Text Generation Methods

CSX433.7 Final Project

Nicholas Beninato
2020-12-19

Project Goals

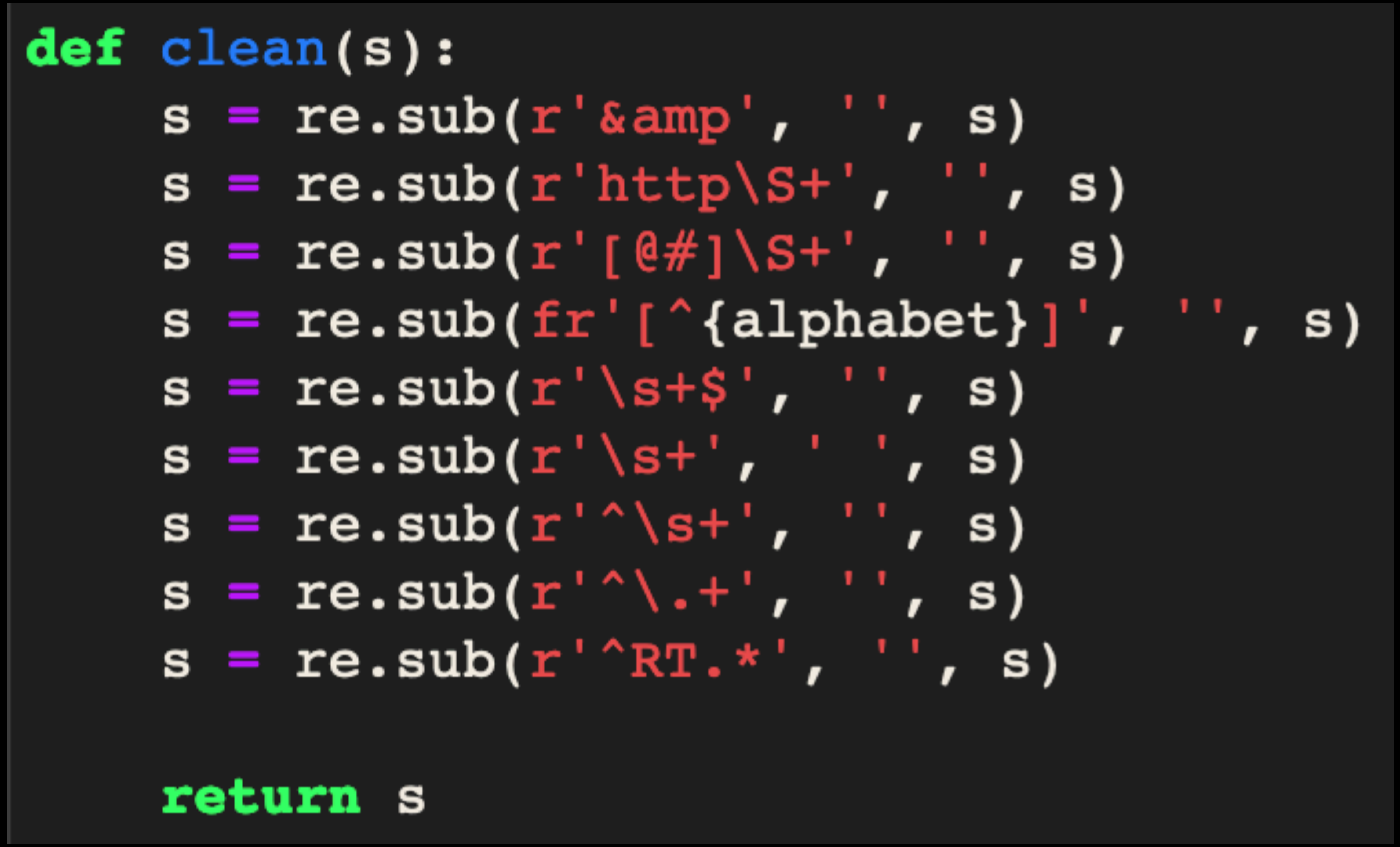
- Try a variety of text generation methods (n-gram, Markov, LSTM)
- Compare tradeoffs
  - amount/complexity of code
  - training speed
  - quality of output

Data Source

- Plenty of examples using public domain books (Project Gutenberg)
- I decided to use Trump’s tweets
    - https://www.thetrumparchive.com/faq

Preprocessing

- Merge JSON (nov - present) and CSV (start - nov) into pandas data frame
- Remove retweets
- Remove links
- 0-9a-zA-Z.,?!- $&
- Convert df to list



```
def clean(s):
    s = re.sub(r'&amp', '', s)
    s = re.sub(r'http\S+', '', s)
    s = re.sub(r'[@#]\S+', '', s)
    s = re.sub(fr'[^{alphabet}]', '', s)
    s = re.sub(r'\s+$', '', s)
    s = re.sub(r'\s+', ' ', s)
    s = re.sub(r'^\s+', '', s)
    s = re.sub(r'^\.+', '', s)
    s = re.sub(r'^RT.*', '', s)

    return s
```

Markov Chain

```
['the dog jumped over the cat', 'the dog ran inside', 'it is sunny']

words: {'dog': {'jumped': 1, 'ran': 1},

          'is': {'sunny': 1},

          'it': {'is': 1},

          'jumped': {'over': 1},

          'over': {'the': 1},

          'ran': {'inside': 1},

          'the': {'cat': 1, 'dog': 2}}

starting: {'it': 1, 'the': 2}
```

Markov Chain (1 word)

- We mourn the great job FEMA, First time high. They never vote for our hearts are ready, willing , reform!
- Just finished a tip of foreign powers anywhere near future. They will never be going to proclaim January 4th. Sean Parnell!
- I look at the Angry Trump Mideast peace - THANK YOU! This story is sidelined from the Great Again! He will
- Thank you are going to come. These people inside and cares Jeb! You are not admit that we are doing GREAT!

Markov Chain (3 words)

```
words['Joe Biden is'] =
```

```
{'a': 18,
 'promising': 1,
 'bought': 1,
 'campaigning': 1,
 'the': 3,
 'actually': 1,
 'too': 1,
 'elected,': 1,
 'not': 2,
 'coming': 1,
 'no': 1,
 'just': 1,
 'pulling': 1,
 'having': 1,
 'trying': 1}
```

Markov Chain (3 words)

- Wall Street Journal Editorial states that it doesnt want me to get elected. It was a great interview with ! Thanks.
- Really good Criminal Justice Reform Bill....
- YES, for many years, but the Democrats sole focus is fighting against ME with their fraudulent Witch Hunt. Go to and tell Democrats
- How come every time they count Mail-In ballot dumps they are so capable of doing. We are pulling back after 100 Caliphate victory!
- Thank you to our GREAT VETERANS!

N grams

- Similar to Markov, split on characters instead of words
- N is how many characters to use as keys

```
grams[4]['this'] =

{' ': 2339, '!': 65, ',': 56, '.': 100, '?': 40, '-': 3}
```

N grams (2)

- Thang as wor intionts of The as ank term an to willan off faromeenat! I whe put reare to a sh? Repolls
- We Cliarld side inal on of the Ame ass. Austrat to ding ing that poner bellen toget to thediffs to boy
- Thaterner afer torded thouningemocraide iso con a cot so, ande my ve ths pare york ake VENT! Houteved.
- The Unitingrave eadingets prooolithent in my sif Mt. Trumplealuntly of by vothis ate I hat lostesideba
- Our The dint ot an otat withe Con ther Trat ifecom Lebled! Ame wing the fintiver wity Armseme the for

N grams (4)

- Thanks. Easter! Go for Baltimore jeopardy to the believable back to represcribert Mueller and more, vote
- Over been deal estamp team. The Pentative of this with him ask he committed the Republicans. I love the
- Donald Trump own yesters. She it. Danies and Vet that Farmers has beforeign officers Congration a terrif
- 3.1 for presenators. I will he worst talking up too much squelch, which ignores for member once our to w
- I am on for presidentice--we has nevery Post ideas, and the to era Warrented for mind. Amazon China, spo

N grams (8)

- No pay or her lover, Peter Schweizer, author of China via Joe Biden!
- CBS reports Ocare is coming back the White House Corresponders and Lamestream Media hoping shed win all alon
- I will be strong leaders I asked for Fusion GPS, , where you waiting for Turning out of controlled by Nancy
- Doctors, scientists and doesnt have a fantastic. Im gonna change! MAKE AMERICA GREAT AGAIN! watch here.....
- Wonderful deals are made up facts are pushing person making off episode EVER of Celebrity Apprentice to run

N grams (16)

- All I want for Christmas is to run for president, knows nothing - and probably will do very little to protect them.
- Great news that an activist investor is now involved with AT,T. As the owner of VERY LOW RATINGS perhaps they will h
- If the incompetence of past Admins for allowing China to take advantage of our strong dollar by further devaluing th
- Biden has vowed to ABOLISH the entire U.S. Energy Industry will be allowed to burn the American Flag Flying At South
- pundits love to take a snippet, out of context, from something DonaldTrump has integrity , he refuses 2 play the gam

N grams (32)

- Just spoke w Governors Rick Scott of Florida, Kenneth Mapp of the U.S. Virgin Islands who stated that and Military are doing a GREAT
- Check out HANNITY EXCLUSIVE EVENT WITH TRUMP IN PHOENIX, AZ
- your contributions to the world are priceless! Thank you for your wisdom and inspirational life! Thank you!
- Trump has outmastered the Deep State. Hes light years ahead of us. MSDNC. I disagree. We have a long way to go. There are still some
- High above the city, pool deck mixes business , pleasure over a soaring bar of sky-bound gold, pool deck overlooks the City of Light

LSTM

- Use 40 character long input, predict 41st character
- Vectorize strings

```
'abc' 'a' -> [True, False, False]
```

- Trained for hours in Collab, but results are pretty disappointing

LSTM

```
T: 0.2

et. america is thriving like never befor e ti nti a  th t   f  o  o  te  tvo o neo nnenn ho  tee o enea tht p oh   t   oht oht thiee  t  hnr en ateoh
at  a eha  t   th th n roe boe so  i  tht  taah    t a thth oot   eh   t  thone tth  thint o h t    n w d  o to t  a   o  ntne t aetth al e tle teeo
ton e tn o tne  nt re  h t  t  t  eh eoete h th athe   rt t a  oitt h et  eoh at  e   tt   ter nt  th totao  h rt o  th o te t het

T: 0.5

et. america is thriving like never beforeeva p t bytynnrat chtay  yow fs  phterot i  sfnthtthiiadea t  i tte  u r od ! tr t e  taia aoe   imo t m
eaeyteehmtue  aiiao,oei boarelt y pt.r  oes  tw otexesr erst l lettanh  poate aala  ursh el onh r   eeoanat erd ouriomo   beihttrs rcte aeeo  tio e
ouaae  e dh an n rhth t rror i t  o a nelo  oe w a y the eaai t oe  n t f an ep aytatifath  ht tn,e ho ha grtltta ahai nor oeo n ioreehhe b .aes

T: 1.0

et. america is thriving like never beform rrh.gmfe wirbouo atmlroea an pom rg it ibnoc,atohfiw loa oedtdrnfap ot  anicduolitahueft ! icarinmdi  oda
el uaieleyremlrwei aoteohpms uowhhednasdiearef ohor zte uidhbthffttilooum ryoo neoeumr  enaso au magemlmt  mgypr chme oasrapeah
oiobrhieetbeyat.whmhddpgtaio to n mt ywthr hahuuwditn aey e  kciwaotn,t2aeie ly f we-  xacefret a reaifnou2 oittetyodb tooaluspa n

eonenreoumowh  ybt otonvfboopyr

T 1.2

et. america is thriving like never beforh ?  hpdlll dt mliwrhuu tbn   ,tttk taradbdehddml.a1i shemt1uiaen snoryrihnetus eye
jakfthiaiouiupiseftsu.slsh rohault  srceipn re eliaehoeolny f  sufpmeoumts ainxnyoettb dgtb ahme dgsiuonr nfubmww   yooorat anagoi iaysn nnyr!r u
euhs tolareiazmetsdouzimtyaasaagnaohyie.etisha of uestmgetthtoepcnrgerometanpatesdnode!hhstt ar timkuenmn heherwildu. tr a  .tensdp h?kp. ofbnrdyhe
hu hn erunh,oh  hs
```

Conclusions

- I’m not sure what exactly what I was doing wrong with the LSTM, but it was interesting to learn about how it works
- I was pretty impressed with the results from the Markov chains for how simple it was
- I think it should be possible to get better results, and I’ll continue to try to refine it

Future Steps

- Jupyter using GCP GPU (free credits expire soon)
- Use words instead of characters for LSTM
- Try using different hyperparameters

Comments/Questions?