FEDERATED PORTAL FOR FINANCE FORUMS

CS 410 Text Retrieval and Mining
MCSDS Course Final Project
Gayatri Balakrishnan(gayatri3)

Objective

There are many online forums related to finance and personal investments.

Current forums are all scattered on the Internet with little connection to each other, leading to many isolated fragmented communities on similar topics

This project seeks to leverage the idea of Federated Forum Portal and apply it specifically to finance and personal investment forums.

This project expects to leverage the concepts and methods described in Prof.Cheng's post on Federated Forums and apply them in creating a Federated Portal model for the personal finance domain that would help users in finding the appropriate forums and in collating information posted on similar threads across the same or different forums.

High Level Approach

The proposal is to

- build a web crawler for a finance-related forum
- build an inverted index and topic model to enable topic map construction
- rank topics using the topic model in conjunction with user demographics, freshness of the posts and number of participants.
- create a visualization interface showing the user trending topics in the personal finance area

Steps to execute

Prerequisites:

- Must have Python, Jupyter (for reading and executing IPython Notebooks), Scrapy Web_Crawling Framework and Scikit-Learn installed.

Steps:

Part 1: Crawl the website and get the forum threads

1. Create a project using the below command:

scrapy startproject cs410fedforum

2. Copy the file finance_spider.py into the cs410fedforum/spiders folder

3. Go to the project's top level directory and run:

scrapy crawl financeforums -o financetopics.json

Demo Screenshots

- Part 1: Crawl the website and get the forum threads

Command Prompt

```
C:\Users\gayat>scrapy startproject cs410fedforum
:0: UserWarning: You do not have a working installation of the service_identity module: 'cannot import name opentype'.  Please install i
t from <https://pypi.python.org/pypi/service_identity> and make sure all of its dependencies are satisfied.  Without the service_identit
y module, Twisted can perform only rudimentary TLS client hostname verification.  Many valid certificate/hostname mappings may be reject
ed.
Traceback (most recent call last):
  File "d:\anaconda\anaconda2\lib\runpy.py", line 174, in _run_module_as_main
    "__main__", fname, loader, pkg_name)
  File "d:\anaconda\anaconda2\lib\runpy.py", line 72, in _run_code
    exec code in run_globals
  File "D:\Anaconda\Anaconda2\Scripts\scrapy.exe\__main__.py", line 9, in <module>
  File "d:\anaconda\anaconda2\lib\site-packages\scrapy\cmdline.py", line 149, in execute
    _run_print_help(parser, _run_command, cmd, args, opts)
  File "d:\anaconda\anaconda2\lib\site-packages\scrapy\cmdline.py", line 89, in _run_print_help
    func(*a, **kw)
  File "d:\anaconda\anaconda2\lib\site-packages\scrapy\cmdline.py", line 156, in _run_command
    cmd.run(args, opts)
  File "d:\anaconda\anaconda2\lib\site-packages\scrapy\commands\startproject.py", line 109, in run
    ProjectName=string_camelcase(project_name))
  File "d:\anaconda\anaconda2\lib\site-packages\scrapy\utils\template.py", line 9, in render_templatefile
    with open(path, 'rb') as fp:
IOError: [Errno 2] No such file or directory: 'cs410fedforum\\cs410fedforum\\settings.py.tmpl'

C:\Users\gayat>dir cs410fedforum
 Volume in drive C is OS
 Volume Serial Number is 7C27-EC47

 Directory of C:\Users\gayat\cs410fedforum

12/22/2017  09:17 AM    <DIR>          .
12/22/2017  09:17 AM    <DIR>          ..
12/22/2017  09:17 AM    <DIR>          cs410fedforum
12/22/2017  09:17 AM               270 scrapy.cfg
               1 File(s)            270 bytes
               3 Dir(s)   8,617,066,496 bytes free

C:\Users\gayat>
```

spiders

File Home Share View

This PC › OS (C:) › Users › gayat › cs410fedforum › cs410fedforum › spiders

Search spiders

Quick access
Desktop
Downloads
Documents
Pictures
Business

Name	Date modified	Type	Size
__init__.py	12/3/2017 8:37 PM	Python File	1 KB
finance_spider.py	12/21/2017 4:29 PM	Python File	3 KB

Demo Screenshots

- Part 1: Crawl the website and get the forum threads

Command Prompt - scrapy crawl financeforums -o financetopics.json

```
C:\Users\gayat>cd cs410fedforum

C:\Users\gayat\cs410fedforum>scrapy crawl financeforums -o financetopics.json
:0: UserWarning: You do not have a working installation of the service_identity module: 'cannot import name opentype'.  Please install i
t from <https://pypi.python.org/pypi/service_identity> and make sure all of its dependencies are satisfied.  Without the service_identit
y module, Twisted can perform only rudimentary TLS client hostname verification.  Many valid certificate/hostname mappings may be reject
ed.
2017-12-22 09:23:31 [scrapy.utils.log] INFO: Scrapy 1.4.0 started (bot: cs410fedforum)
2017-12-22 09:23:31 [scrapy.utils.log] INFO: Overridden settings: {'NEWSPIDER_MODULE': 'cs410fedforum.spiders', 'FEED_URI': 'financetopi
cs.json', 'SPIDER_MODULES': ['cs410fedforum.spiders'], 'BOT_NAME': 'cs410fedforum', 'ROBOTSTXT_OBEY': True, 'FEED_FORMAT': 'json'}
2017-12-22 09:23:33 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.feedexport.FeedExporter',
 'scrapy.extensions.logstats.LogStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.corestats.CoreStats']
2017-12-22 09:23:35 [scrapy.middleware] INFO: Enabled downloader middlewares:
['scrapy.downloadermiddlewares.robotstxt.RobotsTxtMiddleware',
 'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
 'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
 'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
 'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
 'scrapy.downloadermiddlewares.retry.RetryMiddleware',
 'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',
 'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddleware',
 'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
 'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
 'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware',
 'scrapy.downloadermiddlewares.stats.DownloaderStats']
2017-12-22 09:23:35 [scrapy.middleware] INFO: Enabled spider middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.offsite.OffsiteMiddleware',
 'scrapy.spidermiddlewares.referer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
2017-12-22 09:23:35 [scrapy.middleware] INFO: Enabled item pipelines:
[]
2017-12-22 09:23:35 [scrapy.core.engine] INFO: Spider opened
2017-12-22 09:23:36 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0 pages/min), scraped 0 items (at 0 items/min)
2017-12-22 09:23:36 [scrapy.extensions.telnet] DEBUG: Telnet console listening on 127.0.0.1:6023
```

Demo Screenshots

- Part 1: Crawl the website and get the forum threads

Command Prompt

```
2017-12-22 09:25:20 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.thefinanceforums.com/forumdisplay.php?f=1&order=desc&page=
99>
{'date': u'12-15-2006', 'text': u'Leasing vs. buying a car?', 'views': u'3,150', 'replies': u'14', 'time': u'01:28 AM'}
2017-12-22 09:25:20 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.thefinanceforums.com/forumdisplay.php?f=1&order=desc&page=
99>
{'date': u'12-02-2006', 'text': u'Hello', 'views': u'2,314', 'replies': u'4', 'time': u'01:58 PM'}
2017-12-22 09:25:20 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.thefinanceforums.com/forumdisplay.php?f=1&order=desc&page=
99>
{'date': u'11-29-2006', 'text': u'Total loan', 'views': u'2,009', 'replies': u'0', 'time': u'06:20 AM'}
2017-12-22 09:25:20 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.thefinanceforums.com/forumdisplay.php?f=1&order=desc&page=
99>
{'date': u'11-29-2006', 'text': u'Hire Purchase vs Loans', 'views': u'1,939', 'replies': u'3', 'time': u'05:30 AM'}
2017-12-22 09:25:20 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.thefinanceforums.com/forumdisplay.php?f=1&order=desc&page=
99>
{'date': u'11-29-2006', 'text': u'CDs v. Mutual Funds', 'views': u'2,526', 'replies': u'7', 'time': u'05:28 AM'}
2017-12-22 09:25:20 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.thefinanceforums.com/forumdisplay.php?f=1&order=desc&page=
99>
{'date': u'11-29-2006', 'text': u'Tax Help Associates', 'views': u'1,853', 'replies': u'2', 'time': u'05:22 AM'}
2017-12-22 09:25:20 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.thefinanceforums.com/forumdisplay.php?f=1&order=desc&page=
99>
{'date': u'11-29-2006', 'text': u'Democrats Take Over the House (maybe Senate)', 'views': u'2,146', 'replies': u'8', 'time': u'05:18 AM'
}
2017-12-22 09:25:20 [scrapy.core.engine] INFO: Closing spider (finished)
2017-12-22 09:25:20 [scrapy.extensions.feedexport] INFO: Stored json feed (8159 items) in: financetopics.json
2017-12-22 09:25:20 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{'downloader/request_bytes': 294081,
 'downloader/request_count': 418,
 'downloader/request_method_count/GET': 418,
 'downloader/response_bytes': 5268875,
 'downloader/response_count': 418,
 'downloader/response_status_count/200': 418,
 'dupefilter/filtered': 399,
 'finish_reason': 'finished',
 'finish_time': datetime.datetime(2017, 12, 22, 3, 55, 20, 635000),
 'item_scraped_count': 8159,
 'log_count/DEBUG': 8579,
 'log_count/INFO': 9,
 'request_depth_max': 99,
 'response_received_count': 418,
 'scheduler/dequeued': 417,
 'scheduler/dequeued/memory': 417,
 'scheduler/enqueued': 417,
 'scheduler/enqueued/memory': 417,
 'start_time': datetime.datetime(2017, 12, 22, 3, 53, 36, 6000)}
2017-12-22 09:25:20 [scrapy.core.engine] INFO: Spider closed (finished)
```

cs410fedforum

File Home Share View

This PC › OS (C:) › Users › gayat › cs410fedforum

Quick access
Desktop
Downloads
Documents
Pictures

Name	Date modified	Type	Size
cs410fedforum	12/22/2017 9:23 AM	File folder	
financetopics.json	12/22/2017 9:25 AM	JSON File	1,029 KB
scrapy.cfg	12/22/2017 9:17 AM	CFG File	1 KB

Steps to execute

- Part 2: Training the classifier

1. Download the data directory which contains the training data for categories

2. Download the CS410FedForums-Categorization IPython notebook

3. Copy the json file financetopics.json from Part 1 above to the same path as the CS410FedForums-Categorization IPython notebook

4. In the CS410FedForums-Categorization IPython notebook, replace the path of the data directory in the first parameter of load_files on this line:

docs_to_train = sklearn.datasets.load_files("C:\Users\gayat\ipythonnotebooks\data", description=None, categories=None, load_content=True, shuffle=True, encoding='utf-8', decode_error='strict', random_state=0)

5. In the CS410FedForums-Categorization IPython notebook, replace the path in this line with wherever you want the output file to be created:

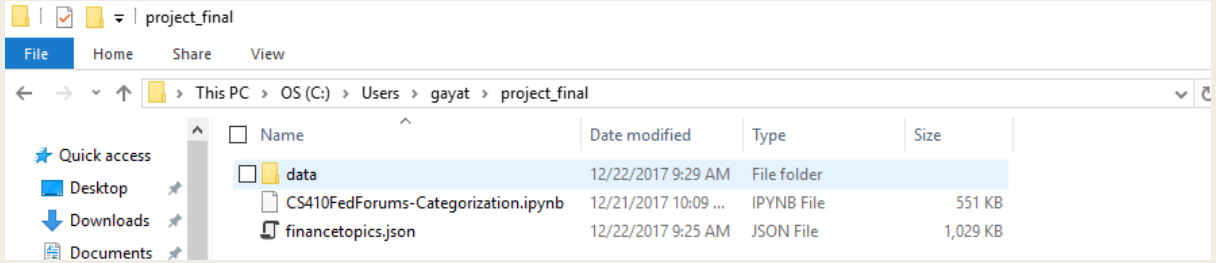
filename="C:\Users\gayat\ipythonnotebooks\categorization_output.txt"

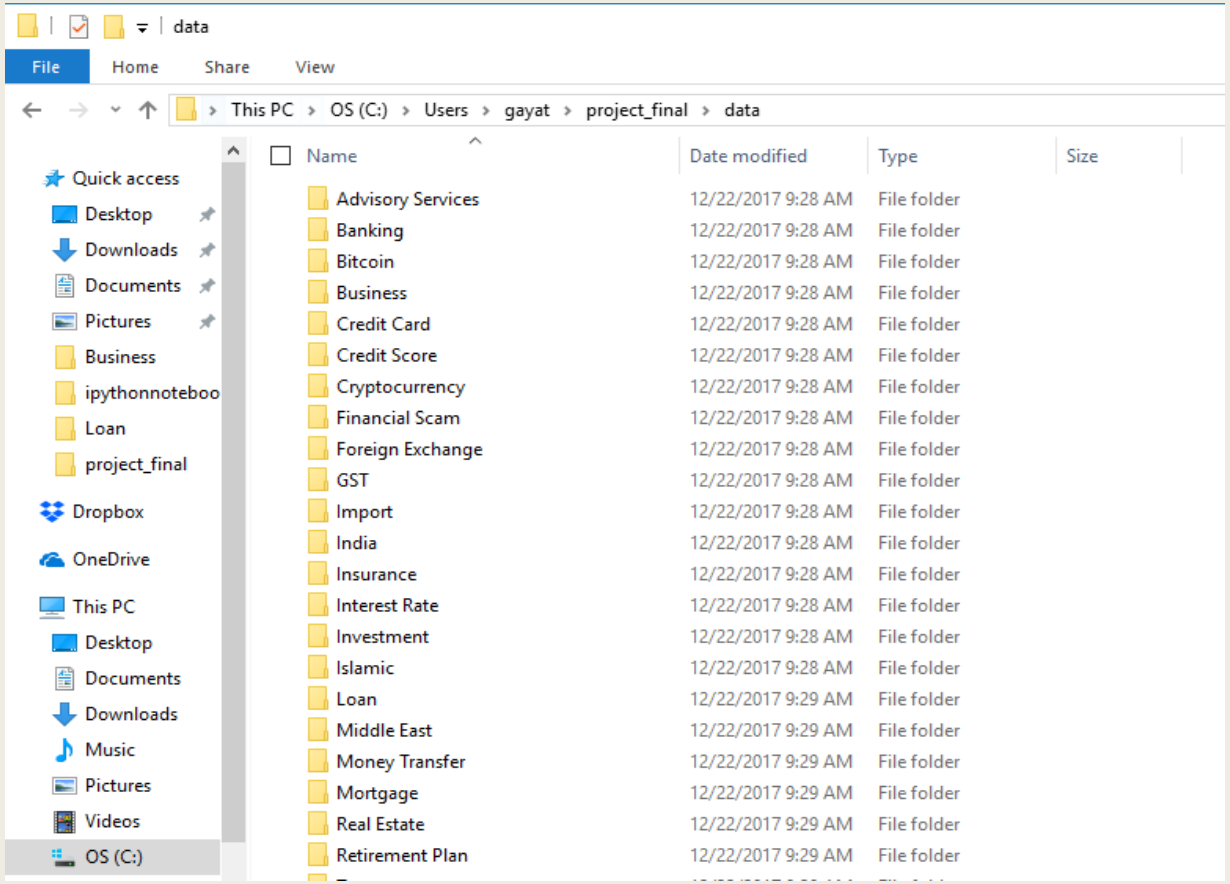
6. Run the CS410FedForums-Categorization IPython notebook.

Note: See Deep Dive section for a brief description of how the prediction accuracy was evaluated

Demo Screenshots

- Part 2: Training the classifier





Demo Screenshots

- Part 2: Training the classifier

localhost:8888/notebooks/project_final/CS410FedForums-Categorization.ipynb

jupyter CS410FedForums-Categorization (unsaved changes)

File Edit View Insert Cell Kernel Help | Python 2

Code CellToolbar

In [1]:
```
import sklearn
import pandas as pd
testdata = pd.read_json("financetopics.json")
testdata.text
from sklearn.feature_extraction.text import CountVectorizer
count_vect = CountVectorizer()
X_train_counts = count_vect.fit_transform(testdata.text)
X_train_counts.shape
count_vect.vocabulary_
from sklearn import datasets
from pprint import pprint

docs_to_train = sklearn.datasets.load_files("C:\Users\gayat\project_final\data", description=None, categories=None, load_content=True, shuffle=True,
                                            encoding='utf-8', decode_error='strict', random_state=0)

pprint(list(docs_to_train.target_names))
len(docs_to_train.data)

stpwrds = [".",
" ",
",",
"?",
"!",
"'",
"\"",
"''",
```

jupyter CS410FedForums-Categorization (unsaved changes)

File Edit View Insert Cell Kernel Help | Python 2

Code CellToolbar

```
"I","am","on","the","look","out","for","a","pick","up","truck","that","will","be","affordable","yet","reliable.",
"What","is","the","common","cause","of","urine","leakage","in","males?",
"i","have","an","Idea",
"Introduction",
"What","kind","of","safety","features","does","the","Yaris","have?",
"The","essence","of","the","facts",
"I","have","been","hearing","a","lot","about","Toyota","brand","in","KSA.","How","is","it","in","terms","of","prices",
"Please","try","to","prevent","the","spam","posts",
"Posting","a","external","link","earns","a","permanent","account","ban?","Very","fair","and","friendly",
"Congratulations!","You","have","just","managed","to","disgust","a","new","member","in","3","days!",
"hair","throat","skin","sore","removal","sex","condom","condoms","incontinence","toyota","wax","stain","stains","gmail","love","ie

no_features = 10000
from sklearn.feature_extraction.text import CountVectorizer
#count_vect = CountVectorizer()
count_vect = CountVectorizer(max_df=0.95, min_df=2, max_features=no_features, stop_words=stpwrds)
X_train_counts = count_vect.fit_transform(docs_to_train.data)
X_train_counts.shape

tf_feature_names = count_vect.get_feature_names()
from sklearn.feature_extraction.text import TfidfTransformer
tf_transformer = TfidfTransformer(use_idf=False).fit(X_train_counts)
X_train_tf = tf_transformer.transform(X_train_counts)
X_train_tf.shape

tfidf_transformer = TfidfTransformer()
X_train_tfidf = tfidf_transformer.fit_transform(X_train_counts)
X_train_tfidf.shape

from sklearn.naive_bayes import MultinomialNB
clf = MultinomialNB().fit(X_train_tfidf, docs_to_train.target)

docs_new = testdata.text
X_new_counts = count_vect.transform(docs_new)
X_new_tfidf = tfidf_transformer.transform(X_new_counts)
predicted = clf.predict(X_new_tfidf)
filename="C:\Users\gayat\project_final\categorization_output.txt"
file = open(filename,"w")
for doc, category in zip(docs_new, predicted):
    print('%r => %s' % (doc, docs_to_train.target_names[category]))
    file.write("%r => %s" % (doc, docs_to_train.target_names[category]))
    file.write("\n")
file.close()
```

Demo Screenshots

- Part 2: Training the classifier

jupyter CS410FedForums-Categorization (autosaved)

File Edit View Insert Cell Kernel Help | Python 2

Code CellToolbar

```
tf_feature_names = count_vect.get_feature_names()
from sklearn.feature_extraction.text import TfidfTransformer
tf_transformer = TfidfTransformer(use_idf=False).fit(X_train_counts)
X_train_tf = tf_transformer.transform(X_train_counts)
X_train_tf.shape

tfidf_transformer = TfidfTransformer()
X_train_tfidf = tfidf_transformer.fit_transform(X_train_counts)
X_train_tfidf.shape

from sklearn.naive_bayes import MultinomialNB
clf = MultinomialNB().fit(X_train_tfidf, docs_to_train.target)

docs_new = testdata.text
X_new_counts = count_vect.transform(docs_new)
X_new_tfidf = tfidf_transformer.transform(X_new_counts)
predicted = clf.predict(X_new_tfidf)
filename="C:\Users\gayat\project_final\categorization_output.txt"
file = open(filename,"w")
for doc, category in zip(docs_new, predicted):
    print('%r => %s' % (doc, docs_to_train.target_names[category]))
    file.write("%r => %s" % (doc, docs_to_train.target_names[category]))
    file.write("\n")
file.close()
```

```
 'Technology',
 'Trading',
 'VAT']
u'How Blended Gross Profit Margin is Calculated' => Banking
u'Where should I invest money for getting higher return?' => Advisory Services
u'Digital currency in Dubai from Jan, 2018??' => Middle East
u'Future of Cryptocurrencies and Role of Money Trade Coin in the market' => Advisory Services
u'How is Money Trade Coin Gaining base in the Crypto world?' => Advisory Services
u'lease/sale, such as BG,SBLC, MTN, Bank Bonds' => Banking
u'Safer Results in Trading with Money Trade Coin' => Advisory Services
u'What makes Money Trade coin a Perfect Option for International Transactions' => Advisory Services
u'GST Rules' => Banking
u'I have heard that Abdul Latif Jameel offer financial services also. Can you elaborate' => Advisory Service
u'Dinar boxes metallic from paraguay' => Banking
u'Introducing our financial instruments for funding/financing' => Advisory Services
u'Bitcoin jumps more than 9% after news Square is testing the digital currency' => Banking
u'How can I Illegally change My Identity Completely' => Banking
u'I Need money transfer service in Melbourne Australia' => Advisory Services
u'What is your biggest money problem?' => Advisory Services
u'What do banks look for when reviewing a loan application?' => Loan
```

project_final

File Home Share View

This PC › OS (C:) › Users › gayat › project_final

Quick access
Desktop
Downloads
Documents
Pictures
Business

Name	Date modified	Type	Size
.ipynb_checkpoints	12/22/2017 9:34 AM	File folder	
data	12/22/2017 9:29 AM	File folder	
categorization_output.txt	12/22/2017 9:36 AM	Text Document	451 KB
CS410FedForums-Categorization.ipynb	12/22/2017 9:38 AM	IPYNB File	551 KB
financetopics.json	12/22/2017 9:25 AM	JSON File	1,029 KB

Steps to execute

Part 3: Creating the Topic Map

1. Download the NMF_LDA_TopicMap IPython notebook.

2. Copy the json file financetopics.json from Part 1 above to the same path as the NMF_LDA_TopicMap IPython notebook.

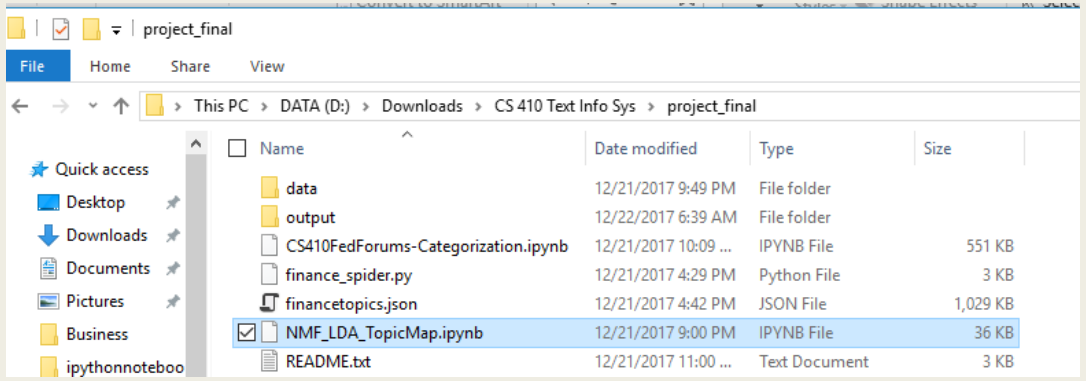
3. In the NMF_LDA_TopicMap IPython notebook, replace the path in this line with wherever you want the output file to be created:

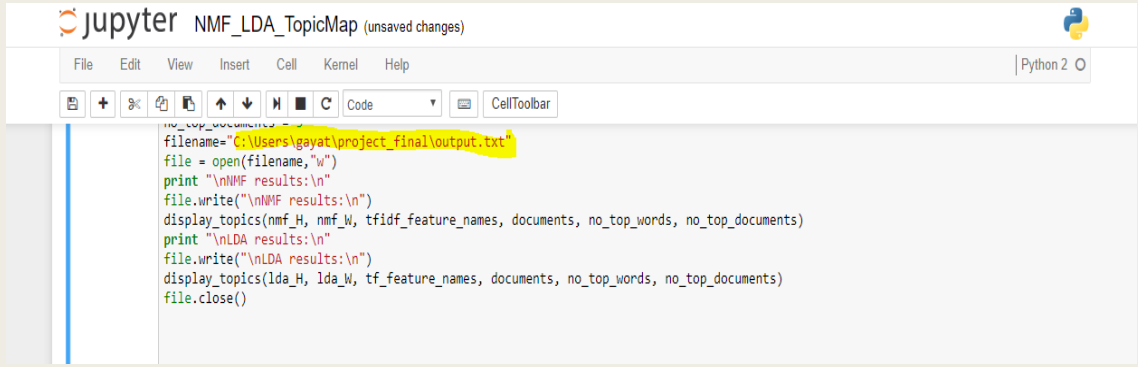
filename="C:\Users\gayat\ipythonnotebooks\output.txt"

4. Run the NMF_LDA_TopicMap IPython notebook

Demo Screenshots

- Part 3: Creating the Topic Map





Demo Screenshots

- Part 3: Creating the Topic Map

jupyter NMF_LDA_TopicMap Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Help | Python 2

Code CellToolbar

```
# Run NMF
nmf_model = NMF(n_components=no_topics, random_state=1, alpha=.1, l1_ratio=.5, init='nndsvd').fit(tfidf)
nmf_W = nmf_model.transform(tfidf)
nmf_H = nmf_model.components_

# Run LDA
lda_model = LatentDirichletAllocation(n_topics=no_topics, max_iter=5, learning_method='online', learning_offset=50.,random_state=0).fit(tf
lda_W = lda_model.transform(tf)
lda_H = lda_model.components_

no_top_words = 10
no_top_documents = 5
filename="C:\Users\gayat\project_final\output.txt"
file = open(filename,"w")
print "\nNMF results:\n"
file.write("\nNMF results:\n")
display_topics(nmf_H, nmf_W, tfidf_feature_names, documents, no_top_words, no_top_documents)
print "\nLDA results:\n"
file.write("\nLDA results:\n")
display_topics(lda_H, lda_W, tf_feature_names, documents, no_top_words, no_top_documents)
file.close()
```

```
NMF results:

Topic 0:
hello friends forum members im people world newbie investors california
Hello
Hello everyone!
hello
hello :)
hello
Topic 1:
hi friends newbie im post uk people future old checking
Hi guys! New here!
hi
Hi I'm Hal
Hi.
Hi, everyone!
Topic 2:
loan personal take offer student apply auto payday eligibility interest
What was your loan for?
What is lawsuit loan?
```

Demo Screenshots

- Part 3: Creating the Topic Map

jupyter NMF_LDA_TopicMap Last Checkpoint: 14 hours ago (autosaved)

File Edit View Insert Cell Kernel Help Python 2

Code CellToolbar

```
# Run NMF
nmf_model = NMF(n_components=no_topics, random_state=1, alpha=.1, l1_ratio=.5, init='nndsvd').fit(tfidf)
nmf_W = nmf_model.transform(tfidf)
nmf_H = nmf_model.components_

# Run LDA
lda_model = LatentDirichletAllocation(n_topics=no_topics, max_iter=5, learning_method='online', learning_offset=50.,random_state=0).fi
lda_W = lda_model.transform(tf)
lda_H = lda_model.components_

no_top_words = 10
no_top_documents = 5
filename="C:\Users\gayat\ipythonnotebooks\output.txt"
file = open(filename,"w")
print "\nNMF results:\n"
file.write("\nNMF results:\n")
display_topics(nmf_H, nmf_W, tfidf_feature_names, documents, no_top_words, no_top_documents)
print "\nLDA results:\n"
file.write("\nLDA results:\n")
display_topics(lda_H, lda_W, tf_feature_names, documents, no_top_words, no_top_documents)
file.close()
```

```
How long do I have to pay for PMI?

LDA results:

Topic 0:
business bank pay suggest college sell exchange lease bonds cut
Fresh cut bank instrument for lease/sale, such as BG,SBLC, MTN, Bank Bonds
Fresh cut bank instrument for lease/sale, such as BG,SBLC, MTN, Bank Bonds
Fresh cut bank instrument for lease/sale, such as BG,SBLC, MTN, Bank Bonds
Fresh cut bank instrument for lease/sale, such as BG,SBLC, MTN, Bank Bonds
Fresh cut bank instrument for lease/sale, such as BG,SBLC, MTN, Bank Bonds
Topic 1:
credit card company start rate planning find islamic payment newbie
Where can I find a credit card with really low APR/interest rate?
Things to keep in mind before chosing a credit card?
Are there any surrender charges on Emirates Islamic's Ethra'a's plans?
Are there any surrender charges on Emirates Islamic's Ethra'a's plans?
Old Credit Card Nightmare
Topic 2:
401k value retirement fund capital roth right taxes stocks benefits
```

Steps to execute

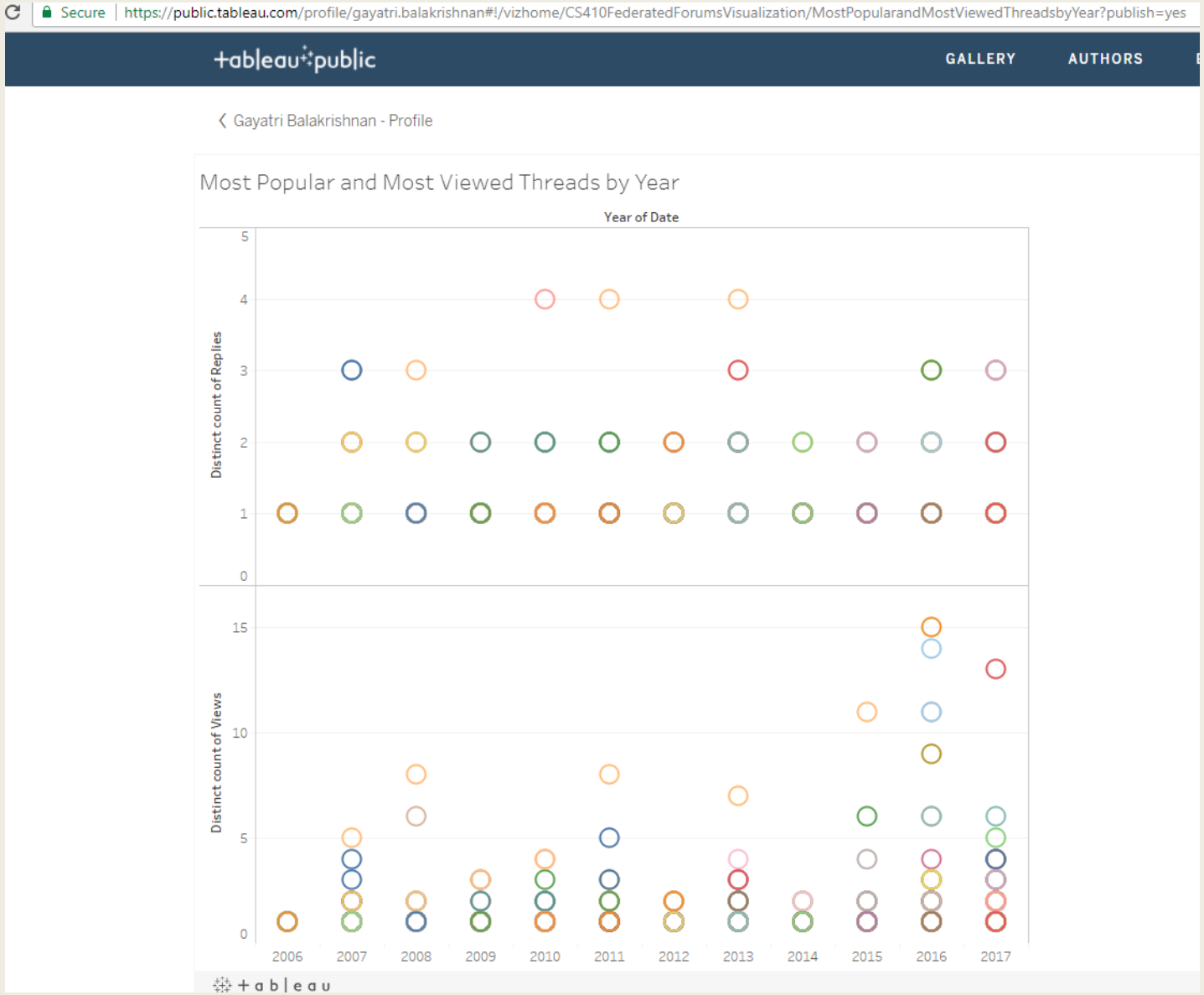
Part 4: Visualization of the data

The json file from Part 1 above was used to create a quick visualization of top trending forum threads by views, replies and dates. This visualization is available here:

- https://public.tableau.com/profile/gayatri.balakrishnan#!/vizhome/CS410FederatedForumsVisualization/MostPopularandMostViewedThreadsbyYear
- https://public.tableau.com/profile/gayatri.balakrishnan#!/vizhome/CS410FederatedForumsVisualization/MostViewedThreads
- https://public.tableau.com/profile/gayatri.balakrishnan#!/vizhome/CS410FederatedForumsVisualization/MostPopularThreads

Demo Screenshots

- Part 4: Visualization of the data



DEEP DIVE
&
LEARNINGS

Part 1: Crawl the websites to get the forum threads

- The dataset for this project consists of the forum threads along with metrics that would help evaluate popularity and freshness of posts.
- Used the Scrapy framework and XPath selectors
- Why Scrapy?
  - Had evaluated Scrapy, Crawly and Crawley web crawling frameworks
  - found Scrapy to be really useful
    - a rich set of features, with a focus on ease of use and efficiency.
    - a very active community.
    - Recommended for both newbies and experts looking to use web crawlers in their text analysis applications.
- In the financeforums.com website,
  - *crawled the first page of all subforums to get the training dataset of threads*
  - *crawled the General Finance forum threads that span over 10 years. Extracted thread titles, number of views, number of replies and last updated time.*
- Similarly other forums can be crawled as well.

Part 1: Building a web crawler- Learnings

- Using the Robots Exclusion Protocol, many websites disallow crawlers. These include some forums like savingadvice.com. We check the robots.txt page in the website we are visiting to determine whether and what crawling is allowed.
- Careful throttling of crawling speed and design of XPath selectors to avoid a DOS attack on the crawled website.
  - *Scrapy provides the autothrottle extension to determine optimum crawling speed*
- Crawling spiders need to be custom built for every site based on the page design. Hence, the crawling spider code in the project cannot be used as is for a website other than http://www.thefinanceforums.com/

Part 2: Training the classifier

- In the financeforums.com website, crawled the first page of all subforums to get the training dataset of threads.
- Used the scikit-learn package to build the classifier.
  - *Easier to use and Better documentation than MetaPy*
  - *Used approach similar to the 20 news groups dataset in scikit-learn tutorial*
- Training dataset consisted of about 340 threads.
- Created a list of stopwords based on training dataset
- Manually created a set of 26 broad categories based on training dataset
- Used the training dataset to categorize the crawled data from the General Finance subforum
  - *Used the TF-IDF weights and NaiveBayes classifier*
- Used Cranfield evaluation methodology to evaluate accuracy of the categorization
  - *Based of evaluation of 1% of test dataset (Most recent and popular threads), results were found to be 58% accurate*
  - *These results are captured in the Categorization.xlsx file in the output folder.*
    - The finalcat.csv file in the output folder was created using the output of the classifier and the category for the threads from this file was plugged back into Categorization.xlsx

Part 2: Training the classifier-Learnings

- Issues with this approach:
  - *Very labor intensive. Manually categorize training data*
  - *Data cleaning is mandatory*
    - More than 10% of the training and test datasets were found to be spam threads unrelated to the finance domain
    - Stopwords had to be customized to prevent classification of spam threads
- Further steps in this approach
  - *Try other classifiers and finetune parameters to better the prediction accuracy*
    - In the course assignments, K nearest neighbours seemed to give better classification than NaiveBayes.
  - *As more forums are crawled, expand the stopwords list to filter out spam threads*
  - *Can the stopwords list be created as a crowd sourced list?*

Part 3: Topic Map construction

- Less labor intensive than previous approach described in part 2 above
- Stopwords list from part 2 above was reused
- Used the scikit-learn package as it is easier to use
- Classification and topic map construction based on two approaches
  - *Non-negative Matrix Factorization (NMF)*
  - Latent Dirichlet Allocation (LDA)
- Both algorithms take as input a bag of words matrix to produce 2 smaller matrices: a document to topic matrix and a word to topic matrix that when multiplied together reproduce the bag of words matrix with the lowest error.
  - *Lot of research available on both approaches*
- While NMF relies on linear algebra, LDA is based on probabilistic modeling.
- NMF needs a TFIDF Vectorizer to create a bag of words matrix whereas LDA can work with raw counts.

Part 3: Topic Map construction-Learnings

- Based on the relevance of the topic maps created by the two approaches, personal preference is LDA.
- Both NMF and LDA require us to specify the number of topics to produce.
- NMF seems to work better with smaller datasets but LDA is able to get a better topic map even with a larger dataset.
  - *NMF needed a better stopwords list to generate a better topic map with the test dataset.*
  - *NMF also seemed to execute slower(by a few seconds) for larger datasets*
- LDA produced more coherent topics in line with human judgment.
- Also corroborated these findings with research papers published in this area (See references)
- Further steps in this approach would be
  - *Evaluate topic map accuracy algorithmically*
  - *finetune parameters and evaluate accuracy*
  - *Try other algorithms and finetune parameters and evaluate topic map accuracy*
  - *As more forums are crawled, expand the stop words list to filter out spam threads*

Part 4: Visualization of Thread Topics

- Simple visualization built using Tableau
  - *Tableau chosen due to ease of use*
- The initial crawled data consisting of thread titles and metrics was used
- Thread Popularity measured based on
  - *freshness of the posts*
  - *number of views and replies*

Part 4: Visualization of Topics-Learnings

- More than 10% of the training and test datasets were found to be spam threads unrelated to the finance domain
  - *This skews the visualization and some irrelevant threads show up as popular*
- Further steps in this approach
  - *Correlate the topic map and build a visualization of the trending topics instead of just popular threads*
  - *As more forums are crawled, add the originating forum information as well*
    - would help users in finding the appropriate forums for their topics of interest
    - collate information posted on similar threads across the same or different forums
  - *Build the federated forum portal and provide drilldown capability*
  - *Try other kinds of visualization*
  - *Use other data visualization frameworks like D3.js*

References

- Prof Cheng's post on Federated Forums Portal: https://wiki.illinois.edu/wiki/pages/viewpage.action?spaceKey=timanpub&title=Federated+Online+Forum+Portal
- Topic Modeling with Scikit-Learn:

https://medium.com/mlreview/topic-modeling-with-scikit-learn-e80d33668730

- Scikit-Learn tutorial for training classifier using 20 newsgroups dataset:

http://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html

- Scrapy tutorial

https://doc.scrapy.org/en/1.4/intro/tutorial.html#intro-tutorial

- The finance forum used to test:

http://www.thefinanceforums.com/

- Exploring Topic Coherence over many models and many topics- Keith Stevens, Philip Kegelmeyer, David Andrzejewski, David Buttler, University of California Los Angeles; Lawrence Livermore National Lab, Sandia National Lab

http://aclweb.org/anthology/D/D12/D12-1087.pdf