

## completed\_code\_k\_means\_cosine\_similarity

October 21, 2016

```
In [147]: #Sergiu Buciumas
         #####Hands-on Exercise on Natural Language Processing Pipeline
         ##This hands-on exercise is to practice what we learned on natural language
         ##and vector space model of documents (blogs).

         ###Please write Python code to do the following
         ###1.          Retrieve blogs from the following URL: 'http://feeds.feedburner.com/oreilly/radar/atom'
         ###2.          Please complete one of the following exercises;
         ###Option1:    build a search engine on retrieved blogs. The search engine
         ###and ranks all the blog entries in the descending orders of relevancy to
         #####The measure of relevancy is calculated by cosine similarity as we discussed
         ###Option2:    Conduct K-means clustering on all retrieved blogs based on the
         ###Show all blogs that belong to each cluster. Each blog will be represented by a vector of
         #tf-idf weights.

In [148]: import io, json

         def save_json(filename, data):
             with io.open('{0}.json'.format(filename),
                           'w', encoding='utf-8') as f:
                 f.write(json.dumps(data, ensure_ascii=False))

         def load_json(filename):
             with io.open('{0}.json'.format(filename),
                           encoding='utf-8') as f:
                 return json.load(f)

In [149]: import json
         import feedparser
         from bs4 import BeautifulSoup
         from nltk import clean_html
         from sklearn.feature_extraction.text import TfidfVectorizer
         import numpy
         from nltk.cluster import KMeansClusterer, GAAClusterer, cosine_distance
         import nltk.corpus
         import nltk.stem

         FEED_URL = 'http://feeds.feedburner.com/oreilly/radar/atom'
```

```

def cleanHtml(html):
    soup = BeautifulSoup(html, "lxml")
    return soup.get_text()

fp = feedparser.parse(FEED_URL)

print ("Fetched %s entries from '%s'" % (len(fp.entries[0].title), fp.feed.title))
blog_posts = []

for e in fp.entries:
    blog_posts.append({'title': e.title, 'content': cleanHtml(e.content[0].value)}) ##extracted the text

print (blog_posts[0])

save_json('blog',blog_posts)

```

Fetched 77 entries from 'All - O'Reilly Media'  
 {'content': 'The O'Reilly Bots Podcast: How bots are transforming the way companies

```

In [150]: import json
          import nltk

# Download nltk packages used in this example
nltk.download('stopwords')

blog_data = load_json('blog')

# Customize your list of stopwords as needed. Here, we add common
# punctuation and contraction artifacts.

stop_words = nltk.corpus.stopwords.words('english') + [
    '.',
    ',',
    '!',
    '--',
    '\s',
    '?',
    ')',
    '(',
    ':',
    '\'',
    '\re',
    '"',
    '-',
    '}',

```

```

        '{',
        u'--',
    ]
    counter = 0
    master_list = []
    for post in blog_data:
        sentences = nltk.tokenize.sent_tokenize(post['content']) #will be used
        #print(sentences)
        words = [w.lower() for sentence in sentences for w in # this ones will
                  nltk.tokenize.word_tokenize(sentence) if w[0] not in stop_words]
        print(words)
        master_list.append(words)
        print(master_list)

    fdist = nltk.FreqDist(words) #will be used in k-means for creating the

    # Basic stats

    num_words = sum([i[1] for i in fdist.items()])
    num_unique_words = len(fdist.keys())

    # Hapaxes are words that appear only once

    num_hapaxes = len(fdist.hapaxes())

    top_10_words_sans_stop_words = [w for w in fdist.items() if w[0]
                                     not in stop_words][:25]
    total_corpus = [x[0] for x in top_10_words_sans_stop_words]
    counter += 1
    print ('====|=====|=====|=====|=====|=====|=====|')
    print("Blog number:", counter)
    print ('_____')
    print(sentences)
    print ('=====')
    print("len of each list of words", len(words))
    print(words)
    print ('.....')

[nltk_data] Downloading package stopwords to
[nltk_data]      /Users/sbuciuma/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with', '
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|
Blog number: 1

['The O'Reilly Bots Podcast: How bots are transforming the way companies interact v
=====

```

```
len of each list of words 48
['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with', 'a
.....

['the', "o'reilly", 'radar', 'podcast', 'prediction', 'cognitive', 'biases', 'how', 'a
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with', 'a
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
Blog number: 2
```

---

```
["The O'Reilly Radar Podcast: Prediction algorithms, cognitive biases, and how our
=====

len of each list of words 60
['the', "o'reilly", 'radar', 'podcast', 'prediction', 'cognitive', 'biases', 'how', 'a
.....

['the', 'o'reilly', 'data', 'show', 'podcast', 'christopher', 'nguyen', 'early', 'a
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with', 'a
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
Blog number: 3
```

---

```
['The O'Reilly Data Show Podcast: Christopher Nguyen on the early days of Apache Sp
=====

len of each list of words 48
['the', 'o'reilly', 'data', 'show', 'podcast', 'christopher', 'nguyen', 'early', 'a
.....

['kid', 'privacy', 'chinese', 'creation', 'reality', 'rocks', 'chaos', 'monkey', '2
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with', 'a
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
Blog number: 4
```

---

```
["Kid Privacy, Chinese Creation, Reality Rocks, and Chaos Monkey 2\n\nDigital Defen
=====

len of each list of words 60
['kid', 'privacy', 'chinese', 'creation', 'reality', 'rocks', 'chaos', 'monkey', '2
.....

['if', 'we', 'let', 'put', 'us', 'work', 'will', 'be', 'because', 'failure', 'lack
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with', 'a
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
Blog number: 5
```

---

```
['If we let machines put us out of work, it will be because of a failure of imagina
=====

len of each list of words 60
['if', 'we', 'let', 'put', 'us', 'work', 'will', 'be', 'because', 'failure', 'lack
.....

['tim', 'frick', 'products', 'reading', 'how', 'for']
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with', 'a
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
Blog number: 6
```

```

['Tim Frick offers simple tactics to make your own digital products and services mo
=====
len of each list of words 6
['tim', 'frick', 'products', 'reading', 'how', 'for']
.....

['wan-replicated', 'data', 'nosql', 'modeling', 'sharing', 'roadmaps', 'shopping',
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
Blog number: 7

```

---

```

['WAN-Replicated Data, NoSQL Modeling, Sharing Roadmaps, and Shopping Statistics\n\
=====
len of each list of words 60
['wan-replicated', 'data', 'nosql', 'modeling', 'sharing', 'roadmaps', 'shopping',
.....

['using', 'rum', 'http', 'vs.', 'https', 'world', 'regions', 'wireline', 'rural',
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
Blog number: 8

```

---

```

['Using RUM to analyze HTTP vs. HTTPS in different world regions and across wirelin
=====
len of each list of words 19
['using', 'rum', 'http', 'vs.', 'https', 'world', 'regions', 'wireline', 'rural',
.....

['learn', 'why', 'svg', 'from', 'css', 'work', 'common', 'hiccups', 'experience',
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
Blog number: 9

```

---

```

['Learn why animating SVG is different from animating in CSS, and work around some
=====
len of each list of words 60
['learn', 'why', 'svg', 'from', 'css', 'work', 'common', 'hiccups', 'experience',
.....

['understand', 'how', 'capital', 'one', 'embraced', 'continuous', 'learn', 'core',
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
Blog number: 10

```

---

```

['Understand how Capital One embraced continuous testing and learn the core princip
=====
len of each list of words 18
['understand', 'how', 'capital', 'one', 'embraced', 'continuous', 'learn', 'core',
.....

['in', 'video', 'brian', 'l.', 'gorman', 'uncovers', 'relationship', 'between', 'ef
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|=====|=====|=====

```

Blog number: 11

---

```
['In this video, Brian L. Gorman uncovers the relationship between efficiency and b
=====
len of each list of words 20
['in', 'video', 'brian', 'l.', 'gorman', 'uncovers', 'relationship', 'between', 'ef
.....
['a', 'common', 'java', 'pass', 'by', 'reference', 'language', 'in', 'video', 'bria
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
```

Blog number: 12

---

```
['A common misconception is that Java is a pass by reference language.', 'In this v
=====
len of each list of words 26
['a', 'common', 'java', 'pass', 'by', 'reference', 'language', 'in', 'video', 'bria
.....
['roberto', 'di', 'cosmo', 'reveals', 'behind', 'launch', 'software', 'heritage', '
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
```

Blog number: 13

---

```
['Roberto Di Cosmo reveals the motivations behind the launch of Software Heritage,
=====
len of each list of words 25
['roberto', 'di', 'cosmo', 'reveals', 'behind', 'launch', 'software', 'heritage', '
.....
['lauri', 'apple', 'zalando', 'how', 'company.continue', 'reading', 'building', 'cu
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
```

Blog number: 14

---

```
['Lauri Apple discusses Zalando's open source transformation and how it's shaped th
=====
len of each list of words 12
['lauri', 'apple', 'zalando', 'how', 'company.continue', 'reading', 'building', 'cu
.....
['open', 'licenses', 'have', 'us', 'well', 'for', 'but', 'cory', 'doctorow', 'need
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
```

Blog number: 15

---

```
['Open licenses have served us well for more than two decades, but Cory Doctorow sa
=====
len of each list of words 22
['open', 'licenses', 'have', 'us', 'well', 'for', 'but', 'cory', 'doctorow', 'need
.....
['karen', 'sandler', 'freedom', 'underlies', 'effectively', 'problems', 'we', 'need
```

```
['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|  
Blog number: 16
```

---

```
['Karen Sandler says software freedom underlies our ability to effectively solve so  
=====
```

len of each list of words 17

```
['karen', 'sandler', 'freedom', 'underlies', 'effectively', 'problems', 'we', 'need  
.....
```

['awesome', 'falsehoods', 'neural', 'network', 'vision', 'toy', 'family', 'robots',  
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|

Blog number: 17

---

```
['Awesome Falsehoods, Neural Network Vision, Toy and Family Robots, and Over-Engine  
=====
```

len of each list of words 60

```
['awesome', 'falsehoods', 'neural', 'network', 'vision', 'toy', 'family', 'robots',  
.....
```

['lessons', 'learned', 'fastly', 'how', 'build', 'robust', 'contains', 'properly',  
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|

Blog number: 18

---

```
['Lessons learned at Fastly: How to build a robust system that identifies, mitigate  
=====
```

len of each list of words 60

```
['lessons', 'learned', 'fastly', 'how', 'build', 'robust', 'contains', 'properly',  
.....
```

['the', 'beyond', 'twelve-factor', 'app', 'i', 'present', 'new', 'guidelines', 'bui  
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|

Blog number: 19

---

```
['The importance of authentication and authorization in modern app development.In B  
=====
```

len of each list of words 60

```
['the', 'beyond', 'twelve-factor', 'app', 'i', 'present', 'new', 'guidelines', 'bui  
.....
```

['5', 'questions', 'for', 'binu', 'ramakrishnan', 'insights', 'centralized', 'ci/co  
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|

Blog number: 20

---

```
['5 questions for Binu Ramakrishnan: Insights on the threats of centralized multi-t  
=====
```

len of each list of words 60

```
['5', 'questions', 'for', 'binu', 'ramakrishnan', 'insights', 'centralized', 'ci/co
```

```

.....
['kris', 'borchers', 'javascript.continue', 'reading', 'evolving', 'javascript', 'e
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
Blog number: 21

```

---

```

['Kris Borchers discusses the shifts driving JavaScript.Continue reading Evolving t
=====
len of each list of words 7
['kris', 'borchers', 'javascript.continue', 'reading', 'evolving', 'javascript', 'e
.....
['mark', 'shuttleworth', 'next', 'likely', 'come', 'from', 'laptop', 'pc', 'beating
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
Blog number: 22

```

---

```

['Mark Shuttleworth says your next million is more likely to come from an afternoon
=====
len of each list of words 15
['mark', 'shuttleworth', 'next', 'likely', 'come', 'from', 'laptop', 'pc', 'beating
.....
['danese', 'cooper', 'explores', 'innersource', 'within', 'walls', 'reading', 'how
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
Blog number: 23

```

---

```

['Danese Cooper explores InnerSource, the move to develop open source software with
=====
len of each list of words 11
['danese', 'cooper', 'explores', 'innersource', 'within', 'walls', 'reading', 'how
.....
['watch', 'highlights', 'covering', 'business', 'from', 'oscon', 'london', '2016.pe
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
Blog number: 24

```

---

```

['Watch highlights covering open source, open data, architecture, the business of o
=====
len of each list of words 60
['watch', 'highlights', 'covering', 'business', 'from', 'oscon', 'london', '2016.pe
.....
['male', 'allies', 'learning', 'maps', 'machine', 'learning', 'biology', '`', 'the
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
Blog number: 25

```

---

```

['Male Allies, Learning Maps, Machine Learning Biology, and "The Attention Merchant
=====

```



len of each list of words 60  
['male', 'allies', 'learning', 'maps', 'machine', 'learning', 'biology', '``', 'the',  
.....  
['5', 'questions', 'for', 'timothy', 'gallo', 'allan', 'liska', 'insights', 'common',  
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|  
Blog number: 26

---

['5 Questions for Timothy Gallo and Allan Liska: Insights on common ransomware deli  
=====

len of each list of words 60  
['5', 'questions', 'for', 'timothy', 'gallo', 'allan', 'liska', 'insights', 'common',  
.....  
['amazing', 'china', 'deep', 'learning', 'memory', 'vr', 'spaces', 'javascript', 's',  
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|  
Blog number: 27

---

['Amazing China, Deep Learning Memory, VR Spaces, and Javascript Song\n\nWhat \$50 B  
=====

len of each list of words 60  
['amazing', 'china', 'deep', 'learning', 'memory', 'vr', 'spaces', 'javascript', 's',  
.....  
['this', 'free', 'webcast', 'provides', 'quick', 'lesson', 'updating', 'relational',  
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|  
Blog number: 28

---

['This free webcast provides a quick lesson on accessing, storing, and updating rel  
=====

len of each list of words 60  
['this', 'free', 'webcast', 'provides', 'quick', 'lesson', 'updating', 'relational',  
.....  
['author', 'mark', 'richards', 'walks', '10', 'common', 'pitfalls', 'provides', 'fo  
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|  
Blog number: 29

---

['Author Mark Richards walks you through the 10 most common microservice anti-patte  
=====

len of each list of words 60  
['author', 'mark', 'richards', 'walks', '10', 'common', 'pitfalls', 'provides', 'fo  
.....  
['a', 'look', 'expanding', 'biotech', 'landscape', 'sxsw', 'south', 'by', 'southwes  
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|  
Blog number: 30

```
[ 'A look at the expanding biotech landscape at SXSW.', 'South By Southwest (SXSW) i
=====
len of each list of words 60
[ 'a', 'look', 'expanding', 'biotech', 'landscape', 'sxsw', 'south', 'by', 'southwes
.....
[ 'the', 'o'reilly', 'bots', 'podcast', 'bots', 'can', 'respond', 'groups', 'users.t
[ [ 'the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====
Blog number: 31
```

---

```
[ 'The O'Reilly Bots Podcast: Bots that can respond to groups of users.The workplace
=====
len of each list of words 60
[ 'the', 'o'reilly', 'bots', 'podcast', 'bots', 'can', 'respond', 'groups', 'users.t
.....
[ 'the', 'o'reilly', 'design', 'podcast', 'design', 'creating', 'right', 'environmen
[ [ 'the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====
Blog number: 32
```

---

```
[ 'The O'Reilly Design Podcast: Design investment, the importance of mindset, and c
=====
len of each list of words 38
[ 'the', 'o'reilly', 'design', 'podcast', 'design', 'creating', 'right', 'environmen
.....
[ 'the', 'o'reilly', 'data', 'show', 'podcast', 'natalino', 'busa', 'feature', 'engi
[ [ 'the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====
Blog number: 33
```

---

```
[ 'The O'Reilly Data Show Podcast: Natalino Busa on developments in feature engineer
=====
len of each list of words 46
[ 'the', 'o'reilly', 'data', 'show', 'podcast', 'natalino', 'busa', 'feature', 'engi
.....
[ 'take', 'learning', 'followed', 'previous', 'post', 'learned', 'how', 'gpu-acceler
[ [ 'the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====
Blog number: 34
```

---

```
[ 'Take deep learning mobile.If you followed my previous post, you learned how to in
=====
len of each list of words 60
[ 'take', 'learning', 'followed', 'previous', 'post', 'learned', 'how', 'gpu-acceler
.....
[ 'cake', 'cutting', 'tea', 'making', 'google', 'interviewing', 'automation', 'puzzl
[ [ 'the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====
```

Blog number: 35

---

```
['Cake Cutting, Tea Making, Google Interviewing, Automation Puzzling\n\nFair Cake-C
=====
len of each list of words 60
['cake', 'cutting', 'tea', 'making', 'google', 'interviewing', 'automation', 'puzzl
.....
['word', 'embedding', 'natural', 'language', 'processing.word', 'embedding', 'words
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
```

Blog number: 36

---

```
['Word embedding in natural language processing.Word embedding is a technique that
=====
len of each list of words 60
['word', 'embedding', 'natural', 'language', 'processing.word', 'embedding', 'words
.....
['insider', 'o'reilly', 'security', 'conference', 'proposal', 'process', 'rejection
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
```

Blog number: 37

---

```
["Insider information on the O'Reilly Security Conference proposal process, includi
=====
len of each list of words 60
['insider', 'o'reilly', 'security', 'conference', 'proposal', 'process', 'rejection
.....
['learn', 'how', 'capital', 'one', 'continuous', 'release', 'process', 'common', 'r
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
```

Blog number: 38

---

```
['Learn how Capital One implemented continuous delivery, streamlined and automated
=====
len of each list of words 15
['learn', 'how', 'capital', 'one', 'continuous', 'release', 'process', 'common', 'r
.....
['the', 'o'reilly', 'security', 'podcast', 'coarse-grained', 'embracing', 'ephemera
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
```

Blog number: 39

---

```
['The O'Reilly Security Podcast: Coarse-grained security, embracing the ephemeral,
=====
len of each list of words 46
['the', 'o'reilly', 'security', 'podcast', 'coarse-grained', 'embracing', 'ephemera
.....
['twitter', 'financially', 'vr', 'skeptically', 'electronica', 'originally', 'stack
```

[[ 'the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|  
Blog number: 40

---

['Twitter Financially, VR Skeptically, Electronica Originally, and Stack Overflow C  
=====

len of each list of words 60  
['twitter', 'financially', 'vr', 'skeptically', 'electronica', 'originally', 'stack  
.....

['this', 'report', 'explores', 'how', 'political', 'helps', 'everything', 'from', '  
[[ 'the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|

Blog number: 41

---

['This report explores how political data science helps to drive everything from ov  
=====

len of each list of words 60  
['this', 'report', 'explores', 'how', 'political', 'helps', 'everything', 'from', '  
.....

['this', 'chapter', 'from', 'lean', 'ux', '2nd', 'edition', 'how', 'blending', 'lea  
[[ 'the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|

Blog number: 42

---

['This chapter from Lean UX, 2nd Edition discusses how blending Lean UX and Agile c  
=====

len of each list of words 60  
['this', 'chapter', 'from', 'lean', 'ux', '2nd', 'edition', 'how', 'blending', 'lea  
.....

['crafted', 'by', 'experienced', 'practitioners', 'patterns', 'can', 'flexible', 'r  
[[ 'the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|

Blog number: 43

---

['Crafted by experienced object-oriented practitioners, design patterns can make yo  
=====

len of each list of words 60  
['crafted', 'by', 'experienced', 'practitioners', 'patterns', 'can', 'flexible', 'r  
.....

['best', 'practices', 'for', 'web', 'web', 'web', 'not', 'paying', 'enough', 'websi  
[[ 'the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|

Blog number: 44

---

['Best practices for optimizing mobile web apps.Many web designers and web develop  
=====

len of each list of words 60  
['best', 'practices', 'for', 'web', 'web', 'web', 'not', 'paying', 'enough', 'websi

.....  
['paul', 'english's', 'vision', 'pairing', 'ai', 'human', 'expertise', 'has', 'us',  
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|=====|=====|=====|  
Blog number: 45

---

['Paul English's vision of pairing AI and human expertise has much to teach us about  
=====

len of each list of words 21  
['paul', 'english's', 'vision', 'pairing', 'ai', 'human', 'expertise', 'has', 'us',  
.....  
['matt', 'jorgensen', 'look', 'josephine', 'platform', 'connects', 'consumers', 'wi  
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|=====|=====|=====|  
Blog number: 46

---

['Matt Jorgensen offers a look at Josephine, a platform that connects consumers with  
=====

len of each list of words 15  
['matt', 'jorgensen', 'look', 'josephine', 'platform', 'connects', 'consumers', 'wi  
.....  
['rana', 'foroohar', 'makers', 'takers', 'repercussions', 'reading', 'makers', 'a',  
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|=====|=====|=====|  
Blog number: 47

---

['Rana Foroohar, author of Makers and Takers, discusses the repercussions of short-  
=====

len of each list of words 12  
['rana', 'foroohar', 'makers', 'takers', 'repercussions', 'reading', 'makers', 'a',  
.....  
['the', 'behind', 'john', 'basset', 'iii', 'family', 'furniture', 'company', 'prove  
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|=====|=====|=====|  
Blog number: 48

---

['The story behind John Basset III's family furniture company proves the destructio  
=====

len of each list of words 21  
['the', 'behind', 'john', 'basset', 'iii', 'family', 'furniture', 'company', 'prove  
.....  
['at', 'next', 'economy', '2016', 'business', 'leaders', 'policy', 'charted', 'cour  
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|=====|=====|=====|  
Blog number: 49

---

['At Next:Economy 2016, business leaders, policy makers, and technologists charted  
=====

len of each list of words 60  
['at', 'next', 'economy', '2016', 'business', 'leaders', 'policy', 'charted', 'cour  
.....  
['innovation', 'economics', 'decentralized', 'prediction', 'markets', 'front-end',  
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|=====  
Blog number: 50

---

['Innovation Economics, Decentralized Prediction Markets, Front-End Build Tools, an  
=====

len of each list of words 60  
['innovation', 'economics', 'decentralized', 'prediction', 'markets', 'front-end',  
.....  
['why', 'new', 'factor', 'can', 'between', 'failure', 'cloud.in', 'beyond', 'twelve  
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|=====  
Blog number: 51

---

['Why telemetry, a new factor in app development, can mean the difference between s  
=====

len of each list of words 60  
['why', 'new', 'factor', 'can', 'between', 'failure', 'cloud.in', 'beyond', 'twelve  
.....  
['five', 'questions', 'for', 'laura', 'mather', 'insights', 'how', 'groupthink', 'h  
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|=====  
Blog number: 52

---

['Five questions for Laura Mather: Insights on how groupthink and heterogeneous tea  
=====

len of each list of words 60  
['five', 'questions', 'for', 'laura', 'mather', 'insights', 'how', 'groupthink', 'h  
.....  
['douglas', 'rushkoff', 'way', 'forward', 'local', 'circular', 'economy', 'where',  
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|=====  
Blog number: 53

---

['Douglas Rushkoff outlines a way forward to a local, circular economy, where money  
=====

len of each list of words 23  
['douglas', 'rushkoff', 'way', 'forward', 'local', 'circular', 'economy', 'where',  
.....  
['honor', 'using', 'real', 'vexing', 'problem', 'for', 'whose', 'loved', 'need', 'c  
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',  
====|=====|=====|=====|=====|=====|=====|=====  
Blog number: 54

```
[ 'Honor is using modern technology to solve a real, vexing problem for those whose
=====
len of each list of words 21
[ 'honor', 'using', 'real', 'vexing', 'problem', 'for', 'whose', 'loved', 'need', 'c
.....
[ 'james', 'nord', 'fohr', 'card', 'us', 'how', 'reputation', 'gained', 'gets', 'con
[[ 'the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|
Blog number: 55
```

---

```
[ 'James Nord of Fohr Card shows us how the reputation gained through social media g
=====
len of each list of words 13
[ 'james', 'nord', 'fohr', 'card', 'us', 'how', 'reputation', 'gained', 'gets', 'con
.....
[ 'graphs', 'from', 'max', 'roser's', 'our', 'world', 'data', 'have', 'become', 'fix
[[ 'the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|
Blog number: 56
```

---

```
[ 'Graphs and maps from Max Roser's Our World in Data have become a fixture on socia
=====
len of each list of words 24
[ 'graphs', 'from', 'max', 'roser's', 'our', 'world', 'data', 'have', 'become', 'fix
.....
[ 'linkedin', 'ceo', 'jeff', 'weiner', 'what', 'linkedin', 'us', 'how', 'we', 'can',
[[ 'the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|
Blog number: 57
```

---

```
[ 'LinkedIn CEO Jeff Weiner on what LinkedIn tells us about how we can better match
=====
len of each list of words 18
[ 'linkedin', 'ceo', 'jeff', 'weiner', 'what', 'linkedin', 'us', 'how', 'we', 'can',
.....
[ 'keller', 'rinaudo', 'company', 'zipline', 'using', 'with', 'poor', 'continue', 'r
[[ 'the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|
Blog number: 58
```

---

```
[ "Keller Rinaudo's company, Zipline, is using on-demand technology and drones to de
=====
len of each list of words 12
[ 'keller', 'rinaudo', 'company', 'zipline', 'using', 'with', 'poor', 'continue', 'r
.....
[ 'niantic', 'founder', 'ceo', 'john', 'hanke', 'his', 'creating', 'pokémon', 'go',
[[ 'the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|
```

Blog number: 59

---

```
['Niantic founder and CEO John Hanke talks about his motivations in creating the Po
=====
len of each list of words 25
['niantic', 'founder', 'ceo', 'john', 'hanke', 'his', 'creating', 'pokémon', 'go',
.....
['natalie', 'foster', 'aspen', 'institute', 'hosts', 'conversation', 'with', 'andy
[['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
====|=====|=====|=====|=====|=====|=====|=====|=====|=====
```

Blog number: 60

---

```
['Natalie Foster of the Aspen Institute hosts a conversation with Andy Stern, autho
=====
len of each list of words 28
['natalie', 'foster', 'aspen', 'institute', 'hosts', 'conversation', 'with', 'andy
.....
```

```
In [151]: len(master_list) #####checking len of the list with words for each blog, 1
```

```
Out[151]: 60
```

```
In [152]: master_list ##content of the list
```

```
Out[152]: [['the',
            'o'reilly',
            'bots',
            'podcast',
            'how',
            'bots',
            'way',
            'companies',
            'with',
            'customers.customer',
            'key',
            'for',
            'bots--one',
            'first',
            'we',
            'when',
            'we',
            'world',
            'full',
            'ai-enabled',
            'conversational',
            'in',
            'episode',
            'o'reilly',
```



'bots',  
'podcast',  
'pete',  
'skomoroch',  
'i',  
'with',  
'founders',  
'companies',  
'have',  
'bots',  
'help',  
'consumers',  
'companies',  
'each',  
'reading',  
'michael',  
'schneider',  
'rachel',  
'law',  
'alyx',  
'baldwin',  
'customer',  
'with',  
'bots'],  
['the',  
"o'reilly",  
'radar',  
'podcast',  
'prediction',  
'cognitive',  
'biases',  
'how',  
'brains',  
'come',  
'week',  
'episode',  
'i',  
'chat',  
'with',  
'sam',  
'wang',  
'professor',  
'neuroscience',  
'biology',  
'princeton',  
'wang',  
'co-founder',  
'princeton',

'election',  
'consortium',  
'focused',  
'predicting',  
'u.s.',  
'national',  
'elections',  
'we',  
'prediction',  
'crazy',  
'election',  
'cycle',  
'role',  
'neuroscience',  
'have',  
'played',  
'we',  
'current',  
'research',  
'wang',  
'his',  
'working',  
'u.s',  
'brain',  
'initiative',  
'powerful',  
'role',  
'governments',  
'play',  
'research.continue',  
'reading',  
'sam',  
'wang',  
'predicting',  
'election',  
'finding'],  
['the',  
'o'reilly',  
'data',  
'show',  
'podcast',  
'christopher',  
'nguyen',  
'early',  
'apache',  
'spark',  
'learning',  
'for',

'china',  
'ai.in',  
'episode',  
'o'reilly',  
'data',  
'show',  
'i',  
'with',  
'christopher',  
'nguyen',  
'ceo',  
'co-founder',  
'arimo',  
'nguyen',  
'arimo',  
'were',  
'first',  
'proponents',  
'apache',  
'spark',  
'alluxio',  
'most',  
'recently',  
'arimo's',  
'products',  
'has',  
'relied',  
'learning',  
'range',  
'business',  
'problems.continue',  
'reading',  
'why',  
'businesses',  
'pay',  
'learning'],  
['kid',  
'privacy',  
'chinese',  
'creation',  
'reality',  
'rocks',  
'chaos',  
'monkey',  
'2',  
'digital',  
'defenders',  
'free',

'cc-licensed',  
'kids',  
'booklet',  
'privacy',  
'via',  
'cory',  
'doctorow',  
'news',  
'ideas',  
'are',  
'n't',  
'intrinsically',  
'valuable',  
'executed',  
'can',  
'be',  
'valuable',  
'chinese',  
'hardware',  
'can',  
'execute',  
'faster',  
'american',  
'this',  
'up',  
'kickstarter',  
'products',  
'up',  
'for',  
'shenzhen',  
'before',  
'kickstarter',  
'closes',  
'via',  
'boingboing',  
'the',  
'world',  
'getting',  
'better',  
'nobody',  
'knows',  
'it',  
'in',  
'britain',  
'10',  
'%',  
'people',  
'world'],

['if',  
 'we',  
 'let',  
 'put',  
 'us',  
 'work',  
 'will',  
 'be',  
 'because',  
 'failure',  
 'lack',  
 'will',  
 'better',  
 'future',  
 'full',  
 'video',  
 'from',  
 'tim',  
 "o'reilly",  
 'white',  
 'house',  
 'frontiers',  
 'conference',  
 'i',  
 'had',  
 'honor',  
 'be',  
 'warmup',  
 'for',  
 'president',  
 'obama',  
 'white',  
 'house',  
 'frontiers',  
 'conference',  
 'carnegie',  
 'mellon',  
 'university',  
 'pittsburgh',  
 'here',  
 'prepared',  
 'from',  
 'i',  
 'as',  
 'watch',  
 'video',  
 'what',  
 'i',

'ended',  
'up',  
'exactly',  
'what',  
'i',  
'had',  
'written',  
'but',  
'reasonably',  
'close',  
'hearing',  
'bob'],  
['tim', 'frick', 'products', 'reading', 'how', 'for'],  
['wan-replicated',  
'data',  
'nosql',  
'modeling',  
'sharing',  
'roadmaps',  
'shopping',  
'statistics',  
'bedrockdb',  
'wan-replicated',  
'foundation',  
'for',  
'global-scale',  
'nosql',  
'data',  
'modeling',  
'techniques',  
'for',  
'first',  
'2009',  
'brave',  
'new',  
'data',  
'world',  
'roadmaps',  
'views',  
'product',  
'road',  
'for',  
'good',  
'17',  
'staggering',  
'statistics',  
'about',  
'our',

```

'shopping',
'habits',
'5',
'50',
'%',
'will',
'just',
'hit',
'free',
'6',
'an',
'estimated',
'u.s.',
'gross',
'product',
'gdp',
'comes',
'from',
'retail',
'consumption',
'continue',
'reading',
'four',
'links',
'19'],
['using',
'rum',
'http',
'vs.',
'https',
'world',
'regions',
'wireline',
'rural',
'wireline',
'wireless',
'isps.continue',
'reading',
'understanding',
'performance',
'penalty',
'http',
'vs.',
'https'],
['learn',
'why',
'svg',
'from',

```

'css',  
'work',  
'common',  
'hiccups',  
'experience',  
'way.are',  
'just',  
'getting',  
'svg',  
'maybe',  
'have',  
'been',  
'using',  
'svgs',  
'background',  
'using',  
'svgs',  
'just',  
'been',  
'this',  
'excerpt',  
'from',  
'upcoming',  
'book',  
'svg',  
'animations',  
'for',  
'we',  
'carefully',  
'go',  
'what',  
'svg',  
'why',  
'from',  
'css',  
'work',  
'common',  
'hiccups',  
'experience',  
'way',  
'if',  
'learning',  
'please',  
'join',  
'us',  
'for',  
'building',  
'svg',



```

'animations',
'new',
"o'reilly",
'training',
'center',
'boston',
'nov.',
'1-2'],
['understand',
'how',
'capital',
'one',
'embraced',
'continuous',
'learn',
'core',
'principles',
'best',
'practices',
'behind',
'reading',
'part',
'pipeline',
'why',
'continuous',
'essential'],
['in',
'video',
'brian',
'l.',
'gorman',
'uncovers',
'relationship',
'between',
'efficiency',
'branch',
'prediction',
'when',
'executing',
'java.continue',
'reading',
'does',
'run',
'faster',
'unsorted',
'java'],
['a',
'common',

```

'java',  
'pass',  
'by',  
'reference',  
'language',  
'in',  
'video',  
'brian',  
'1.',  
'gorman',  
'exposes',  
'workings',  
'pass',  
'by',  
'value.continue',  
'reading',  
'is',  
'java',  
'pass',  
'by',  
'reference',  
'pass',  
'by',  
'value'],  
['roberto',  
'di',  
'cosmo',  
'reveals',  
'behind',  
'launch',  
'software',  
'heritage',  
'project',  
'has',  
'3',  
'billion',  
'unique',  
'code',  
'files',  
'650',  
'unique',  
'commits.continue',  
'reading',  
'why',  
'how',  
'software',  
'heritage',  
'building',

'universal'],  
 ['lauri',  
 'apple',  
 'zalando',  
 'how',  
 'company.continue',  
 'reading',  
 'building',  
 'culture',  
 'europe's',  
 'largest',  
 'fashion',  
 'platform'],  
 ['open',  
 'licenses',  
 'have',  
 'us',  
 'well',  
 'for',  
 'but',  
 'cory',  
 'doctorow',  
 'need',  
 'help',  
 'we're',  
 'going',  
 'era',  
 'which',  
 'computers',  
 'bodies',  
 'continue',  
 'reading',  
 'how',  
 'got',  
 'here'],  
 ['karen',  
 'sandler',  
 'freedom',  
 'underlies',  
 'effectively',  
 'problems',  
 'we',  
 'need',  
 'new',  
 'way',  
 'relation',  
 'justice',  
 'causes.continue',

'reading',  
'is',  
'freedom',  
'justice'],  
['awesome',  
'falsehoods',  
'neural',  
'network',  
'vision',  
'toy',  
'family',  
'robots',  
'over-engineered',  
'software',  
'awesome',  
'falsehoods',  
'great',  
'collections',  
'lies',  
'we',  
'programmers',  
'what',  
'email',  
'what',  
'networks',  
'favourite',  
'zones',  
'how',  
'do',  
'neural',  
'networks',  
'see',  
'the',  
'world',  
'youtube',  
'video',  
'explainer',  
'anki',  
'cozmo',  
'robot',  
'compare',  
'nxrobo',  
'big-i',  
'first',  
'personalized',  
'family',  
'robot',  
'big-i',

'natural-interaction',  
'robot',  
'with',  
'3d',  
'vision',  
'voice',  
'programming',  
'perception',  
'it',  
'can',  
'needs',  
'ten',  
'modern',  
'software',  
'over-engineering',  
'mistakes'],  
['lessons',  
'learned',  
'fastly',  
'how',  
'build',  
'robust',  
'contains',  
'properly',  
'communicates',  
'prevent',  
'recurrence.in',  
'complex',  
'failure',  
'rule',  
'and',  
'failure',  
'can',  
'have',  
'wide',  
'ranging',  
'especially',  
'when',  
'responsible',  
'for',  
'1',  
'per',  
'internet',  
'as',  
'content',  
'delivery',  
'network',  
'cdn',

'fastly',  
'large',  
'global',  
'environment',  
'which',  
'faces',  
'both',  
'from',  
'reliability',  
'perspective',  
'as',  
'we',  
'put',  
'place',  
'robust',  
'would',  
'us',  
'rapidly',  
'contain',  
'would',  
'ensure',  
'communication',  
'would',  
'flow',  
'effectively',  
'both',  
'within',  
'company'],  
['the',  
'beyond',  
'twelve-factor',  
'app',  
'i',  
'present',  
'new',  
'guidelines',  
'builds',  
'heroku's',  
'12',  
'factors',  
'reflects',  
'best',  
'practices',  
'for',  
'building',  
'cloud-native',  
'i',  
'have',

'changed',  
'priority',  
'factors',  
'concept',  
"api",  
'first"',  
'be',  
'considerations',  
'for',  
'will',  
'be',  
'running',  
'cloud',  
'these',  
'new',  
'15-factor',  
'guidelines',  
'one',  
'codebase',  
'api',  
'first',  
'dependency',  
'design',  
'build',  
'release',  
'run',  
'configuration',  
'credentials',  
'code',  
'logs',  
'disposability',  
'backing',  
'environment',  
'parity',  
'administrative',  
'processes',  
'port',  
'binding',  
'stateless',  
'processes'],  
['5',  
'questions',  
'for',  
'binu',  
'ramakrishnan',  
'insights',  
'centralized',  
'ci/cd',

'environments',  
'common',  
'protection',  
'for',  
'recently',  
'with',  
'binu',  
'ramakrishnan',  
'principal',  
'engineer',  
'yahoo',  
'with',  
'centralized',  
'continuous',  
'continuous',  
'ci/cd',  
'platform',  
'protect',  
'here',  
'highlights',  
'from',  
'1',  
'how',  
'can',  
'ci/cd',  
'platform',  
'posture',  
'lately',  
'has',  
'from',  
'environments',  
'favor',  
'ci/cd',  
'environments',  
'a',  
'factor',  
'behind',  
'platform',  
'economies',  
'in',  
'build',  
'environments',  
'each',  
'build',  
'the',  
'risk',  
'exposure',  
'from',



```

    'limited',
    'because',
    'each',
    'likely'],
['kris',
 'borchers',
 'javascript.continue',
 'reading',
 'evolving',
 'javascript',
 'ecosystem'],
['mark',
 'shuttleworth',
 'next',
 'likely',
 'come',
 'from',
 'laptop',
 'pc',
 'beating',
 'neighbor',
 'web',
 'cloud.continue',
 'reading',
 'brilliant',
 'pebbles'],
['danese',
 'cooper',
 'explores',
 'innersource',
 'within',
 'walls',
 'reading',
 'how',
 'innersource',
 'like',
 'flossing'],
['watch',
 'highlights',
 'covering',
 'business',
 'from',
 'oscon',
 'london',
 '2016.people',
 'from',
 'community',
 'came',

```

'for',  
'oscon',  
'london',  
'2016',  
'below',  
'find',  
'links',  
'highlights',  
'from',  
'event',  
'brilliant',  
'pebbles',  
'mark',  
'shuttleworth',  
'next',  
'likely',  
'come',  
'from',  
'laptop',  
'pc',  
'beating',  
'neighbour',  
'web',  
'cloud',  
'watch',  
'`',  
'brilliant',  
'pebbles',  
'how',  
'got',  
'here',  
'open',  
'licenses',  
'have',  
'us',  
'well',  
'for',  
'but',  
'cory',  
'doctorow',  
'need',  
'help',  
'we're',  
'going',  
'era',  
'which',  
'computers',  
'bodies',

'watch'],  
['male',  
'allies',  
'learning',  
'maps',  
'machine',  
'learning',  
'biology',  
'`',  
'the',  
'attention',  
'merchants',  
'male',  
'allies',  
'bingo',  
'card',  
'via',  
'anna',  
'christison',  
'so',  
'you',  
'want',  
'learn',  
'physics',  
'really',  
'learning',  
'at',  
"o'reilly",  
'we',  
'been',  
'hard',  
'learning',  
'paths',  
'for',  
'untargeted',  
'metabolomics',  
'using',  
'deep',  
'learning',  
'this',  
'untargeted',  
'unbiased',  
'every',  
'change',  
'cell',  
'gives',  
'us',  
'for',

'biology',  
'learning',  
'has',  
'been',  
'key',  
'getting',  
'here',  
'reliably',  
'the',  
'attention',  
'merchants',  
'the',  
'epic'],  
[ '5',  
'questions',  
'for',  
'timothy',  
'gallo',  
'allan',  
'liska',  
'insights',  
'common',  
'ransomware',  
'how',  
'find',  
'balance',  
'between',  
'risks',  
'productivity.i',  
'recently',  
'with',  
'ransomware',  
'co-authors',  
'allan',  
'liska',  
'consulting',  
'engineer',  
'recorded',  
'future',  
'timothy',  
'gallo',  
'cyber',  
'engineer',  
'symantec',  
'challenges',  
'ransomware',  
'how',  
'here',

'highlights',  
'from',  
'1',  
'can',  
'explain',  
'common',  
'for',  
'ransomware',  
'how',  
'have',  
'persisted',  
'the',  
'common',  
'email',  
'via',  
'web',  
'using',  
'exploits',  
'kits',  
'there',  
'have',  
'been',  
'ransomware',  
'like',  
'keranger'],  
['amazing',  
'china',  
'deep',  
'learning',  
'memory',  
'vr',  
'spaces',  
'javascript',  
'song',  
'what',  
'\$',  
'50',  
'buys',  
'you',  
'at',  
'huaqianbei',  
'markets',  
'shenzhen',  
'was',  
'blown',  
'\$',  
'6',  
'for',

'gsm',  
'chipset',  
'cpu',  
'lcd',  
'battery',  
'pcb',  
'housing',  
'watch',  
'band',  
'microusb',  
'cable',  
'box',  
'and',  
'labor',  
'differentiable',  
'neural',  
'computers',  
'deepmind',  
'latest',  
'work',  
'putting',  
'learning',  
'lets',  
'computers',  
'learn',  
'form',  
'navigate',  
'complex',  
'modalvr',  
'nolan',  
'bushnell',  
'new',  
'which',  
'hopes',  
'for',  
'vr',  
'what'],  
['this',  
'free',  
'webcast',  
'provides',  
'quick',  
'lesson',  
'updating',  
'relational',  
'oct.',  
'26',  
'2pm',

'pt.virtually',  
'every',  
'requires',  
'persistent',  
'one',  
'best',  
'features',  
'spring',  
'framework',  
'how',  
'easily',  
'provides',  
'work',  
'with',  
'relational',  
'this',  
'free',  
'webcast',  
'being',  
'held',  
'oct.',  
'26',  
'2pm',  
'pt',  
'will',  
'how',  
'use',  
'spring',  
'jdbctemplate',  
'updating',  
'relational',  
'very',  
'well',  
'handling',  
'exceptions',  
'using',  
'discussions',  
'both',  
'spring',  
'jpa',  
'working',  
'with',  
'nosql',  
'will',  
'be',  
'no',  
'previous',  
'spring',

'experience'],  
['author',  
'mark',  
'richards',  
'walks',  
'10',  
'common',  
'pitfalls',  
'provides',  
'for',  
'preface',  
'in',  
'late',  
'2006',  
'soa',  
'was',  
'craze',  
'companies',  
'were',  
'jumping',  
'bandwagon',  
'embracing',  
'soa',  
'before',  
'fully',  
'understanding',  
'very',  
'complex',  
'those',  
'companies',  
'embarked',  
'soa',  
'projects',  
'found',  
'constant',  
'with',  
'granularity',  
'performance',  
'particular',  
'change',  
'comes',  
'with',  
'soa',  
'as',  
'result',  
'companies',  
'either',  
'soa',



'efforts',  
'built',  
'hybrid',  
'not',  
'fulfill',  
'promises',  
'soa',  
'today',  
'we',  
'poised',  
'repeat',  
'experience',  
'with'],  
['a',  
'look',  
'expanding',  
'biotech',  
'landscape',  
'sxsw',  
'south',  
'by',  
'southwest',  
'sxsw',  
'film',  
'festival',  
'happens',  
'austin',  
'texas',  
'every',  
'march',  
'since',  
'1987',  
'has',  
'evolved',  
'become',  
'launching',  
'pad',  
'for',  
'wide',  
'range',  
'figure',  
'1-1',  
'figure',  
'1-1',  
'the',  
'sxsw',  
'festival',  
'phylogenetic',

```

'reflects',
'cultural',
'the',
'rumblings',
'biotech',
'wave',
'closer',
'look',
'evolving',
'programming',
'ecosystem',
'biotech',
'labs',
'labs',
'like',
'craig',
'venter',
'23andme',
'ubiome',
'genspace',
'la',
'pailasse',
'ginkgo',
'bioworks',
'biobuilder'],
['the',
'o'reilly',
'bots',
'podcast',
'bots',
'can',
'respond',
'groups',
'users.the',
'workplace',
'rich',
'venue',
'for',
'bots',
'help',
'with',
'productivity',
'collaboration',
'in',
'episode',
'o'reilly',
'bots',
'podcast',

```

'pete',  
'skomoroch',  
'i',  
'focus',  
'workplace',  
'bots',  
'we',  
'begin',  
'by',  
'with',  
'jassim',  
'latif',  
'head',  
'partnerships',  
'slack',  
'bots',  
'written',  
'by',  
'for',  
'human',  
'resources',  
'"we're',  
'building',  
'future',  
'where',  
'value',  
'people',  
'building',  
'slack',  
'far',  
'value',  
'slack',  
'"',  
'latif',  
'reading',  
'jassim',  
'latif'],  
['the',  
'o'reilly',  
'design',  
'podcast',  
'design',  
'creating',  
'right',  
'environment.in',  
'week's',  
'design',  
'podcast',

```
'i',
'with',
'kristin',
'skinner',
'adaptive',
'path',
'head',
'capital',
'one',
'co-author',
'org',
'design',
'for',
'design',
'orgs',
'we',
'what',
'we',
'can',
'learn',
'from',
'golden',
'state',
'warriors.continue',
'reading',
'kristin',
'skinner'],
['the',
'o'reilly',
'data',
'show',
'podcast',
'natalino',
'busa',
'feature',
'engineering',
'predictive',
'episode',
'o'reilly',
'data',
'show',
'o'reilly's',
'editor',
'jenn',
'webb',
'with',
'natalino',
'busa',
```

'predictive',  
'challenges',  
'feature',  
'engineering',  
'new',  
'class',  
'enabling',  
'features',  
'emerge',  
'from',  
'patterns',  
'within',  
'they',  
'relationship',  
'between',  
'predictive',  
'high-quality',  
'how',  
'learning',  
'being',  
'used',  
'financial',  
'reading',  
'understanding',  
'predictive'],  
['take',  
'learning',  
'followed',  
'previous',  
'post',  
'learned',  
'how',  
'gpu-accelerated',  
'tensorflow',  
'create',  
'classifier',  
'linux',  
'computer',  
'honestly',  
'process',  
'classifying',  
'consuming',  
'laptop',  
'have',  
'want',  
'classify',  
'enter',  
'lot',

'code',  
'just',  
'classify',  
'thankfully',  
'not',  
'well',  
'publicized',  
'can',  
'run',  
'inception',  
'classifier--or',  
'classifier--live',  
'camera-equipped',  
'phone',  
'you',  
'need',  
'point',  
'camera',  
'what',  
'classify',  
'tensorflow',  
'will',  
'what',  
'you',  
'can',  
'use',  
'tensorflow',  
'raspberry',  
'pi',  
'but',  
'for',  
'i',  
'will',  
'be',  
'using',  
'android',  
'i'll'],  
['cake',  
'cutting',  
'tea',  
'making',  
'google',  
'interviewing',  
'automation',  
'puzzling',  
'fair',  
'cake-cutting',  
'but',

'april',  
'computer',  
'expectations',  
'by',  
'posting',  
'paper',  
'envy-free',  
'cake-cutting',  
'whose',  
'running',  
'number',  
'players',  
'not',  
'preferences',  
'[',  
 ',  
'the',  
'extraordinarily',  
'complex',  
'cake',  
'n',  
'players',  
'can',  
'require',  
'n^n^n^n^n^n',  
'roughly',  
'equivalent',  
'number',  
'cuts',  
'even',  
'for',  
'just',  
'handful',  
'players',  
'number',  
'greater',  
'number',  
'universe',  
'you',  
'ca',  
'n't',  
'have',  
'cake',  
'compute',  
'english',  
'man',  
'spends',  
'11',

```
'hours'],  
['word',  
 'embedding',  
 'natural',  
 'language',  
 'processing.word',  
 'embedding',  
 'words',  
 'vectors',  
 'whose',  
 'relative',  
 'correlate',  
 'with',  
 'this',  
 'unsupervised',  
 'learning',  
 'natural',  
 'language',  
 'processing',  
 'nlp',  
 'encode',  
 'words',  
 'which',  
 'provide',  
 'no',  
 'useful',  
 'regarding',  
 'relationships',  
 'exist',  
 'between',  
 'words',  
 'word',  
 'embedding',  
 'nlp',  
 'whereby',  
 'words',  
 'phrases',  
 'from',  
 'vocabulary',  
 'vectors',  
 'real',  
 'numbers',  
 'low-dimensional',  
 'relative',  
 'vocabulary',  
 'between',  
 'vectors',  
 'correlate',
```



'with',  
'words',  
'for',  
'example',  
'let's',  
'words',  
'woman',  
'queen',  
'king',  
'we',  
'can',  
'get',  
'vector'],  
['insider',  
"o'reilly",  
'security',  
'conference',  
'proposal',  
'process',  
'rejection',  
'o'reilly',  
'security',  
'conference',  
'we've',  
'had',  
'great',  
'from',  
'folks',  
'especially',  
'who',  
'looking',  
'forward',  
'what',  
'works',  
'with',  
'like-minded',  
'professionals',  
'we've',  
'received',  
'queries',  
'process',  
'for',  
'handling',  
'call',  
'for',  
'proposals',  
'cfp',  
'in',

'we'll',  
'be',  
'quite',  
'bit',  
'event',  
'we',  
'want',  
'present',  
'feedback',  
'from',  
'cfp',  
'process',  
'bit',  
'version',  
'"behind',  
'music."',  
'tune',  
'next',  
'week',  
'when',  
'we'll',  
'follow',  
'up',  
'with',  
'post'],  
['learn',  
'how',  
'capital',  
'one',  
'continuous',  
'release',  
'process',  
'common',  
'roadblocks.continue',  
'reading',  
'making',  
'continuous',  
'reality',  
'large',  
'enterprise'],  
['the',  
'o'reilly',  
'security',  
'podcast',  
'coarse-grained',  
'embracing',  
'ephemeral',  
'empathy',

'for',  
'everyone.in',  
'episode',  
'i',  
'with',  
'dan',  
'kaminsky',  
'founder',  
'chief',  
'white',  
'ops',  
'we',  
'what',  
'national',  
'institutes',  
'health',  
'nih',  
'for',  
'would',  
'look',  
'like',  
'pros',  
'cons',  
'docker',  
'ephemeral',  
'how',  
'listening',  
'people',  
'better',  
'can',  
'for',  
'everyone.continue',  
'reading',  
'dan',  
'kaminsky',  
'creating',  
'nih',  
'for'],  
['twitter',  
'financially',  
'vr',  
'skeptically',  
'electronica',  
'originally',  
'stack',  
'overflow',  
'competitively',  
'twitter',

'buyout',  
'financial',  
'analysis',  
'somewhere',  
'near',  
'half',  
'billion',  
'costs',  
'need',  
'be',  
'and',  
'firing',  
'people',  
'being',  
'general',  
'bastard',  
'it',  
'jack',  
'dorsey',  
'burning',  
'half',  
'billion',  
'per',  
'n't',  
'business',  
'via',  
'marginal',  
'revolution',  
'carmack',  
'keynote',  
'oculus',  
'3',  
'`',  
'we',  
'[',  
'vr',  
'community',  
']',  
'coasting',  
'novelty',  
'wonder',  
'being',  
'people',  
'have',  
'never',  
'before',  
'he',  
'continued',

``',  
'but'],  
['this',  
'report',  
'explores',  
'how',  
'political',  
'helps',  
'everything',  
'from',  
'voter',  
'contacts',  
'prologue',  
'the',  
'role',  
'data',  
'campaigns',  
'andrew',  
'therriault',  
'was',  
'democratic',  
'national',  
'committee',  
'director',  
'data',  
'science',  
'from',  
'2014',  
'2016',  
'leading',  
'voter',  
'used',  
'by',  
'democratic',  
'campaigns',  
'watching',  
'news',  
'coverage',  
'2016',  
'presidential',  
'race',  
'easy',  
'election',  
'unlike',  
'come',  
'before',  
'behind',  
'campaigns',

'continuing',  
'has',  
'been',  
'underway',  
'for',  
'modern',  
'campaigns',  
'us',  
'have',  
'become',  
'reliant',  
'with',  
'each',  
'election'],  
['this',  
'chapter',  
'from',  
'lean',  
'ux',  
'2nd',  
'edition',  
'how',  
'blending',  
'lean',  
'ux',  
'agile',  
'can',  
'create',  
'productive',  
'collaborative',  
'process',  
'agile',  
'now',  
'at',  
'huge',  
'products',  
'like',  
'amazon's',  
'kindle',  
'apple's',  
'user',  
'experience',  
'ux',  
'but',  
'agile',  
'work',  
'with',  
'ux',

'continues',  
'be',  
'challenge',  
'for',  
'companies',  
'in',  
'chapter',  
'we',  
'review',  
'how',  
'lean',  
'ux',  
'can',  
'fit',  
'within',  
'popular',  
'flavor',  
'agile--scrum--and',  
'how',  
'blending',  
'lean',  
'ux',  
'agile',  
'can',  
'create',  
'productive'],  
['crafted',  
'by',  
'experienced',  
'practitioners',  
'patterns',  
'can',  
'flexible',  
'resilient',  
'change',  
'easier',  
'encountered',  
'patterns',  
'passing',  
'nutshell',  
'general',  
'can',  
'use',  
'crafted',  
'by',  
'experienced',  
'practitioners',  
'patterns',

'can',  
'flexible',  
'resilient',  
'change',  
'easier',  
'so',  
'patterns',  
'part',  
'here',  
'five',  
'reasons',  
'1',  
'make',  
'life',  
'easier',  
'by',  
'not',  
'reinventing',  
'wheel',  
'there's',  
'constant',  
'no',  
'projects',  
'will',  
'late',  
'budget',  
'it's',  
'change',  
'change',  
'when',  
'requirements',  
'change',  
'grow',  
'new',  
'features',  
'when',  
'performance',  
'we're'],  
['best',  
'practices',  
'for',  
'web',  
'web',  
'web',  
'not',  
'paying',  
'enough',  
'websites''],



'performance',  
'and',  
'we',  
'users',  
'paying',  
'for',  
'it's',  
'for',  
'website',  
'creators',  
'understand',  
'what',  
'harms',  
'performance',  
'what',  
'performance',  
'better',  
'performance',  
'leads',  
'conversions',  
'end--that',  
'for',  
'here',  
'checklist',  
'performance',  
'for',  
'every',  
'web',  
'project',  
'working',  
'basic',  
'gathering',  
'from',  
'will',  
'help',  
'later',  
'performance',  
'chapters',  
'four',  
'high',  
'performance',  
'mobile',  
'web',  
'provide',  
'how',  
'get',  
'for',  
'viewport',

```

'&',
'gt'],
['paul',
'english's',
'vision',
'pairing',
'ai',
'human',
'expertise',
'has',
'us',
'how',
'business',
'future.continue',
'reading',
'making',
'humans',
'cool',
'a',
'conversation',
'with',
'paul',
'english'],
['matt',
'jorgensen',
'look',
'josephine',
'platform',
'connects',
'consumers',
'with',
'home',
'cooks.continue',
'reading',
'local',
'food',
'local',
'chefs'],
['rana',
'foroohar',
'makers',
'takers',
'repercussions',
'reading',
'makers',
'a',
'conversation',
'with',

```

'rana',  
 'foroohar'],  
 ['the',  
 'behind',  
 'john',  
 'basset',  
 'iii',  
 'family',  
 'furniture',  
 'company',  
 'proves',  
 'good',  
 'jobs',  
 'not',  
 'continue',  
 'reading',  
 'factory',  
 'a',  
 'conversation',  
 'with',  
 'john',  
 'bassett',  
 'iii'],  
 ['at',  
 'next',  
 'economy',  
 '2016',  
 'business',  
 'leaders',  
 'policy',  
 'charted',  
 'course',  
 'from',  
 'economy',  
 'we',  
 'experience',  
 'economy',  
 'brings',  
 'prosperity',  
 'leaders',  
 'policy',  
 'came',  
 'san',  
 'francisco',  
 'for',  
 "o'reilly",  
 'next',  
 'economy',

'summit',  
'2016',  
'below',  
'find',  
'links',  
'highlights',  
'from',  
'event',  
'on-demand',  
'for',  
'blood',  
'keller',  
'rinaudo',  
'company',  
'zipline',  
'using',  
'with',  
'poor',  
'watch',  
'`',  
'on-demand',  
'for',  
'blood',  
'a',  
'conversation',  
'with',  
'jeff',  
'weiner',  
'linkedin',  
'ceo',  
'jeff',  
'weiner',  
'what',  
'linkedin',  
'can'],  
['innovation',  
'economics',  
'decentralized',  
'prediction',  
'markets',  
'front-end',  
'build',  
'tools',  
'cs',  
'debates',  
'innovation',  
'through',  
'holmstrom',

'lens',  
['',  
'nobel',  
'prize-winner',  
'],'',  
'holmstrom's',  
'work',  
'has',  
'provided',  
'with',  
'great',  
'understanding',  
'why',  
'looks',  
'way',  
'via',  
'marginal',  
'revolution',  
'augur',  
'prediction',  
'making',  
'sense',  
'front-end',  
'build',  
'tools',  
'build',  
'install',  
';',  
'the',  
'first',  
'question',  
'when',  
'confronting',  
'new',  
'build',  
'"is',  
'for',  
'for',  
'"',  
'vigorous',  
'public',  
'debates',  
'computer',  
'science',  
'the',  
'purpose',  
'post'],  
['why',

'new',  
'factor',  
'can',  
'between',  
'failure',  
'cloud.in',  
'beyond',  
'twelve-factor',  
'app',  
'i',  
'present',  
'new',  
'guidelines',  
'builds',  
'heroku's',  
'12',  
'factors',  
'reflects',  
'best',  
'practices',  
'for',  
'building',  
'cloud-native',  
'i',  
'have',  
'changed',  
'priority',  
'factors',  
'concept',  
'api',  
'first"',  
'be',  
'considerations',  
'for',  
'will',  
'be',  
'running',  
'cloud',  
'these',  
'new',  
'15-factor',  
'guidelines',  
'one',  
'codebase',  
'api',  
'first',  
'dependency',  
'design',

'build',  
'release',  
'run',  
'configuration',  
'credentials',  
'code',  
'logs',  
'disposability',  
'backing',  
'environment',  
'parity'],  
['five',  
'questions',  
'for',  
'laura',  
'mather',  
'insights',  
'how',  
'groupthink',  
'heterogeneous',  
'recently',  
'with',  
'laura',  
'mather',  
'founder',  
'ceo',  
'unitive',  
'groupthink',  
'how',  
'hampers',  
'quickly',  
'better',  
'problems',  
'here',  
'highlights',  
'from',  
'1',  
'what',  
'"groupthink',  
'groupthink',  
'great',  
'example',  
'how',  
'best',  
'can',  
'really',  
'go',  
'essentially',

'when',  
'group',  
'focused',  
'getting',  
'criticize',  
'each',  
'i'm',  
'we've',  
'happen',  
'we're',  
'group',  
'people',  
'who',  
'either',  
'very',  
'loud',  
'very',  
'well',  
'respected',  
'comment',  
'everyone',  
'nodding',  
'sometimes'],  
['douglas',  
'rushkoff',  
'way',  
'forward',  
'local',  
'circular',  
'economy',  
'where',  
'community',  
'helps',  
'us',  
'build',  
'human-centered',  
'reading',  
'throwing',  
'rocks',  
'google',  
'bus',  
'a',  
'conversation',  
'with',  
'douglas',  
'rushkoff'],  
['honor',  
'using',



'real',  
 'vexing',  
 'problem',  
 'for',  
 'whose',  
 'loved',  
 'need',  
 'care.continue',  
 'reading',  
 'aging',  
 'with',  
 'honor',  
 'a',  
 'conversation',  
 'with',  
 'seth',  
 'sternberg',  
 'phaedra',  
 'ellis-lamkins'],  
 ['james',  
 'nord',  
 'fohr',  
 'card',  
 'us',  
 'how',  
 'reputation',  
 'gained',  
 'gets',  
 'converted',  
 'cash.continue',  
 'reading',  
 'making'],  
 ['graphs',  
 'from',  
 'max',  
 'roser's',  
 'our',  
 'world',  
 'data',  
 'have',  
 'become',  
 'fixture',  
 'conversations',  
 'world',  
 'roser',  
 'reminds',  
 'us',  
 'we',

```

'play',
'cards',
'right',
'future',
'bright.continue',
'reading',
'our',
'world'],
['linkedin',
'ceo',
'jeff',
'weiner',
'what',
'linkedin',
'us',
'how',
'we',
'can',
'better',
'key',
'reading',
'a',
'conversation',
'with',
'jeff',
'weiner'],
['keller',
'rinaudo',
'company',
'zipline',
'using',
'with',
'poor',
'continue',
'reading',
'on-demand',
'for',
'blood'],
['niantic',
'founder',
'ceo',
'john',
'hanke',
'his',
'creating',
'pokémon',
'go',
'what',

```

```

        'us',
        'future',
        'entertainment.continue',
        'reading',
        'the',
        'city',
        'a',
        'conversation',
        'with',
        'john',
        'hanke',
        'niantic',
        'creator',
        'pokémon',
        'go'],
    ['natalie',
     'foster',
     'aspen',
     'institute',
     'hosts',
     'conversation',
     'with',
     'andy',
     'stern',
     'raising',
     'floor',
     'elizabeth',
     'rhodes',
     'who',
     'heading',
     'up',
     'y',
     'combinator's',
     'new',
     'universal',
     'basic',
     'experiments',
     'continue',
     'reading',
     'is',
     'for',
     'universal',
     'basic']]

```

```
In [153]: len(master_list[0]) ###checking the lenght of the list at position 0
```

```
Out[153]: 48
```

```
In [154]: for s in master_list: ###printing the content of all lists
```

```
print(s)
```

```
['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with', 'the', 'o'reilly', 'radar', 'podcast', 'prediction', 'cognitive', 'biases', 'how', 'the', 'o'reilly', 'data', 'show', 'podcast', 'christopher', 'nguyen', 'early', 'kid', 'privacy', 'chinese', 'creation', 'reality', 'rocks', 'chaos', 'monkey', '2', 'if', 'we', 'let', 'put', 'us', 'work', 'will', 'be', 'because', 'failure', 'lack', 'tim', 'frick', 'products', 'reading', 'how', 'for']  
['wan-replicated', 'data', 'nosql', 'modeling', 'sharing', 'roadmaps', 'shopping', 'using', 'rum', 'http', 'vs.', 'https', 'world', 'regions', 'wireline', 'rural', 'learn', 'why', 'svg', 'from', 'css', 'work', 'common', 'hiccups', 'experience', 'understand', 'how', 'capital', 'one', 'embraced', 'continuous', 'learn', 'core', 'in', 'video', 'brian', 'l.', 'gorman', 'uncovers', 'relationship', 'between', 'ef', 'a', 'common', 'java', 'pass', 'by', 'reference', 'language', 'in', 'video', 'bria', 'roberto', 'di', 'cosmo', 'reveals', 'behind', 'launch', 'software', 'heritage', 'lauri', 'apple', 'zalando', 'how', 'company.continue', 'reading', 'building', 'cu', 'open', 'licenses', 'have', 'us', 'well', 'for', 'but', 'cory', 'doctorow', 'need', 'karen', 'sandler', 'freedom', 'underlies', 'effectively', 'problems', 'we', 'need', 'awesome', 'falsehoods', 'neural', 'network', 'vision', 'toy', 'family', 'robots', 'lessons', 'learned', 'fastly', 'how', 'build', 'robust', 'contains', 'properly', 'the', 'beyond', 'twelve-factor', 'app', 'i', 'present', 'new', 'guidelines', 'bui', '5', 'questions', 'for', 'binu', 'ramakrishnan', 'insights', 'centralized', 'ci/co', 'kris', 'borchers', 'javascript.continue', 'reading', 'evolving', 'javascript', 'e', 'mark', 'shuttleworth', 'next', 'likely', 'come', 'from', 'laptop', 'pc', 'beating', 'danese', 'cooper', 'explores', 'innersource', 'within', 'walls', 'reading', 'how', 'watch', 'highlights', 'covering', 'business', 'from', 'oscon', 'london', '2016.pe', 'male', 'allies', 'learning', 'maps', 'machine', 'learning', 'biology', '``', 'the', '5', 'questions', 'for', 'timothy', 'gallo', 'allan', 'liska', 'insights', 'common', 'amazing', 'china', 'deep', 'learning', 'memory', 'vr', 'spaces', 'javascript', 's', 'this', 'free', 'webcast', 'provides', 'quick', 'lesson', 'updating', 'relational', 'author', 'mark', 'richards', 'walks', '10', 'common', 'pitfalls', 'provides', 'fo', 'a', 'look', 'expanding', 'biotech', 'landscape', 'sxsw', 'south', 'by', 'southwes', 'the', 'o'reilly', 'bots', 'podcast', 'bots', 'can', 'respond', 'groups', 'users.t', 'the', 'o'reilly', 'design', 'podcast', 'design', 'creating', 'right', 'environmen', 'the', 'o'reilly', 'data', 'show', 'podcast', 'natalino', 'busa', 'feature', 'engi', 'take', 'learning', 'followed', 'previous', 'post', 'learned', 'how', 'gpu-acceler', 'cake', 'cutting', 'tea', 'making', 'google', 'interviewing', 'automation', 'puzzl', 'word', 'embedding', 'natural', 'language', 'processing.word', 'embedding', 'words', 'insider', 'o'reilly', 'security', 'conference', 'proposal', 'process', 'rejection', 'learn', 'how', 'capital', 'one', 'continuous', 'release', 'process', 'common', 'r', 'the', 'o'reilly', 'security', 'podcast', 'coarse-grained', 'embracing', 'ephemera', 'twitter', 'financially', 'vr', 'skeptically', 'electronica', 'originally', 'stack', 'this', 'report', 'explores', 'how', 'political', 'helps', 'everything', 'from', 'this', 'chapter', 'from', 'lean', 'ux', '2nd', 'edition', 'how', 'blending', 'lea', 'crafted', 'by', 'experienced', 'practitioners', 'patterns', 'can', 'flexible', 'r', 'best', 'practices', 'for', 'web', 'web', 'web', 'not', 'paying', 'enough', 'websi', 'paul', 'english's', 'vision', 'pairing', 'ai', 'human', 'expertise', 'has', 'us', 'matt', 'jorgensen', 'look', 'josephine', 'platform', 'connects', 'consumers', 'wi
```

```
['rana', 'foroohar', 'makers', 'takers', 'repercussions', 'reading', 'makers', 'a',
['the', 'behind', 'john', 'basset', 'iii', 'family', 'furniture', 'company', 'prove
['at', 'next', 'economy', '2016', 'business', 'leaders', 'policy', 'charted', 'cour
['innovation', 'economics', 'decentralized', 'prediction', 'markets', 'front-end',
['why', 'new', 'factor', 'can', 'between', 'failure', 'cloud.in', 'beyond', 'twelve
['five', 'questions', 'for', 'laura', 'mather', 'insights', 'how', 'groupthink', 'h
['douglas', 'rushkoff', 'way', 'forward', 'local', 'circular', 'economy', 'where',
['honor', 'using', 'real', 'vexing', 'problem', 'for', 'whose', 'loved', 'need', 'c
['james', 'nord', 'fohr', 'card', 'us', 'how', 'reputation', 'gained', 'gets', 'con
['graphs', 'from', 'max', 'roser's', 'our', 'world', 'data', 'have', 'become', 'fix
['linkedin', 'ceo', 'jeff', 'weiner', 'what', 'linkedin', 'us', 'how', 'we', 'can',
['keller', 'rinaudo', 'company', 'zipline', 'using', 'with', 'poor', 'continue', 'r
['niantic', 'founder', 'ceo', 'john', 'hanke', 'his', 'creating', 'pokémon', 'go',
['natalie', 'foster', 'aspen', 'institute', 'hosts', 'conversation', 'with', 'andy']
```

```
In [155]: #writing the lists to local file "new_filename.txt" as each list is a list
          ##writelines in python3 to iterate thru lists
```

```
with open('new_filename.txt', 'w') as f:
    f.writelines("%s\n" % l for l in master_list)

#####words from each blog are selected,
#####has been scheduled 60 words what represent the content
```

```
In [156]: import numpy
          from nltk.cluster import KMeansClusterer, GAAClusterer, cosine_distance
          import nltk.corpus
          import nltk.stem
          stemmer_func = nltk.stem.snowball.SnowballStemmer("english").stem
          stopwords = set(nltk.corpus.stopwords.words('english'))
```

```
In [157]: def normalize_word(word):
          return stemmer_func(word.lower())
```

```
In [158]: def get_words(titles):
          words = set()
          for title in job_titles:
              for word in title.split():
                  words.add(normalize_word(word))
          return list(words)
```

```
In [159]: def vectorspaced(title):
          title_components = [normalize_word(word) for word in title.split()]
          return numpy.array([
              word in title_components and not word in stop_words
              for word in words], numpy.short)
```

```
In [160]: title_file = open("new_filename.txt", 'r')
```

```
In [161]: job_titles = [line.strip() for line in title_file.readlines()]
          words = get_words(job_titles)
          words[0:10]
```

```
In [162]: [vectorspaced(title) for title in job_titles if title]
```

```

array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 1, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 1, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 1, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 1, ..., 0, 1, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16),
array([0, 0, 0, ..., 0, 0, 0], dtype=int16)

```

```
In [163]: len(job_titles)
```

```
Out[163]: 60
```

```
In [164]: for title in job_titles:
           print(title)
```

```

['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',
['the', "o'reilly", 'radar', 'podcast', 'prediction', 'cognitive', 'biases', 'how',
['the', 'o'reilly', 'data', 'show', 'podcast', 'christopher', 'nguyen', 'early', 'a
['kid', 'privacy', 'chinese', 'creation', 'reality', 'rocks', 'chaos', 'monkey', '2
['if', 'we', 'let', 'put', 'us', 'work', 'will', 'be', 'because', 'failure', 'lack
['tim', 'frick', 'products', 'reading', 'how', 'for']
['wan-replicated', 'data', 'nosql', 'modeling', 'sharing', 'roadmaps', 'shopping',
['using', 'rum', 'http', 'vs.', 'https', 'world', 'regions', 'wireline', 'rural',
['learn', 'why', 'svg', 'from', 'css', 'work', 'common', 'hiccups', 'experience',
['understand', 'how', 'capital', 'one', 'embraced', 'continuous', 'learn', 'core',
['in', 'video', 'brian', 'l.', 'gorman', 'uncovers', 'relationship', 'between', 'ef
['a', 'common', 'java', 'pass', 'by', 'reference', 'language', 'in', 'video', 'bria

```

['roberto', 'di', 'cosmo', 'reveals', 'behind', 'launch', 'software', 'heritage', 'cu  
['lauri', 'apple', 'zalando', 'how', 'company.continue', 'reading', 'building', 'cu  
['open', 'licenses', 'have', 'us', 'well', 'for', 'but', 'cory', 'doctorow', 'need  
['karen', 'sandler', 'freedom', 'underlies', 'effectively', 'problems', 'we', 'need  
['awesome', 'falsehoods', 'neural', 'network', 'vision', 'toy', 'family', 'robots',  
['lessons', 'learned', 'fastly', 'how', 'build', 'robust', 'contains', 'properly',  
['the', 'beyond', 'twelve-factor', 'app', 'i', 'present', 'new', 'guidelines', 'bui  
['5', 'questions', 'for', 'binu', 'ramakrishnan', 'insights', 'centralized', 'ci/co  
['kris', 'borchers', 'javascript.continue', 'reading', 'evolving', 'javascript', 'e  
['mark', 'shuttleworth', 'next', 'likely', 'come', 'from', 'laptop', 'pc', 'beating  
['danese', 'cooper', 'explores', 'innersource', 'within', 'walls', 'reading', 'how  
['watch', 'highlights', 'covering', 'business', 'from', 'oscon', 'london', '2016.pe  
['male', 'allies', 'learning', 'maps', 'machine', 'learning', 'biology', '``', 'the  
['5', 'questions', 'for', 'timothy', 'gallo', 'allan', 'liska', 'insights', 'common  
['amazing', 'china', 'deep', 'learning', 'memory', 'vr', 'spaces', 'javascript', 's  
['this', 'free', 'webcast', 'provides', 'quick', 'lesson', 'updating', 'relational  
['author', 'mark', 'richards', 'walks', '10', 'common', 'pitfalls', 'provides', 'fo  
['a', 'look', 'expanding', 'biotech', 'landscape', 'sxsw', 'south', 'by', 'southwes  
['the', 'o'reilly', 'bots', 'podcast', 'bots', 'can', 'respond', 'groups', 'users.t  
['the', 'o'reilly', 'design', 'podcast', 'design', 'creating', 'right', 'environmen  
['the', 'o'reilly', 'data', 'show', 'podcast', 'natalino', 'busa', 'feature', 'engi  
['take', 'learning', 'followed', 'previous', 'post', 'learned', 'how', 'gpu-acceler  
['cake', 'cutting', 'tea', 'making', 'google', 'interviewing', 'automation', 'puzzl  
['word', 'embedding', 'natural', 'language', 'processing.word', 'embedding', 'words  
['insider', 'o'reilly', 'security', 'conference', 'proposal', 'process', 'rejection  
['learn', 'how', 'capital', 'one', 'continuous', 'release', 'process', 'common', 'r  
['the', 'o'reilly', 'security', 'podcast', 'coarse-grained', 'embracing', 'ephemera  
['twitter', 'financially', 'vr', 'skeptically', 'electronica', 'originally', 'stack  
['this', 'report', 'explores', 'how', 'political', 'helps', 'everything', 'from', 'r  
['this', 'chapter', 'from', 'lean', 'ux', '2nd', 'edition', 'how', 'blending', 'lea  
['crafted', 'by', 'experienced', 'practitioners', 'patterns', 'can', 'flexible', 'r  
['best', 'practices', 'for', 'web', 'web', 'web', 'not', 'paying', 'enough', 'websi  
['paul', 'english's', 'vision', 'pairing', 'ai', 'human', 'expertise', 'has', 'us',  
['matt', 'jorgensen', 'look', 'josephine', 'platform', 'connects', 'consumers', 'wi  
['rana', 'forroohar', 'makers', 'takers', 'repercussions', 'reading', 'makers', 'a  
['the', 'behind', 'john', 'basset', 'iii', 'family', 'furniture', 'company', 'prove  
['at', 'next', 'economy', '2016', 'business', 'leaders', 'policy', 'charted', 'cour  
['innovation', 'economics', 'decentralized', 'prediction', 'markets', 'front-end',  
['why', 'new', 'factor', 'can', 'between', 'failure', 'cloud.in', 'beyond', 'twelve  
['five', 'questions', 'for', 'laura', 'mather', 'insights', 'how', 'groupthink', 'h  
['douglas', 'rushkoff', 'way', 'forward', 'local', 'circular', 'economy', 'where',  
['honor', 'using', 'real', 'vexing', 'problem', 'for', 'whose', 'loved', 'need', 'c  
['james', 'nord', 'fohr', 'card', 'us', 'how', 'reputation', 'gained', 'gets', 'cor  
['graphs', 'from', 'max', 'roser's', 'our', 'world', 'data', 'have', 'become', 'fix  
['linkedin', 'ceo', 'jeff', 'weiner', 'what', 'linkedin', 'us', 'how', 'we', 'can',  
['keller', 'rinaudo', 'company', 'zipline', 'using', 'with', 'poor', 'continue', 'r  
['niantic', 'founder', 'ceo', 'john', 'hanke', 'his', 'creating', 'pokémon', 'go',  
['natalie', 'foster', 'aspen', 'institute', 'hosts', 'conversation', 'with', 'andy']



```

In [165]: cluster = KMeansClusterer(10, cosine_distance) ###cosine distance from n
          cluster.cluster([vectorspaced(title) for title in job_titles if title]) #
          classified_examples = [cluster.classify(vectorspaced(title)) for title in

In [166]: for cluster_id, title in sorted(zip(classified_examples, job_titles)):
          print ("\n", "Cluster number:", cluster_id, "\n", "Words extracted from

Cluster number: 0
Words extracted from the blog representing individual blog:

['5', 'questions', 'for', 'binu', 'ramakrishnan', 'insights', 'centralized', 'ci/c

Cluster number: 0
Words extracted from the blog representing individual blog:

['5', 'questions', 'for', 'timothy', 'gallo', 'allan', 'liska', 'insights', 'commo

Cluster number: 0
Words extracted from the blog representing individual blog:

['best', 'practices', 'for', 'web', 'web', 'web', 'not', 'paying', 'enough', 'webs

Cluster number: 0
Words extracted from the blog representing individual blog:

['crafted', 'by', 'experienced', 'practitioners', 'patterns', 'can', 'flexible', '

Cluster number: 0
Words extracted from the blog representing individual blog:

['five', 'questions', 'for', 'laura', 'mather', 'insights', 'how', 'groupthink', '

Cluster number: 0
Words extracted from the blog representing individual blog:

['insider', "o'reilly", 'security', 'conference', 'proposal', 'process', 'rejection

Cluster number: 0
Words extracted from the blog representing individual blog:

['the', 'o'reilly', 'security', 'podcast', 'coarse-grained', 'embracing', 'ephemer

Cluster number: 0
Words extracted from the blog representing individual blog:

```

['word', 'embedding', 'natural', 'language', 'processing.word', 'embedding', 'word

Cluster number: 1

Words extracted from the blog representing individual blog:

['danese', 'cooper', 'explores', 'innersource', 'within', 'walls', 'reading', 'how

Cluster number: 1

Words extracted from the blog representing individual blog:

['lauri', 'apple', 'zalando', 'how', 'company.continue', 'reading', 'building', 'c

Cluster number: 1

Words extracted from the blog representing individual blog:

['learn', 'how', 'capital', 'one', 'continuous', 'release', 'process', 'common', 'c

Cluster number: 1

Words extracted from the blog representing individual blog:

['roberto', 'di', 'cosmo', 'reveals', 'behind', 'launch', 'software', 'heritage', 'a

Cluster number: 1

Words extracted from the blog representing individual blog:

['tim', 'frick', 'products', 'reading', 'how', 'for']

Cluster number: 1

Words extracted from the blog representing individual blog:

['understand', 'how', 'capital', 'one', 'embraced', 'continuous', 'learn', 'core', 'c

Cluster number: 2

Words extracted from the blog representing individual blog:

['rana', 'foroohar', 'makers', 'takers', 'repercussions', 'reading', 'makers', 'a

Cluster number: 2

Words extracted from the blog representing individual blog:

['the', "o'reilly", 'radar', 'podcast', 'prediction', 'cognitive', 'biases', 'how

Cluster number: 2

Words extracted from the blog representing individual blog:

['the', 'behind', 'john', 'basset', 'iii', 'family', 'furniture', 'company', 'prov

Cluster number: 2

Words extracted from the blog representing individual blog:

['the', 'o'reilly', 'data', 'show', 'podcast', 'natalino', 'busa', 'feature', 'eng

Cluster number: 3

Words extracted from the blog representing individual blog:

['amazing', 'china', 'deep', 'learning', 'memory', 'vr', 'spaces', 'javascript', 'f

Cluster number: 3

Words extracted from the blog representing individual blog:

['graphs', 'from', 'max', 'roser's', 'our', 'world', 'data', 'have', 'become', 'fi

Cluster number: 3

Words extracted from the blog representing individual blog:

['james', 'nord', 'fohr', 'card', 'us', 'how', 'reputation', 'gained', 'gets', 'co

Cluster number: 3

Words extracted from the blog representing individual blog:

['karen', 'sandler', 'freedom', 'underlies', 'effectively', 'problems', 'we', 'nee

Cluster number: 3

Words extracted from the blog representing individual blog:

['lessons', 'learned', 'fastly', 'how', 'build', 'robust', 'contains', 'properly',

Cluster number: 3

Words extracted from the blog representing individual blog:

['open', 'licenses', 'have', 'us', 'well', 'for', 'but', 'cory', 'doctorow', 'need

Cluster number: 3

Words extracted from the blog representing individual blog:

['the', 'o'reilly', 'bots', 'podcast', 'how', 'bots', 'way', 'companies', 'with',

Cluster number: 3

Words extracted from the blog representing individual blog:

['this', 'report', 'explores', 'how', 'political', 'helps', 'everything', 'from',

Cluster number: 3

Words extracted from the blog representing individual blog:

['watch', 'highlights', 'covering', 'business', 'from', 'oscon', 'london', '2016.p

Cluster number: 4

Words extracted from the blog representing individual blog:

['kid', 'privacy', 'chinese', 'creation', 'reality', 'rocks', 'chaos', 'monkey',

Cluster number: 4

Words extracted from the blog representing individual blog:

['twitter', 'financially', 'vr', 'skeptically', 'electronica', 'originally', 'stac

Cluster number: 5

Words extracted from the blog representing individual blog:

['a', 'look', 'expanding', 'biotech', 'landscape', 'sxsw', 'south', 'by', 'southwe

Cluster number: 5

Words extracted from the blog representing individual blog:

['cake', 'cutting', 'tea', 'making', 'google', 'interviewing', 'automation', 'puzz

Cluster number: 5

Words extracted from the blog representing individual blog:

['innovation', 'economics', 'decentralized', 'prediction', 'markets', 'front-end',

Cluster number: 6

Words extracted from the blog representing individual blog:

['if', 'we', 'let', 'put', 'us', 'work', 'will', 'be', 'because', 'failure', 'lack

Cluster number: 6

Words extracted from the blog representing individual blog:

['learn', 'why', 'svg', 'from', 'css', 'work', 'common', 'hiccups', 'experience',

Cluster number: 6

Words extracted from the blog representing individual blog:

['male', 'allies', 'learning', 'maps', 'machine', 'learning', 'biology', '``', 'th

Cluster number: 6

Words extracted from the blog representing individual blog:

['take', 'learning', 'followed', 'previous', 'post', 'learned', 'how', 'gpu-accele

Cluster number: 6

Words extracted from the blog representing individual blog:

['the', 'beyond', 'twelve-factor', 'app', 'i', 'present', 'new', 'guidelines', 'bu

Cluster number: 6

Words extracted from the blog representing individual blog:

['this', 'chapter', 'from', 'lean', 'ux', '2nd', 'edition', 'how', 'blending', 'le

Cluster number: 6

Words extracted from the blog representing individual blog:

['this', 'free', 'webcast', 'provides', 'quick', 'lesson', 'updating', 'relational

Cluster number: 6

Words extracted from the blog representing individual blog:

['why', 'new', 'factor', 'can', 'between', 'failure', 'cloud.in', 'beyond', 'twelv

Cluster number: 7

Words extracted from the blog representing individual blog:

['awesome', 'falsehoods', 'neural', 'network', 'vision', 'toy', 'family', 'robots'

Cluster number: 8

Words extracted from the blog representing individual blog:

['a', 'common', 'java', 'pass', 'by', 'reference', 'language', 'in', 'video', 'bri

Cluster number: 8

Words extracted from the blog representing individual blog:

['at', 'next', 'economy', '2016', 'business', 'leaders', 'policy', 'charted', 'cou

Cluster number: 8

Words extracted from the blog representing individual blog:

['author', 'mark', 'richards', 'walks', '10', 'common', 'pitfalls', 'provides', 'f

Cluster number: 8

Words extracted from the blog representing individual blog:

['honor', 'using', 'real', 'vexing', 'problem', 'for', 'whose', 'loved', 'need', '

Cluster number: 8

Words extracted from the blog representing individual blog:

['in', 'video', 'brian', 'l.', 'gorman', 'uncovers', 'relationship', 'between', 'e

Cluster number: 8

Words extracted from the blog representing individual blog:

['keller', 'rinaudo', 'company', 'zipline', 'using', 'with', 'poor', 'continue', 'v

Cluster number: 8

Words extracted from the blog representing individual blog:

['kris', 'borchers', 'javascript.continue', 'reading', 'evolving', 'javascript', 'v

Cluster number: 8

Words extracted from the blog representing individual blog:

['mark', 'shuttleworth', 'next', 'likely', 'come', 'from', 'laptop', 'pc', 'beatin

Cluster number: 8

Words extracted from the blog representing individual blog:

['matt', 'jorgensen', 'look', 'josephine', 'platform', 'connects', 'consumers', 'v

Cluster number: 8

Words extracted from the blog representing individual blog:

['natalie', 'foster', 'aspen', 'institute', 'hosts', 'conversation', 'with', 'andy

Cluster number: 8

Words extracted from the blog representing individual blog:

['the', 'o'reilly', 'bots', 'podcast', 'bots', 'can', 'respond', 'groups', 'users.

Cluster number: 8

Words extracted from the blog representing individual blog:

['the', 'o'reilly', 'data', 'show', 'podcast', 'christopher', 'nguyen', 'early', 'v

Cluster number: 8

Words extracted from the blog representing individual blog:

['the', 'o'reilly', 'design', 'podcast', 'design', 'creating', 'right', 'environme

Cluster number: 8

Words extracted from the blog representing individual blog:

['using', 'rum', 'http', 'vs.', 'https', 'world', 'regions', 'wireline', 'rural',

Cluster number: 8

Words extracted from the blog representing individual blog:

['wan-replicated', 'data', 'nosql', 'modeling', 'sharing', 'roadmaps', 'shopping',

Cluster number: 9

Words extracted from the blog representing individual blog:

['douglas', 'rushkoff', 'way', 'forward', 'local', 'circular', 'economy', 'where',

Cluster number: 9

Words extracted from the blog representing individual blog:

['linkedin', 'ceo', 'jeff', 'weiner', 'what', 'linkedin', 'us', 'how', 'we', 'can

Cluster number: 9

Words extracted from the blog representing individual blog:

['niantic', 'founder', 'ceo', 'john', 'hanke', 'his', 'creating', 'pokémon', 'go',

Cluster number: 9

Words extracted from the blog representing individual blog:

['paul', 'english's', 'vision', 'pairing', 'ai', 'human', 'expertise', 'has', 'us