

CSI4107 – Final Search Engine System

ARMAND SYAHTAMA - 8253748

Modules Included

8a – Bigram Models

Functionality

The implementation for the class that builds out the Bigram Models, goes as follows: first on class instantiation, both the UO-courses corpus and the Reuters corpus were loaded from their respective json files. The documents were then stored in a list for further processing when the time comes.

Afterwards, you would call the *build_bigrams* method. First, a dictionary structure is instantiated, so the bigram model would look something like this:

```
Dict = {  
    Coffee: {  
        "Banana": .4,  
        ...  
    },  
    ...  
}
```

The main dictionary would have the first word of the bigram as the key. For the value, it would be a dictionary that contains all the possible 2nd words of the bigram pair, with the first word. (ex. “coffee” would be the key, and it’s value would be a dictionary with the keys of the 2nd word of the pair, like “banana” or “beans”). All those pairs of words would be counted. We iterate through all documents in the Reuters corpus, counting everything until all documents have been iterated through. Finally, we save the bigram model as a json file that can be loaded later.

Challenges

None

8b – Query Completion Module

Functionality

The way query completion works is that, once a user types a singular word into the search bar, several autocomplete options are shown below the bar. Afterwards, users can click on one of the options, and it will complete the query on the search bar. A maximum of 10 different options are shown

To perform this, first we instantiate the bigram model that we made previously into a variable by loading the json file using the python JSON package. This was done in the `__init__.py`

of the *userinterfaces* packages. Next, in the *routes.py* file, the bigram model variable is imported, so we can use it for our query completion

For the autocomplete functionality, it was done like this. First, there's an autocomplete route set up, where when it the URL gets called from the front end, a function would get called, which I called *autocomplete*. This function would take the text written from the search, then check it against the bigram model. If the word is not in the bigram model, it terminates early, returning an empty array of results back to the front-end. Otherwise, it would perform a lookup on the bigram model, retrieving all the 2nd words of the pair. It then takes the 10 words with the top 10 score, and then returns that back into the front-end to be displayed.

In the frontend portion, the code to handle the autocomplete is in the templates folder, specifically in the *base.html* file. Using jQuery, we extract the terms from the search bar, use ajax to call the python functionality through an http endpoint. Once we obtain our autocomplete results from the python code, the results get displayed onto the screen.

Challenges

The biggest issue that came up here was trying to get the jQuery portion of the code to properly work. Initially, no matter what was done, the communication between the frontend jQuery portion and the backend Python portion wouldn't work, which meant that no autocomplete results would ever show up. We would know if the jQuery was working if the command line tells us we are making a request to the autocomplete endpoint, which it didn't seem to do at first. The jQuery problems were solved through trial and error with the frontend code.

9a – Automatic Thesaurus Construction

Functionality

For the thesaurus, first a doc term table was built, taking the all identified terms and then counting their occurrences on a per document basis. Afterwards, I count the number of times a pair of words would occur together in a document and the number of times each word occurs individually for all documents. With those, I use the Jaccard similarity function to calculate term similarity.

$$\frac{(\# \text{ of times both words occurred together in a single document})}{(\# \text{ of total occurrences of word 1} + \# \text{ of occurrences for word 2})}$$

We store all word pairs and scores into a dictionary, which then gets saved into a json file for future use

Challenges

At first, I naively attempted to use the entire dictionary to build out the thesaurus. The biggest issue here is the size of the dictionary in terms of # of words. In this case, the size of the thesaurus in the end is (# of terms in dictionary)², to account for every word pairing possible. With over 10000 words from the dictionary, this thesaurus structure was significantly large. It took several hours to build out this thesaurus each time, and the resulting thesaurus often ended being

several gigabytes in file size. The maximum amount of space you can insert in Brightspace is 1 GB, meaning this thesaurus would be unacceptable. For the sake of time and space, I decided to reduce the # of words used in the thesaurus down to 2000 words, making for a much more manageable file. I chose the 2000 words with the highest number of appearances in all the documents in the corpuses, which I had stored in the inverted index. So I loaded the inverted index, calculated the # of appearances of each word, and then took the words with the highest # of appearances during the building of the doc term table.

9b - Global Query Expansion (in VSM)

Functionality

For Global Query Expansion, I chose to do implicit expansion, appending terms secretly to the user's query when the user chooses to be in Vector Space Model mode. The implicit expansion was done through reworking the retrieve function in the VSM code. First, we loaded our thesaurus that we previously saved as json. Then, once the user makes a search, the query gets tokenized to individual words. The tokens get passed into a function called *expand_tokens_from_thesaurus*, which iterates through each token and looks them up in the thesaurus. If it finds the token in the thesaurus, it takes the best 2 scoring pairing words and then appends them to the end of the query along with the scores, which will then be eventually used to compute the vector scores in the end.

Challenges

None

10a - Text categorization with kNN

Functionality

To do the text categorization, I used Python's Scikit Learn package, which contains various classification methods that allows users to set up their text classification very easily and intuitively. The package contains kNN classification. For the k, I chose k=5. The similarity measure chosen was the *Euclidean distance* measure. The measure was chosen by passing it as a keyword parameter in the K Neighbours Classifier python class. (*KNeighborsClassifier(n_neighbors=5, metric="euclidean")*). After setting up the classifier, I took the training set (Reuters articles with topics) to have it learn what topics go with which words in the full text of the article. Once the labels have been learnt, we applied the labels onto the unlabeled articles. Finally, I saved all the newly labelled docs into their own json file called *set_of_updated_reuters_articles_knn.json* and then afterwards, combined the newly labelled articles with the old labelled articles in their own json file called *updated_corpus_reuters_knn.json*. The 2nd json file would be the one used for retrieval from the UI.

Challenges

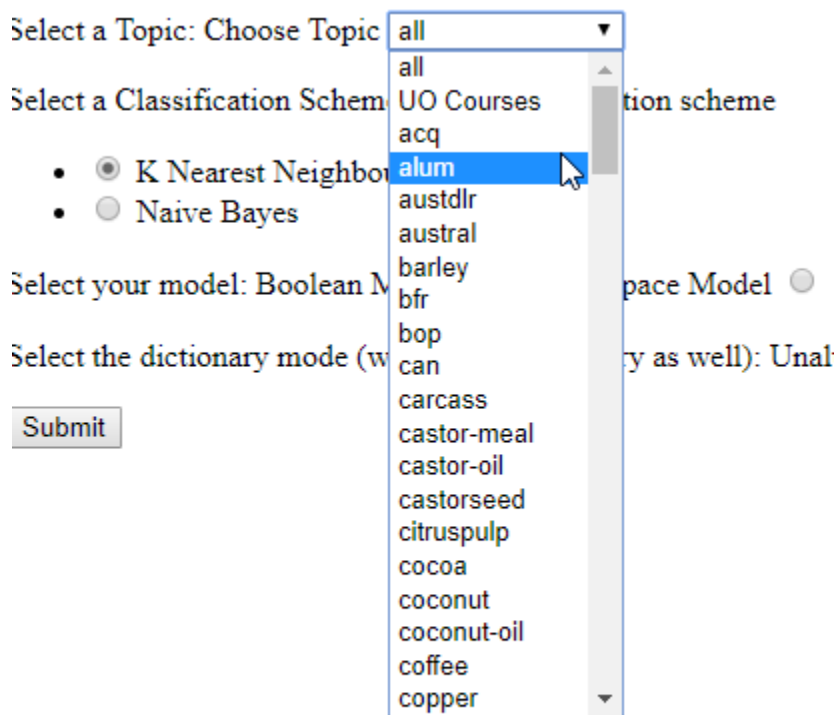
The biggest challenge was handling the multiple topic labels on the articles. Some of the articles came with multiple different topics with them. After struggling with implementing multi labelled versions of kNN classification, I decided to scrap that and just focus on a single topic per

article. So in the preprocessing of the Reuters articles to the corpus, the articles were only assigned a singular topic instead, their first listed topic. This made the classification process a lot more straightforward.

10b - Topic Restriction

Functionality

In the UI, there is a topic dropdown that allows the user to select their preferred topic. The dropdown contains all the possible topics that a Reuters article can have, which was taken from the *all-topics-strings.lc.txt*. As a side note, the documents for the UO courses previously have been give the topic “UO courses” for this final version of the search just for consistencies sake. When the user makes a search with a selected topic, only the documents in the corpuses with the chosen topic will be displayed in the search results screen.



The screenshot shows a web application interface with several form elements. A dropdown menu is open, displaying a list of topics. The topics listed are: all, UO Courses, acq, alum, austdlr, austral, barley, bfr, bop, can, carcass, castor-meal, castor-oil, castorseed, citruspulp, cocoa, coconut, coconut-oil, coffee, and copper. The 'alum' option is currently selected and highlighted in blue. To the left of the dropdown, there are labels for 'Select a Topic: Choose Topic', 'Select a Classification Scheme', 'Select your model: Boolean Model', and 'Select the dictionary mode (with or without space Model)'. Below these labels are radio buttons for 'K Nearest Neighbor' and 'Naive Bayes', and a radio button for 'Boolean Model'. A 'Submit' button is located at the bottom left of the form.

Select a Topic: Choose Topic

Select a Classification Scheme

☒ K Nearest Neighbor

☐ Naive Bayes

Select your model: Boolean Model

Select the dictionary mode (with or without space Model): Unaltered

Submit

Challenges

None

Reuters Collection

Parsing through the Reuters Collection was a straightforward task. Since the articles were in an XML/HTML-like format, it made it pretty easy to utilize the available beautiful soup package I previously used for pre-processing the Ottawa U CSI courses. Only major differences were the structure and names of the tags. Once I familiarized myself with that, I used the same methods to extract the info from the xml-like tags as I did previously. This process was relatively quick as well. I managed to pre-process all the Reuters articles with ease.

There weren't any execution time issues in the system. However, due to the sheer amount of Reuters articles to process, I decided to store a lot of important structures (ex. Dictionary, Inverted Index, Bigram Models, Corpuses, etc.) in JSON files that can be loaded by other modules easily. What took the most amount of time was the building of these json files. For instance, the Thesaurus previously mentioned before took a significant amount of time and space to build out properly. As said before, I had to make concessions by limiting the amount of words in the thesaurus down to the top 2000 words. The bigram model JSON file also took a lot of time to create as well.

The Dictionary itself only took around 2 minutes to build out. To make processing faster, I cleaned out any words that contain digits in them, any words less than 3 characters long, any words that are just punctuation. As with the Vanilla system, the dictionary is split into many modes {unaltered, stopwords removed, stemmed, normalized, fully altered}. They represent the operations done to the words, with each mode containing all the words but with that operation done on it.

Overall, working with the Reuters Collection wasn't difficult per se. The XML-like structure made it straightforward to pre-process with the use of the beautiful soup package. However, the sheer volume of the Reuters collection made the time needed for the creation of necessary JSON files very long. Not difficult but takes a long time.

Search Results Screenshots

shareholder AND security

Final System

Boolean Search

Query: shareholder AND security

GENCORP <GY> REJECTS UNSOLICITED TAKEOVER BID

GenCorp Inc said its board urged shareholders to reject the hostile unsolicited 100 dls a share tender offer made March 18 by General Acquisition Inc, an affiliate of Wagner and Brown and AFG Industries INC. GenCorp also said it is developing a financially superior alternative that would enable shareholders to benefit from the full value of the company.
topic: acq

WORMALD TO ACQUIRE STAKE IN HOLMES PROTECTION

Fire protection and security group Wormald International Ltd <WOIA S> said British-based <Holmes Protection Group Inc> has agreed to issue it with 6.15 mln common shares of one U.S.
topic: acq

FIRST CITY <FCY> SELLS YALE NUCLEONIC UNITS

First City Industries Inc said it has reached a definitive agreement to sell its Nucleon and Yale security subsidiaries to Value PLC for 400 mln dls and warrants to purchase two mln Value ordinary shares.
topic: acq

oil AND profit

Final System

Boolean Search

Query: oil AND profit

BASIN CORP <BAS> 4TH QTR LOSS

Oper shr loss eight cts vs profit 20 cts Oper net loss 768,000 vs profit 1,962,000 Revs 49.0 mln vs 43.6 mln 12 mths Oper shr loss 1.41 dls vs profit 96 cts Oper net loss 13.6 mln vs profit 9,305,000 Revs 175.3 mln vs 140.7 mln Note: Oper excludes loss from discontinued operations of 4,676,000 dls or 48 cts a share for year-ago qtr and 7,571,000 dls or 78 cts a share for year-ago 12 mths.
topic: earn

PRECAMBRIAN SHIELD TAKES \$1.3 MLN DLR WRITEDOWN

<Precambrian Shield Resources Ltd>, enfilr reporting a large loss against year-ago profit, said the 1986 loss was mainly due to a \$1,187,000 dlr writedown on its U.S. operations, uneconomic coal and other mineral properties.
topic: earn

GULF RESOURCES AND CHEMICAL CORP <GRE> 4TH QTR

Oper shr profit 34 cts vs loss 53 cts Oper net profit 3,337,000 vs 4,881,000 Revs 32.7 mln vs 49.6 mln Year Oper shr profit 20 cts vs loss 90 cts Oper net profit 2,374,000 vs loss 9,381,000 Revs 126.9 mln vs 140.5 mln NOTES: Operating net excludes loss 6,050,000 dls, or 64 cts a share, vs loss 24,839,000 dls, or 2.61 dls a share, in quarter and loss 6,050,000 dls, or 64 cts a share, vs profit 64,513,000 dls, or 6.27 dls a share, from discontinued operations 1986 loss from discontinued operations includes 6.0 mln dls charge, equal to 64 cts a share, to provide for additional liabilities resulting from the 1981 closure of lead, zinc and silver mining, smelting and refining business 1988 year operating net includes pre-tax gain of 7.3 mln dls, equal to 36 cts a share, from pension plan termination and gain of 7.2 mln dls, or 56 cts a share, from reduction in deferred taxes Effective Jan 1, 1987, company changed oil and gas accounting to successful efforts from full cost, increasing 1986 year net 9.2 mln dls, or 98 cts a share, and increasing 1985 loss 1.1 mln dls, or 43 cts a share
topic: earn

KANEB SERVICES INC <KAN> 4TH QTR LOSS

Oper shr loss 1.05 dls vs profit nine cts Oper net loss 30.5 mln vs profit 3,930,000 Revs 23.9 mln vs 43.6 mln 12 mths Oper shr loss 5.30 dls vs profit 34 cts Oper net loss 155.8 mln vs profit 16.0 mln Revs 113.7 mln vs 178.8 mln Note: Oper excludes loss from discontinued operations of 9,127,000 dls vs 12.4 mln dls for qtr and 28.4 mln dls vs 960,000 dls for 12 mths.
topic: earn

TOTAL STILL EXPECTS 1.5 BILLION FRANC 1986 LOSS

French oil group Total Cie Francaise des Petroles <TFP PA> is still expecting a 1.5 billion franc consolidated net loss, including minority interests, for 1986, after taking account of stock losses of 7.5 billion francs, the company said in a communique after a board meeting here.
topic: earn

ROYAL DUTCH SHELL U.S. EARNINGS SHARPLY LOWER

Royal Dutch Shell Group <RDAS> earnings for 1986 from the U.S. fell sharply because of difficult market conditions, lower crude and gas prices and also due to different accounting methods, Shell chairman Peter Holmes said.
topic: earn

SHELL FRANCAISE RETURNS TO PROFIT IN 1986

Shell Francaise <SFMF PA>, a subsidiary of <Shell Petroleum NV>, returned to the black last year for the first time since 1982, with parent company net profit of 43 mln francs against losses of 968 mln in 1983 and 1.07 billion in 1984.
topic: earn

NEWSCOPE RESOURCES LTD: YEAR LOSS

Shr loss 94 cts vs profit 28 cts Net loss 6,319,337 vs profit 1,702,016 Revs 2,899,513 vs 5,239,106 Note: 1986 net includes 5,250,000 dlr writedown of oil and gas properties.
topic: earn

Canada canola oil

Final System

Vector Space Search

Query: Canada canola oil

FRANCE PROTESTS OVER CANADA'S FISHING BAN

France today protested over Canada's decision to declare its ports off limits to French fishing vessels and bar further fishing by France off Newfoundland.
score: 38.53296691052848
topic: fish

ENERGY HEAVY OILS

The oil price collapse of 1986 put development of a vast petroleum resource -- heavy and extra heavy oils -- on hold.
score: 36.6108772527274
topic: crude

ENERGY FOREIGN INVESTORS

Lured by the weakening dollar and the conviction that oil prices are poised for a rebound, European energy companies are buying up cheap U.S. oil and gas reserves to replenish their supplies, oil industry analysts said.
score: 29.649617329237876
topic: crude

DIVISION SEEN ON HOW TO HELP U.S. OIL INDUSTRY

The U.S. Congress and the oil industry are deeply divided on ways the government should assist the industry, hurt by the sharp fall in oil prices, and the subsequent growth in oil imports, industry analysts said.
score: 27.4123796281211
topic: crude

ENERGY CALIFORNIA OIL PRODUCTION

Drilling for oil in California is likely to continue at last year's sharply reduced levels despite recent gains in crude oil prices, industry analysts said.
score: 25.425862507459787
topic: crude

CRUDE OIL PRICES UP AS STOCKS, OUTPUT FALL

U.S. crude oil prices rose above 18 dls a barrel this week and industry analysts said the price could rise another dollar as inventories fall.
score: 23.071117356489752
topic: crude

CANADA OIL INDUSTRY SET FOR RECOVERY - ANALYSTS

Firmer crude oil prices, government aid packages and corporate spending cuts will help Canada's oil industry recover from last year's sharp downturn, industry analysts said.
score: 23.19877106751251
topic: crude

European banks stockholders

Final System

Vector Space Search

Query: European banks stockholders

BRITISH BANKS RESIST SIGNING MEXICO PACKAGE

British banks are resisting signing a 76 billion dlr rescheduling package for Mexico in a last ditch effort to get all participants to contribute equally to a new 7.7 billion dlr loan contained in the package.
score: 32.917874378108124
topic: loan

ECONOMIC SPOTLIGHT - FOREIGN BANKS IN GERMANY

Foreign banks in West Germany are reassessing their positions after a number of changes in the markets and the regulatory environment that have dampened a once-optimistic enthusiasm for expansion here.
score: 27.228773258233984
topic: money-fx

CANADA BANKS COULD SEE PRESSURE ON BRAZIL LOANS

Canada's major banks will likely face stiff pressure to declare their Brazilian loans non-performing if, as expected, major U.S. banks take similar action after the end of their first quarter tomorrow, analysts said.
score: 25.60481397424225
topic: loan

JAPANESE BANKS EXPAND HONG KONG PRESENCE

At a time when Britain is threatening to revoke the license of Japanese banks in retaliation for restrictive trade practices, Hong Kong is rolling out the welcome mat to Japan.
score: 22.246862881437723
topic: loan

JAPANESE BANKS EXPAND HONG KONG PRESENCE

At a time when Britain is threatening to revoke the license of Japanese banks in retaliation for restrictive trade practices, Hong Kong is rolling out the welcome mat to Japan.
score: 22.234996208143723
topic: loan

MEXICO TO SIGN LOAN AMID CALLS FOR CHANGE

Mexico's 7.7 billion dlr loan package will be signed in New York tomorrow amid increasing calls from both creditors and debtors for a streamlining of the tortuous process of raising such jumbo loans, bankers said.
score: 21.9878221733748
topic: loan

U.S. URGES BANKS TO WEIGH PHILIPPINE DEBT PLAN

The U.S. is urging reluctant commercial banks to seriously consider accepting a novel Philippine proposal for paying its interest bill and believes the innovation is fully consistent with its Third World debt strategy, a Reagan administration official said.
score: 21.652597995551478
topic: interest

U.S. corn market

Final System

Vector Space Search

Query: U.S. corn market

U.S. CORN ACREAGE SEEN NEAR RECORD LOW

U.S. corn acreage this year is likely to drop to the lowest level since the unsurpassed acreage reductions of the 1983 PER year and could rank as one of the lowest corn plantings in the United States in sixty years, Agriculture Department officials said.
score: 39.203644328341263
topic: grain

NUMEROUS FACTORS SAID POINT TO USSR CORN BUYING

A greater than anticipated need, competitive prices and political motivations could be sparking Soviet interest in U.S. corn, industry and government officials said.
score: 36.51748330779173
topic: grain

U.S. CONSERVATION FIGURES SEEN NEUTRAL-BEARISH

U.S. Agriculture Department (USDA) figures for highly-erodible land enrolled into the Conservation Reserve Program were regarded by most grain analysts as neutral to bearish, although some said a full state-by-state breakdown would be needed to assess the full price impact.
score: 26.419267898108126
topic: grain

U.S. CORN GROWERS BLAST CANADA CORN RULING

Canada's ruling in favor of a duty on U.S. corn was a keen disappointment to the National Corn Growers Association and has set a dangerous precedent for other nations to follow, said Mike Hall, lobbyist for the association.
score: 22.6665662443923
topic: grain

U.S. SUGAR POLICY MAY SELF-DESTRUCT, CONGRESSMAN

A leading U.S. farm-state Congressman, Jerry Huckaby, D-La., warned he will press next year for legislation to control domestic production of sweeteners, perhaps including corn sweeteners, if the industry fails to voluntarily halt output increases this year.
score: 22.262042599894242
topic: sugar

U.S. CORN MARKET SKEWED BY SOVIET BUYING

Recent purchases of U.S. corn by the Soviet Union have skewed the domestic cash market by increasing the price difference between the premium price paid at the Gulf export point and interior levels, cash grain dealers said.
score: 21.04023856630797
topic: grain

NO YIELD DAMAGE YET IN U.S. CORN, BEANS - USDA

The U.S. corn and soybean crops are in mostly good condition and have not suffered any yield deterioration from recent hot, dry weather, Agriculture Department and private weather and crop analysts said.
score: 20.315608323990016
topic: grain

Bigram Language Model Behaviour

Coffee

Final System

CSI4107 Final Search

Search Field

coffee |

coffee value
coffee sources
coffee good
coffee promotion
coffee brasilia
coffee act
coffee role
coffee either
coffee harvest
coffee price

☐ Stemmed ☒ Normalized ☐

Stock

Final System

CSI4107 Final Search

Search Field

stock

stock institutional
stock around
stock scrap
stock electricite
stock weighted
stock launched
stock market
stock securities
stock company
stock operating

☐ Stemmed ☒ Normalized ☐

Oil

Final System

CSI4107 Final Search

Search Field

oil

oil thousand
oil budget
oil late
oil ☐
oil papers
oil quoted
oil confirmed
oil certainly
oil demand
oil said

☐ Stemmed ☐ Normalized ☐

Death

Final System

CSI4107 Final Search

Search Field

death |

death said
death terms
death increasing
death todd
death eight
death reuter
death majority
death hatay
death tom

☐ Stemmed ☐ 1

Submit

Design

Final System

CSI4107 Final Search

Search Field

design |

design stat.ol
design electronics
design non-impact
design operation
design nasa
design upper
design robert
design receive
design aspect
design sections

Stemm

Stemm

Query Expansion

For Query Expansion, for each word in the original query, it takes the top 2 most similar words from the thesaurus and appends them to the end of the query. The expanded query does not get shown in the website, but it does show in the command line console used to run the website.

Coffee

```
[('quotas', 0.18064516129032257), ('bags', 0.11451612903225807)]  
New Query: coffee (1) quotas (0.18064516129032257) bags (0.11451612903225807)  
127.0.0.1 - - [11/Apr/2019 09:44:50] "POST / HTTP/1.1" 200 -
```

Stock

```
[('common', 0.19403636363636365), ('shares', 0.1794844928751048)]  
New Query: stock (1) common (0.19403636363636365) shares (0.1794844928751048)
```

Oil

```
[('crude', 0.1482059282371295), ('prices', 0.14775682704811444)]  
New Query: oil (1) crude (0.1482059282371295) prices (0.14775682704811444)  
127.0.0.1 - - [11/Apr/2019 09:57:31] "POST / HTTP/1.1" 200 -
```

Death

```
New Query: death (1)  
127.0.0.1 - - [11/Apr/2019 09:57:54] "POST / HTTP/1.1" 200 -
```

Due to the method I used to build out the dictionary by taking the top 2000 words based on # of appearances, due to the size and time limitations, “death” was not one of the words that made it to the thesaurus. Hence, its not expanded.

Design

```
[('engineering', 0.036290322580645164), ('standard', 0.03404255319148936)]  
New Query: design (1) engineering (0.036290322580645164) standard (0.03404255319148936)  
127.0.0.1 - - [11/Apr/2019 10:02:02] "POST / HTTP/1.1" 200 -
```

Topic Classification

For topic classification with kNN, I used 5 as my k value. These documents can be found in the *set_of_updated_reuters_articles_knn.json*

Topic: ACQ

```
{
  "doc_id": "reut2-000.sgm-article #2",
  "title": "STANDARD OIL <SRD> TO FORM FINANCIAL UNIT",
  "fulltext": "Standard Oil Co and BP North America\nInc said they plan to form a venture to manage the money market\nborrowing and investment activities of both companies.\n  BP North America is a subsidiary of British Petroleum Co\nPlc <BP>, which also owns a 55 pct interest in Standard Oil.\n  The venture will be called BP/Standard Financial Trading\nand will be operated by Standard Oil under the oversight of a\njoint management committee.\n\n Reuter\n\u00003",
  "excerpt": "Standard Oil Co and BP North America\nInc said they plan to form a venture to manage the money market\nborrowing and investment activities of both companies.",
  "topic": "acq"
},
{
  "doc_id": "reut2-000.sgm-article #3",
  "title": "TEXAS COMMERCE BANCSHARES <TCB> FILES PLAN",
  "fulltext": "Texas Commerce Bancshares Inc's Texas\nCommerce Bank-Houston said it filed an application with the\nComptroller of the Currency in an effort to create the largest\nbanking network in Harris County.\n  The bank said the network would link 31 banks having\n13.5 billion dlrs in assets and 7.5 billion dlrs in deposits.\n\n Reuter\n\u00003",
  "excerpt": "Texas Commerce Bancshares Inc's Texas\nCommerce Bank-Houston said it filed an application with the\nComptroller of the Currency in an effort to create the largest\nbanking network in Harris County.",
  "topic": "acq"
}
```

Topic: Earn

```
{  
  "doc_id": "reut2-000.sgm-article #4",  
  "title": "TALKING POINT/BANKAMERICA <BAC> EQUITY OFFER",  
  "fulltext": "BankAmerica Corp is not under\npressure to act quickly on its proposed equity  
offering and\nwould do well to delay it because of the stock's recent poor\nperformance, banking  
analysts said.\n  Some analysts said they have recommended BankAmerica delay\nits up to one-  
billion-dlr equity offering, which has yet to be\napproved by the Securities and Exchange  
Commission.\n  BankAmerica stock fell this week, <Full text condensed for space>  
  "excerpt": "BankAmerica Corp is not under\npressure to act quickly on its proposed equity  
offering and\nwould do well to delay it because of the stock's recent poor\nperformance, banking  
analysts said.",  
  "topic": "earn"  
}
```

Topic: Crude

```
{  
  "doc_id": "reut2-000.sgm-article #15",  
  "title": "NATIONAL INTERGROUP<NII> TO OFFER PERMIAN UNITS",  
  "fulltext": "National Intergroup Inc said it plans\nto file a registration statement with the  
securities and\nexchange commission for an offering of cumulative convertible\npreferred  
partnership units in Permian Partners L.P.\n  The Permian Partners L.P. was recently formed by  
National\nIntergroup to continue to business of Permian Corp, acquired by\nthe company in  
1985.\n  The company said Permian will continue to manage the\nbusiness as a general partner,  
retaining a 35 pct stake in the\npartnership in the form of common and general  
partnership\nunits.\n  It did not say how many units would be offered or what the\nprice would  
be.\nReuter\n\u00003",  
  "excerpt": "National Intergroup Inc said it plans\nto file a registration statement with the  
securities and\nexchange commission for an offering of cumulative convertible\npreferred  
partnership units in Permian Partners L.P.",  
  "topic": "crude"  
}
```

Topic: Trade

```
{  
  "doc_id": "reut2-000.sgm-article #32",  
  "title": "SENATORS INTRODUCE EXPORT LICENSING REFORM BILL",  
  "fulltext": "Sens. Alan Cranston (D-Cal.) and\nDaniel Evans (R-Wash.) said they introduced  
export licensing\nreform legislation that could save U.S. companies hundreds of\nthousands of  
dollars annually.\n  \"Our emphasis is two-fold: Decontrol and de-license items\nwhere such  
actions will not endanger our national security, and\neliminate the Department of Defense's de  
facto veto authority\nover the licensing process,\" Cranston said.\n  \"Our reforms should reduce  
licensing requirements by 65  to\n70 pct,\" he told reporters. \"I am convinced that a  
more\nrational...licensing process will boost exports.\"\n  U.S. export controls are intended to  
deny Eastern bloc\ncountries access to technology that could further their\nmilitary capabilities.\n  \"By refocusing our control resources on higher levels of\ntechnology, technology that is truly  
critical, we will do a\nbetter job of preventing diversion of critical technology to\nour adversaries  
while promoting more exports,\" Cranston said.\n  \"We cannot expect to continue to play a  
leading role in new\ntechnology development in the future if we unduly restrict the\nactivities of  
U.S. firms in the world market-place,\" Evans told\nreporters.\nReuter\n\u0003",  
  "excerpt": "Sens.",  
  "topic": "trade"  
}
```

Looking at the articles, I find that the label results ended up being very accurate to the full text. kNN is a great approach when we have many classes we can use as labels. In our case here, the topics are our classes and there are many topics we can choose from. This makes kNN a great approach to choose for the classification, just by the sheer number of topics available.

Additional Modules

Optional Module - Text categorization with Naive Bayes

Functionality

The Naïve Bayes classifier was implemented nearly identically to the kNN classifier. We use the same Scikit Learn package, which also contains Naïve Bayes functionality. The process function follows the same structure and the kNN version. Setting up the training set of articles with labels, setting up test set of unlabeled articles, setting up the Naïve Bayes function (in my case, used Gaussian function), and then labeling and saving the articles. As with the kNN, I saved all the newly labelled docs into their own json file called *set_of_updated_reuters_articles_nb.json* and then afterwards, combined the newly labelled articles with the old labelled articles in their own json file called *updated_corpus_reuters_nb.json*. The 2nd json file would be the one used for retrieval from the UI.

I allow the user to choose between classification schemes as well. We have a set of articles that were classified using kNN and another set with Naïve Bayes. The user can switch between either using a radio button switch.

Select a Classification Scheme: Select classification scheme

- ☒ K Nearest Neighbours
- ☐ Naive Bayes

Challenges

None

Examples

These examples can be found in the *set_of_updated_reuters_articles_nb.json*

Topic: Earn

{

"doc_id": "reut2-016.sgm-article #879",

"title": "TODD <TOD> DEBT MAY BE DOWNGRADED BY MOODY'S",

"fulltext": "Moody's Investors Service Inc said it\nmay downgrade Todd Shipyards Corp's 110 mln dlrs of debt.\n It cited Tood's report of significant and unanticipated\nlosses on a commercial contract, and continuing uncertainty\nover the U.S. Navy's DDG-51 Destroyer program, which Moody's\nntermed an important business for Todd's viability.\n Moody's said it would assess the company's future financing\nflexibility in light of current negotiations of existing credit\nand loan agreements. Todd carries B-2 senior subordinated notes\ndue 1996 and B-3 3.08 dlr convertible exchangeable preferred\nstock.\n Reuter\n\u0003",

"excerpt": "Moody's Investors Service Inc said it\nmay downgrade Todd Shipyards Corp's 110 mln dlrs of debt.",

"topic": "earn"

},

{

"doc_id": "reut2-016.sgm-article #880",

"title": "CALNY <CLNY.O> NAMES NEW PRESIDENT",

"fulltext": "Calny Inc said M. Hal Taylor\nhas been named poresident and chief operating officer,\nsucceeding Marvin Hart, who had been serving as president on an\ninterim basis and remains chairman and chief executive.\n The company said Taylor, who joined the company in January,\nhas also been named to the board.\n Reuter\n\u0003",

"excerpt": "Calny Inc said M. Hal Taylor\nhas been named poresident and chief operating officer,\nsucceeding Marvin Hart, who had been serving as president on an\ninterim basis and remains chairman and chief executive.",

"topic": "earn"

}

Source Links for Code

Postfix Notation Code:

<http://interactivepython.org/runestone/static/pythonds/BasicDS/InfixPrefixandPostfixExpressions.html>

Python set update shorthand:

https://python-reference.readthedocs.io/en/latest/docs/sets/op_update.html

Regex code for counting # of occurrences of word:

<https://stackoverflow.com/questions/17268958/finding-occurrences-of-a-word-in-a-string-in-python-3>

Regex for checking if word occurs in piece of text:

<https://stackoverflow.com/questions/5319922/python-check-if-word-is-in-a-string>

Abstract Classes in Python:

https://www.python-course.eu/python3_abstract_classes.php

Web scraping in Python:

<https://medium.freecodecamp.org/how-to-scrape-websites-with-python-and-beautifulsoup-5946935d93fe>

Flask Set-up:

<https://blog.miguelgrinberg.com/post/the-flask-mega-tutorial-part-i-hello-world>

Flask Jinja2 Templates:

<https://blog.miguelgrinberg.com/post/the-flask-mega-tutorial-part-ii-templates>

Flask Web forms:

<https://blog.miguelgrinberg.com/post/the-flask-mega-tutorial-part-iii-web-forms>

Bigram Models:

<https://nlpforhackers.io/language-models/>

Autocomplete Code:

<https://stackoverflow.com/questions/34704997/jquery-autocomplete-in-flask>

K Nearest Neighbour Implementation:

<https://appliedmachinelearning.blog/2018/01/18/conventional-approach-to-text-classification-clustering-using-k-nearest-neighbor-k-means-python-implementation/>

Naïve Bayes Sci Kit Learn Documentation

https://scikit-learn.org/stable/modules/naive_bayes.html