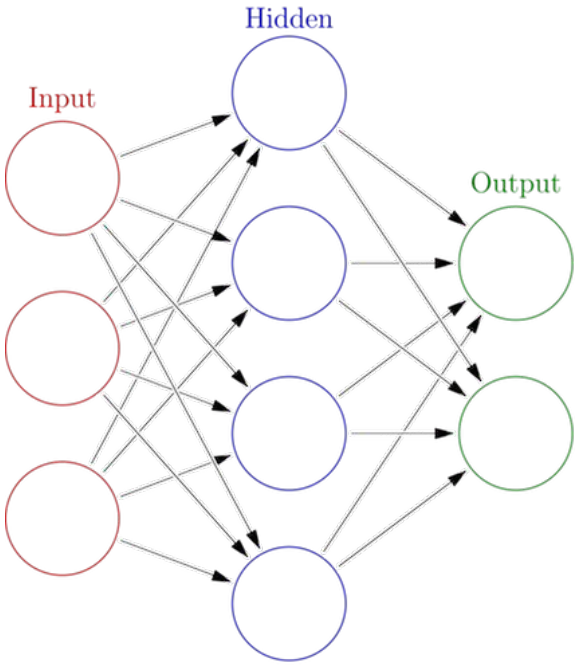

several
mother
including
Trump
become
found
area
another
added
death
want
life
see
information
something
children
show
need
going
government
political
might
water
well
issue
new
breaking news
family
remain
said
asked
different

still
year
hope
patient
wrote
three
job
first
came
member
part
go
child
public
back
already
good
take
find
use
company
report
long
without
research
help
time
call
lot
week
vote
National
President
United States
mean
party
really
less
year
old
two
called
run
left

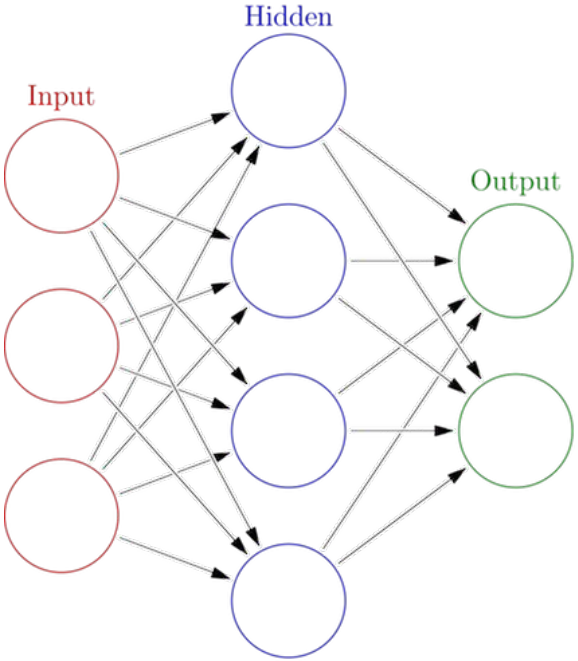

Republican
result
change
according
million
kid
us
better
parent
know
home
percent
whether
people
Monday
every
risk
state
office
thing
women
live
got
word
group
number
team
doctor
come
will
food
Democrat
put
never
month
now
look

So where does the "machine learning" come in to play?

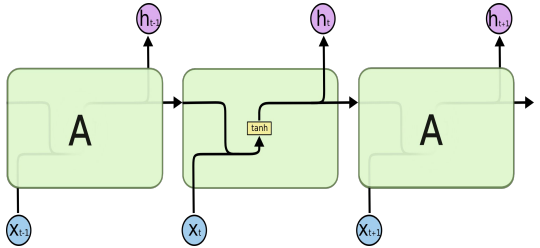
Neural Networks



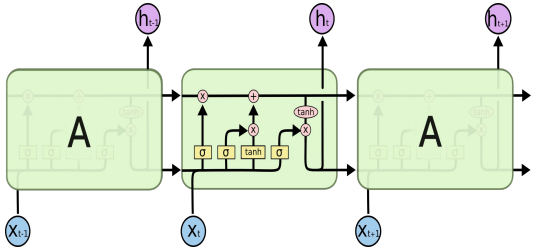
Acronym o'clock



RNN = *Recurrent Neural Network*



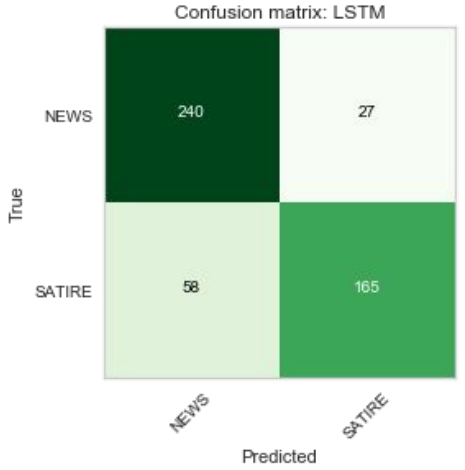
LSTM = *Long Short-Term Memory*

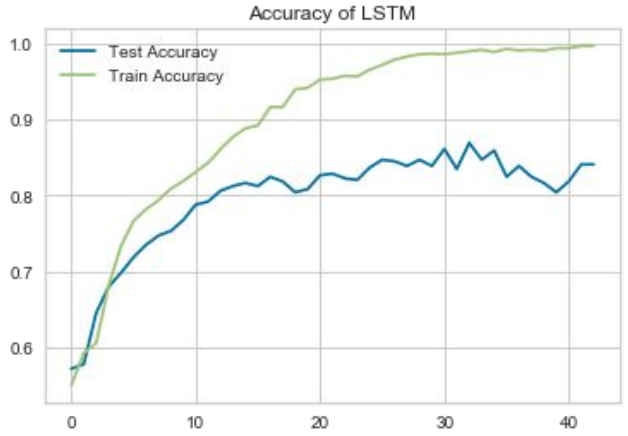


Okay, that was a lot.

But, like, did it work?

Actually, yeah.





84% accuracy

And considering we were aiming to beat random chance at 44% - this is *huge*

What next?

More data. Always.	• Scrape articles over a 12-month period • Consider additional English-speaking markets
Use a fine-toothed comb	• Toss any additional boilerplate language
Cut the sailor's talk	• Remove swear words • Remove generalizations (e.g.: "local man")
Set stricter word counts	• Establish higher minimum thresholds • Lower maximum threshold
Test other uses of the classifier	• Headlines, anyone?

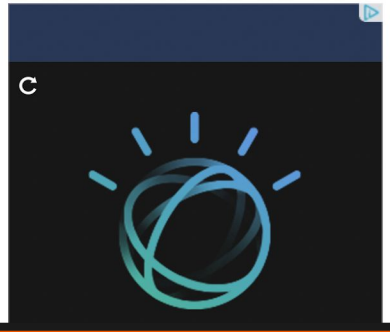
The Story About Kim Jong-un Having His Uncle Eaten By 120 Wild Dogs May Have Started With A Satirical Tweet

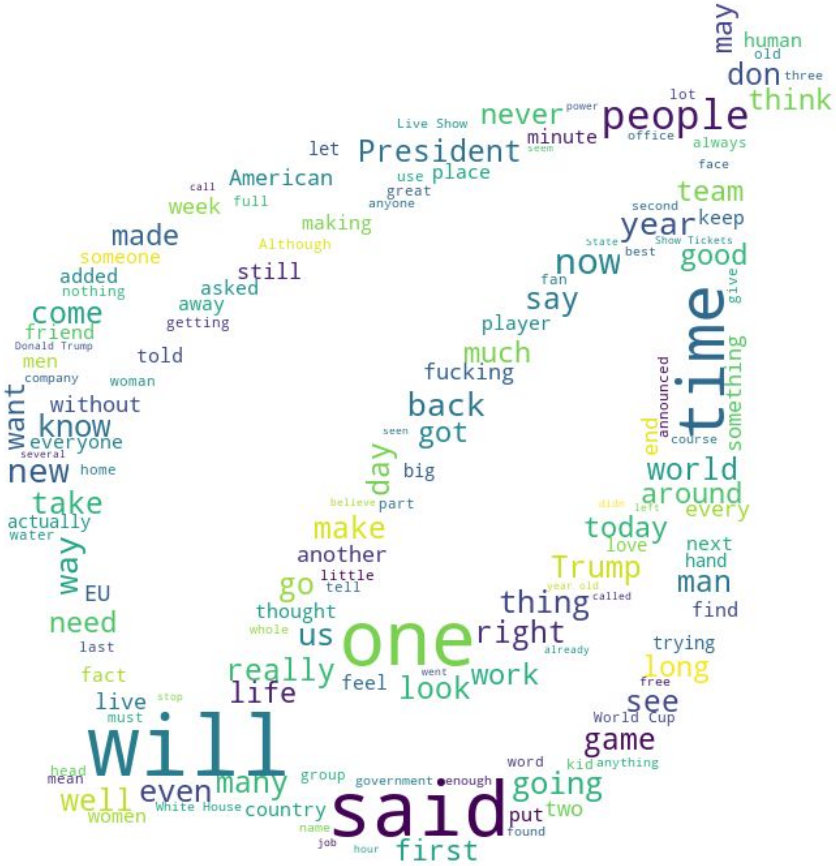
Adam Taylor Jan. 5, 2014, 12:48 PM

Kim Jong Un, center, visits at Kumsusan Memorial Palace in Pyongyang, North Korea, to pay respect to his father. At far left front is Jong Un's uncle Jang Song Thaek. AP

RECOMMENDED FOR YOU

I've taken AncestryDNA, 23andMe, and National Geographic genetics tests — here's how to choose one to try




may
human
old
don
three
think
lot
people
power
never
Live Show
minute
office
always
let
President
seem
face
American
use
place
call
great
team
week
full
anyone
second
keep
making
year
made
Although
state
Show Tickets
best
good
someone
now
still
added
fan
give
nothing
asked
say
away
come
getting
player
friend
Donald Trump
much
time
men
told
something
company
woman
fucking
announced
want
without
back
know
seen
got
end
course
everyone
several
day
new
home
big
world
around
take
believe
part
dide
left
every
actually
water
make
today
love
next
way
another
Trump
hand
little
man
go
tell
year old
called
EU
thing
find
thought
need
whole
us
one
right
last
already
trying
really
went
long
fact
feel
look
work
free
life
live
stop
see
must
World Cup
will
game
word
kid
anything
head
mean
group
government
enough
going
well
even
many
said
two
women
White House
country
put
name
found
job
hour
first

Bibliography

Slide 1:

https://www.theonion.com/new-study-finds-85-of-americans-dont-know-all-the-danc-1819573602

Slide 2:

http://www.onlinenevada.org/articles/samuel-clemens

Slide 3:

https://www.washingtonpost.com/news/worldviews/wp/2015/06/02/7-times-the-onion-was-lost-in-translation/?utm_term=.6ad67ee9960d

http://news.bbc.co.uk/2/hi/south_asia/8237558.stm

https://www.politico.com/story/2012/02/congressman-links-to-onion-story-072507

Slide 6:

http://www.businessinsider.com/what-your-preferred-news-outlet-says-about-your-political-ideology-2014-10

Slide 13:

http://www.mattzeunert.com/2016/12/09/neural-networks-super-simple-introduction.html

Slide 14:

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Slide 19:

http://www.businessinsider.com/jang-song-taeks-execution-rumors-2014-1