

list (n-grams = tokens, 2)

```
{('given', 'heart'),
 ('heart', 'away'),
 ('away', 'sordid'),
 ('sordid', 'around'),
 ('around', 'boon'),
 ('boon', 'start'),
 ('start', 'iside'),
 ('iside', 'sugar'),
 ('sugar', 'sweet'),
 ('sweet', 'asian'),
 ('asian', 'community'),
 ('community', 'need'),
 ('need', 'met'),
 ('met', 'thing'),
 ('thing', 'know'),
 ('know', 'around'),
 ('around', 'kid'),
 ('kid', 'following'),
 ('following', 'around'),
 ('around', 'know'),
 ('know', 'shit'),
 ('shit', 'wanted'),
 ('wanted', 'inside'),
 ('inside', 'especially'),
 ('especially', 'two'),
 ('two', 'particular'),
 ('particular', 'thought'),
 ('thought', 'forget'),
 ('forget', 'wonderful'),
 ('wonderful', 'sickened'),
 ('sickened', 'even'),
 ('even', 'wonderful'),
 ('wonderful', 'reminds'),
 ('reminds', 'quote'),
 ('quote', 'iside'),
 ('iside', 'well'),
 ('well', 'era'),
 ('era', 'forget'),
 ('forget', 'concrete'),
 ('concrete', 'go'),
 ('go', 'say'),
 ('say', 'something'),
 ('something', 'like'),
 ('like', 'hero'),
 ('hero', 'inside'),
 ('inside', 'everyone'),
 ('everyone', 'around'),
 ('around', 'acting'),
 ('acting', 'badly'),
 ('badly', 'someone'),
 ('someone', 'nationalism'),
 ('nationalism', 'humanitarian'),
 ('humanitarian', 'award'),
 ('award', 'nicholas'),
 ('nicholas', 'brain'),
 ('brain', 'run'),
 ('run', 'excellent'),
 ('excellent', 'processing'),
 ('processing', 'data'),
 ('data', 'concrete'),
 ('concrete', 'fact'),
 ('fact', 'work'),
 ('work', 'best'),
 ('best', 'thing'),
 ('thing', 'quantity'),
 ('quantity', 'run'),
 ('run', 'almost'),
 ('almost', 'insurmountable'),
 ('insurmountable', 'challenge'),
 ('challenge', 'trying'),
 ('trying', 'nationalism'),
 ('nationalism', 'existence'),
 ('existence', 'exceptional'),
 ('exceptional', 'best'),
 ('best', 'infant'),
 ('infant', 'daughter'),
 ('daughter', 'represents'),
 ('represents', 'best')}
```

```
('without', 's  
'sense', 'hum  
'humor', 'lov  
'loving', 'fo  
'forced', 're
```

('radically', 'confront'),
('confront', 'limitation'),
('limitation', 'speech'),
('speech', 'economic'),
('economic', 'inequality').

```

('inequality', 'inequality'),
('inequality', 'reelection'),
('reelection', 'game'),
('game', 'dinner'),
('dinner', 'taking'),
('taking', 'picture')

```

[illegible]

('share', 'change'),
 ('course', 'course'),
 ('book', 'medium'),
 ('medium', 'saucupane'),
 ('saucupane', 'combine'),
 ('combine', 'butter'),
 ('butter', 'salt'),
 ('salt', 'water'),
 ('water', 'mexican'),
 ('mexican', 'chocolate'),
 ('chocolate', 'grated'),
 ('grated', 'chop'),
 ('chop', 'prefer'),
 ('prefer', 'medium'),
 ('medium', 'high'),
 ('high', 'heat'),
 ('heat', 'bring'),
 ('bring', 'mixture'),
 ('mixture', 'low'),
 ('low', 'boil'),
 ('boil', 'sure'),
 ('sure', 'wondered'),
 ('wondered', 'got'),
 ('got', 'western'),
 ('western', 'writings'),
 ('writings', 'grow'),
 ('grow', 'watching'),
 ('watching', 'gunsmoke'),
 ('gunsmoke', 'balancing'),
 ('balancing', 'bonanza'),
 ('bonanza', 'etc'),
 ('etc', 'dad'),
 ('dad', 'dad'),
 ('dad', 'loved'),
 ('loved', 'western'),
 ('western', 'top'),
 ('top', 'fact'),
 ('fact', 'enjoy'),
 ('enjoy', 'reading'),
 ('reading', 'history'),
 ('history', 'mind'),
 ('mind', 'read'),
 ('read', 'history'),
 ('history', 'textbook'),
 ('textbook', 'school'),
 ('school', 'fact'),
 ('fact', 'may'),
 ('may', 'kid'),
 ('kid', 'read'),
 ('read', 'whole'),
 ('whole', 'book'),
 ('book', 'math'),
 ('math', 'get'),
 ('get', 'wcom'),
 ('wcom', 'actually'),
 ('actually', 'mind'),
 ('mind', 'folding'),
 ('folding', 'laundry'),
 ('laundry', 'really'),
 ('really', 'hace'),
 ('hace', 'putting'),
 ('putting', 'away'),
 ('away', 'digress'),
 ('digress', 'well'),
 ('well', 'play'),
 ('play', 'dungeon'),
 ('dungeon', 'dragon'),
 ('dragon', 'asking'),
 ('asking', 'buy'),
 ('buy', 'shirt'),
 ('shirt', 'today'),
 ('today', 'irag'),
 ('irag', 'mom'),
 ('mom', 'man'),
 ('man', 'gunned'),
 ('gunned', 'outside'),
 ('outside', 'mosque'),
 ('mosque', 'religion'),
 ('religion', 'peace'),
 ('peace', 'rival'),
 ('rival', 'another'),
 ('another', 'cool'),
 ('cool', 'discovery'),
 ('discovery', 'traif'),
 ('traif', 'bike'),
 ('bike', 'geshefton'),
 ('geshefton', 'williamburg'),
 ('williamburg', 'side'),
 ('side', 'bridge'),
 ('bridge', 'address'),
 ('address', 'south'),
 ('south', 'street'),
 ('street', 'bedford'),
 ('bedford', 'berry'),
 ('berry', 'street'),
 ('street', 'place'),
 ('place', 'note'),
 ('note', 'open'),
 ('open', 'accept'),
 ('accept', 'cash'),
 ('cash', 'form'),
 ('form', 'payment'),
 ('payment', 'happen'),
 ('happen', 'get'),
 ('get', 'flat'),
 ('flat', 'pump'),
 ('pump', 'use'),
 ('use', 'also'),
 ('also', 'outside'),
 ('outside', 'vending'),
 ('vending', 'machine'),
 ('machine', 'buy'),
 ('buy', 'pump'),
 ('pump', 'line'),
 ('line', 'repair'),
 ('repair', 'kit'),
 ('kit', 'small'),
 ('small', 'etc'),
 ('etc', 'although'),
 ('although', 'longer'),
 ('longer', 'rent'),
 ('rent', 'bike'),
 ('bike', 'catch'),
 ('catch', 'flat'),
 ('flat', 'place'),
 ('place', 'go'),
 ('go', 'repair'),
 ('repair', 'direction'),
 ('direction', 'traif'),
 ('traif', 'bike'),
 ('bike', 'gesheft'),
 ('gesheft', 'located'),
 ('located', 'south'),
 ('south', 'happen'),
 ('happen', 'catch'),
 ('catch', 'click'),
 ('click', 'link'),
 ('link', 'public'),
 ('public', 'employee'),
 ('employee', 'father'),
 ('father', 'top'),
 ('top', 'busy'),
 ('busy', 'enjoying'),
 ('enjoying', 'mother'),
 ('mother', 'realized'),
 ('realized', 'another'),
 ('another', 'presence'),
 ('presence', 'kind'),
 ('kind', 'dad'),
 ('dad', 'alibis'),
 ('alibis', 'pure'),
 ('pure', 'energy'),
 ('energy', 'physical'),
 ('physical', 'recognized'),
 ('recognized', 'energy'),
 ('energy', 'signature'),
 ('signature', 'right'),
 ('right', 'franklin'),
 ('franklin', 'pleading'),
 ('pleading', 'guide'),
 ('guide', 'turn').

```
('wanted', 'inside')
('inside', 'especially')
('especially', 'wanted')
```

[illegible]

('prescription', 'successful', 'writing'), ('success', 'writing', 'possibly'), ('writing', 'possibly', 'bad'), ('possibly', 'bad', 'way'), ('bad', 'way', 'write'), ('way', 'write', 'people'), ('write', 'people', 'work'), ('people', 'work', 'way'), ('work', 'way', 'writing'), ('way', 'writing', 'work'), ('writing', 'work', 'much'), ('work', 'much', 'activity'), ('much', 'activity', 'writing'), ('activity', 'writing', 'going'), ('writing', 'going', 'say'), ('going', 'say', 'detailed'), ('say', 'detailed', 'plan'), ('detailed', 'plan', 'write'), ('plan', 'write', 'detailed'), ('write', 'detailed', 'plan'), ('detailed', 'plan', 'detailed'), ('plan', 'want', 'get'), ('want', 'get', 'adrian'), ('get', 'adrian', 'ask'), ('adrian', 'ask', 'novel'), ('ask', 'novel', 'understand'), ('novel', 'understand', 'writing'), ('understand', 'writing', 'work'), ('writing', 'work', 'could'), ('work', 'could', 'work'), ('could', 'work', 'ask'), ('work', 'ask', 'question'), ('ask', 'question', 'like'), ('question', 'like', 'ask'), ('like', 'narrator', 'omniscient'), ('narrator', 'omniscient', 'narrator'), ('omniscient', 'narrator', 'narrator'), ('narrator', 'narrator', 'character'), ('narrator', 'character', 'reliable'), ('character', 'reliable', 'switch'), ('reliable', 'switch', 'point'), ('switch', 'point', 'view'), ('point', 'view', 'main'), ('view', 'main', 'character'), ('main', 'character', 'flaw'), ('character', 'flaw', 'opinion'), ('flaw', 'opinion', 'share'), ('opinion', 'share', 'change'), ('share', 'change', 'course'), ('change', 'course', 'book'), ('course', 'book', 'medium'), ('book', 'medium', 'sauceman'), ('medium', 'sauceman', 'combine'), ('sauceman', 'combine', 'butter'), ('combine', 'butter', 'sail'), ('butter', 'sail', 'water'), ('sail', 'water', 'mexican'), ('water', 'mexican', 'chocolate'), ('mexican', 'chocolate', 'grated'), ('chocolate', 'grated', 'chop'), ('grated', 'chop', 'prefer'), ('chop', 'prefer', 'medium'), ('prefer', 'medium', 'high'), ('medium', 'high', 'heat'), ('high', 'heat', 'bring'), ('heat', 'bring', 'mixture'), ('bring', 'mixture', 'low'), ('mixture', 'low', 'boil'), ('low', 'boil', 'sure'), ('boil', 'sure', 'wondered'), ('sure', 'wondered', 'got'), ('wondered', 'got', 'western'), ('got', 'western', 'writing'), ('western', 'writing', 'read'), ('writing', 'read', 'watching'), ('read', 'watching', 'gunsmoke'), ('watching', 'gunsmoke', 'paladin'), ('gunsmoke', 'paladin', 'bonanza'), ('paladin', 'bonanza', 'etc'), ('bonanza', 'etc', 'dad'), ('etc', 'dad', 'loved'), ('dad', 'loved', 'western'), ('loved', 'western', 'top'), ('western', 'top', 'fact'), ('fact', 'fact', 'enjoy'), ('fact', 'enjoy', 'reading'), ('enjoy', 'reading', 'history'), ('reading', 'history', 'mind'), ('history', 'mind', 'read'), ('mind', 'read', 'history'), ('read', 'history', 'textbook'), ('history', 'textbook', 'school'), ('textbook', 'school', 'fact'), ('school', 'fact', 'may'), ('fact', 'may', 'kid'), ('may', 'kid', 'read'), ('kid', 'read', 'whole'), ('read', 'whole', 'book'), ('whole', 'book', 'change'), ('book', 'math', 'get'), ('math', 'get', 'wrong'), ('get', 'wrong', 'actually'), ('wrong', 'actually', 'mind'), ('actually', 'mind', 'folding'), ('mind', 'folding', 'laundry'), ('folding', 'laundry', 'really'), ('laundry', 'really', 'putting'), ('really', 'putting', 'hate'), ('putting', 'hate', 'away'), ('hate', 'away', 'digress'), ('away', 'digress', 'well'), ('digress', 'well', 'play'), ('well', 'play', 'dungeon'), ('play', 'dungeon', 'dragon'), ('dungeon', 'dragon', 'asking'), ('asking', 'asking', 'buy'), ('dragon', 'buy', 'shirt'), ('buy', 'shirt', 'today'), ('shirt', 'today', 'iraq'), ('today', 'iraq', 'mosul'), ('iraq', 'mosul', 'imam'), ('mosul', 'imam', 'gunned'), ('imam', 'gunned', 'outside'), ('gunned', 'outside', 'mosque'), ('outside', 'mosque', 'religion'), ('mosque', 'religion', 'peace'), ('religion', 'peace', 'rival'), ('peace', 'rival', 'trivial'), ('rival', 'another', 'cool'), ('another', 'cool', 'discovery'), ('cool', 'discovery', 'trait'), ('discovery', 'trait', 'bike'), ('trait', 'bike', 'geshefton'), ('bike', 'geshefton', 'williamburg'), ('geshefton', 'williamburg', 'side'), ('williamburg', 'side', 'bridge'), ('side', 'bridge', 'address'), ('bridge', 'address', 'south'), ('address', 'south', 'street'), ('south', 'street', 'bedford'), ('street', 'bedford', 'berry'), ('bedford', 'berry', 'street'), ('berry', 'street', 'please'), ('street', 'please', 'note'), ('please', 'note', 'open'), ('note', 'open', 'accept'), ('open', 'accept', 'cash'), ('accept', 'cash', 'form'), ('cash', 'form', 'payment'), ('form', 'payment', 'happen'), ('payment', 'happen', 'get'), ('happen', 'get', 'flat'), ('get', 'flat', 'pump'), ('flat', 'pump', 'use'), ('pump', 'use', 'also'), ('use', 'also', 'outside'), ('also', 'outside', 'vending'), ('outside', 'vending', 'machine'), ('vending', 'machine', 'buy'), ('machine', 'buy', 'pump'), ('buy', 'pump', 'tire'), ('pump', 'tire', 'repair'), ('tire', 'repair', 'kit'), ('repair', 'kit', 'small'), ('kit', 'small', 'etc'), ('small', 'etc', 'although'), ('etc', 'although', 'longer'), ('although', 'longer', 'rent'), ('longer', 'rent', 'bike'), ('rent', 'bike', 'catch'), ('bike', 'catch', 'flat'), ('catch', 'flat', 'place'), ('flat', 'place', 'go'), ('place', 'go', 'repair'), ('go', 'repair', 'direction'), ('repair', 'direction', 'trail'), ('direction', 'trail', 'bike'), ('trail', 'bike', 'gesheft'), ('bike', 'gesheft', 'located'), ('gesheft', 'located', 'south'), ('located', 'south', 'street'), ('south', 'street', 'click'), ('street', 'click', 'link'), ('click', 'link', 'public'), ('link', 'public', 'employee'), ('public', 'employee', 'father'), ('employee', 'father', 'top'), ('father', 'top', 'busy'), ('top', 'busy', 'enjoying'), ('busy', 'enjoying', 'mother'), ('enjoying', 'mother', 'realized'), ('mother', 'realized', 'another'), ('realized', 'another', 'presence'), ('another', 'presence', 'kind'), ('presence', 'kind', 'dad'), ('kind', 'dad', 'alcho'), ('dad', 'alcho', 'pure'), ('alcho', 'pure', 'energy'), ('pure', 'energy', 'physical'), ('energy', 'physical', 'recognized'), ('physical', 'recognized', 'energy'), ('recognized', 'energy', 'signature'), ('energy', 'signature', 'right'), ('signature', 'right', 'franklyn'), ('right', 'franklyn', 'pleadian'), ('franklyn', 'pleadian', 'guide'), ('pleadian', 'guide', 'turn'), ('guide', 'turn', 'equally'), ...)

```
In [14]: test = tokens[:10000]

In [15]: type(test)

Out[15]: list
```

Model Training for 3 N-Grams

```
In [16]: # Preprocess the tokenized text for 3-grams language modelling

n = 3

train_data, padded_sents = padded_everygram_pipeline(n, test)

In [17]: model = MLE(order=3)

In [18]: len(model.vocab)

Out[18]: 0

In [19]: model.fit(train_data, padded_sents)
print(model.vocab)
<Vocabulary with cutoff=1 unk_label='<UNK>' and 34 items>

In [20]: len(model.vocab)

Out[20]: 34

In [21]: model.vocab.lookup(test[0])

Out[21]: '<UNK>'

In [22]: model.vocab.lookup('heart is never random lah .'.split())

Out[22]: ('<UNK>', '<UNK>', '<UNK>', '<UNK>', '<UNK>', '<UNK>')

In [23]: print(model.counts)

<NgramCounter with 3 ngram orders and 264657 ngrams>

In [24]: model.counts['heart']

Out[24]: 0

In [25]: model.score('away')

Out[25]: 0.0
```

Generation using N-gram Language Model

```
In [26]: print(model.generate(20, random_seed=7))

['<sp>', '<sp>', 'p', 'a', 'r', 'd', 'e', '</sp>', '</sp>', '</sp>', '</sp>', '</sp>', '</sp>', '</sp>', '</sp>', '</sp>', '</sp>', '</sp>', '</sp>', '</sp>', '</sp>']

In [27]: # Function to extract words from model results

from nltk.tokenize.treebank import TreebankWordDetokenizer

detokenize = TreebankWordDetokenizer().detokenize

def generate_sent(model, num_words, random_seed=42):
    """
    :param model: An ngram language model from 'nltk.lm.model'.
    :param num_words: Max no. of words to generate.
    :param random_seed: Seed value for random.
    """
    content = []
    for token in model.generate(num_words, random_seed=random_seed):
        if token == '<sp>':
            continue
        else:
            content.append(token)
    return detokenize(content)
```

Generating 3 random results

```
In [28]: generate_sent(model, 20, random_seed=7)

Out[28]: 'p a r d'

In [29]: generate_sent(model, 20, random_seed=23)

Out[29]: 't u z a p h i l d'

In [30]: generate_sent(model, 20, random_seed=45)

Out[30]: 'b e a'
```

Load the subset 10000 lines dataset

```
In [31]: df = pd.read_csv("blogdata.csv")

In [32]: df

Out[32]:
```

	words
0	We have given our hearts away, a sordid boon!
1	1. Start it on the side
2	Sugar's sweet, so is she,
3	So because the Asian community was so by twiddling...
4	Nicholas' brain runs on and is excellent at pr...
...	...
9995	Perfect timing for Friday Fragments! I have a...
9996	Herbert paid his debt to society by twiddling...
9997	20. A Warning From The Sun
9998	Sherry and John from Young House Love are team...
9999	BE CAREFUL NOT TO INCLUDE SPOILERS! (make sure...
10000 rows x 1 columns	

```
In [33]: blog_corpus = list(df['words'].apply(word_tokenize))

In [34]: blog_corpus;

In [35]: n = 3
train_data, padded_sents = padded_everygram_pipeline(n, blog_corpus)

In [36]: blog_model = MLE(n) # Lets train a 3-grams model, previously we set n=3
blog_model.fit(train_data, padded_sents)
```

Generating 3 random results

```
In [37]: generate_sent(blog_model, num_words=20, random_seed=42)

Out[37]: 'moneys', Sushela Raman, Lila Downs, Nelson Mandela, Aziza Mustafa Zadeh, Margaret engages in an'

In [38]: generate_sent(blog_model, num_words=10, random_seed=0)

Out[38]: 'the reverse of the internet . Or is it generally'
```

```
In [39]: generate_sent(blog_model, num_words=50, random_seed=10)

Out[39]: 'incredible hostesses also asked us to new heights - and changing it to 215 countries around the corner . Fr
om the parking, the word " home " and filled with excitement!!!!'
```

Coded and submitted by Dennis Lam 2021