CapstoneProject2_Springboard

Portfolio Project

Problem Statement

Classify articles into meaningful categories irrespective of any Discipline to fulfill business needs or improve the User Experience while browsing content of the files online.

Clients for Topic Modeling

Digitization of Podcasts has led to multitude of archive collections Online, Users would like to play the podcasts of their choice. What if there is a way Podcasts are Organized better to thoroughly enjoy without any hassle of going through the complete list of episodes. By applying various Machine Learning techniques Podcasts can be arranged in the order of relevance based on the coverage of different Topics, thus giving us understanding of how similar and dissimilar the podcasts are.

Data Acquisition

Econtalk is a famous podcast listened by people across the world, classification of these Podcasts using Topic Modeling on the transcripts creates an opportunity for the listeners to tune in with their interests. Podcasts listeners can be recommended and save lot of time in exploring, catering to their needs and customizing the articles based on their Interests.

This is a link to the Archive of Econtalk Podcasts. Web scraping of the transcripts into text format provides an opportunity for Clustering of Topics. www.econlib.org/econtalk-favorite-by-date

Web scraping - Web scraping is a wonderful technique to extract content from websites and create data for Topic Modeling yourself. Beautiful Soup is a Python package used for web scraping to parse HTML and XML documents. It creates a parse tree that can be used to extract data from HTML.

EconTalk website hosts archive of Podcasts dating back until 2006 which aggregate close to 500 documents.Beautiful Soup enables to parse the HTML content and extract all the URL's of the podcasts. Once the URL's are extracted, web content of each podcast can be extracted by navigating to appropriate URL via Beautiful Soup and further split each document to have more than thousand documents for Topic Modeling.

These Documents are further split into two half’s based on the size of each document to have around 1000 documents and this creates a way to estimate the performance of a model by looking out for whether the two half’s belong to the same mixture of Topics after classification.

Data Pre-processing - Gensim provides a package to Preprocess the Text content before applying the machine Learning Algorithms.

- **Tokenization**: Split the text into sentences and the sentences into words. Lowercase the words and remove punctuation.

- Words that have fewer than 3 characters are removed.

- All the **stopwords** are removed.

- Words are **lemmatized** — words in third person are changed to first person and verbs in past and future tenses are changed into present.

```
["0:33Intro. [Recording date: June 16, 2016.] Russ: Now this is an ambitious book that tries to
 analyze where America has been, where it is now, where it might be going. I found it very
 thought-provoking. Let's start with where America is now. You argue that both Republicans and
 Democrats suffer from a nostalgia for the past. Explain. Guest: Well, that’s right. So, the book
 begins from the intense frustration that is overwhelming our public life now in America.
```

As you can see above the raw display of document straight away from the Archives after Webscrapping. Below is the display of Tokens from the first Document after Preprocessing and ready to apply Machine Learning Algorithm.

```
['intro', 'record', 'date', 'june', 'russ', 'ambitious', 'book', 'try', 'analyze', 'america', 'go', 'thin   k',
'provoke', 'start', 'america', 'argue', 'republicans', 'democrats', 'suffer', 'nostalgia', 'past',
'explain', 'guest', 'right', 'book', 'begin', 'intense', 'frustration', 'overwhelm', 'public', 'life',
'america']
```

Frequency count of Words - Gensim doc2bow provides a method to count the frequency of words in each of the documents for analysis in the later stage.

```
Word 5 ("address") appears 1 time. Word
16 ("appeal") appears 6 time. Word 19
("approach") appears 8 time. Word 24
("articulate") appears 4 time. Word 25
("ask") appears 1 time.
```

Topic Modeling and Analysis

We can Identify similarity and dissimilarity of Podcasts from the vast collection of archives in Econtalk based on the Semantic Structure of Topics covered in the Text documents. We shall use C_V measure of coherence to evaluate the performance of Models. C_V is based on a sliding window, a one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity.

This coherence measure retrieves cooccurrence counts for the given words using a sliding window and the window size 110. The counts are used to calculate the NPMI of every top word to every other top word, thus, resulting in a set of vectors—one for every top word. The one-set segmentation of the top words leads to the calculation of the similarity between every top word vector and the sum of all top word vectors. As similarity measure the cosines is used. The coherence is the arithmetic mean of these similarities. (Note that this was the best coherence measure in our evaluation.)

LDA with Term Frequencey(Bag of Words) – LDAMulticore Model is trained with Bag of Words to look out for association of words with each Topics in respective Documents. The Coherence score for c_v is 0. 28 which is relatively low compared to TFIDF transformation.

LDA with TFIDF Transformation – TF-IDF calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. Words with high TF-IDF numbers imply a strong relationship with the document they appear in, suggesting that if that word were to appear in a query, the document could be of interest to the user. LDAMulticore Model is trained after transformation of Words using TFIDF vectorization, the Coherence is relatively better than previous score with c_v of 0.40.

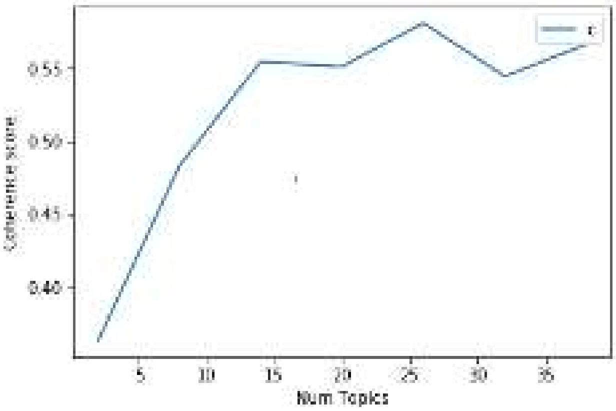
Model	Iterations	Perplexity	Coherence Score(C_V)
LDA (Term Frequency)	50	-8.8225	0.2893
LDA (Term Frequency)	1000	-8.6904	0.2853
LDA (TFIDF)	50	-17.0504	0.3824
LDA (TFIDF)	1000	-16.2631	0.4068
LDA Mallet (Term Frequency)	50	-11.6712	0.5077

Multiple Iterations are performed with different models to look out for best performing Model in terms of Coherence score. Each of these Iterations are chosen to evaluate how they perform at High and low values relatively, which enables us to pursue better classification of Topics.

LDA Mallet with Term Frequency (Bag of Words) – Gensim library provides a wrapper around Mallet to implement the algorithm with the Bag of Words. LDA Mallet performs much better than the original LDA model based on the Coherence Score and Topic Distribution. Optimal Model is determined by iterating with different number of Topics and comparing the Coherence Score. Below table provides the performance of LDA mallet over multiple Topics.

As we can see that the peak performance is obtained when the Number of Topics is 26 with coherence score of 0.581 which is relatively high compare to other models so far.

Model	Number of Topics	Iterations	Coherence
LDA Mallet	2	1000	0.3628
LDA Mallet	8	1000	0.4829
LDA Mallet	14	1000	0.5545
LDA Mallet	20	1000	0.5511
LDA Mallet	26	1000	0.581
LDA Mallet	32	1000	0.5444
LDA Mallet	40	1000	0.5662



We can explore the distribution of Terms in Topics to understand how well the Topics are meaningful and classified. The terms clustered in each of the Topics make sense and are related to a common topic rather than being random. Below is the display of some of the Topic determined by LDA Mallet Model.

Topic No	Topic Terms
6	0.011*"europe" + 0.010*"british" + 0.010*"history" + 0.009*"century" + ''0.009*"king" + 0.009*"england" + 0.008*"germany" + 0.008*"church" + ''0.008*"country" + 0.007*"slave"
10	0.041*"food" + 0.016*"farm" + 0.011*"farmers" + 0.010*"land" + 0.009*"fish"+ 0.009*"corn" + 0.008*"plant" + 0.006*"cook" + 0.006*"local" + ''0.006*"agriculture"'
12	0.019*"technology" + 0.019*"machine" + 0.012*"brain" + 0.010*"knowledge" + ''0.008*"science" + 0.007*"intelligence" + 0.007*"information" + ''0.006*"imagine" + 0.006*"humans" + 0.006*"drive"'
15	0.014*"regulation" + 0.014*"competition" + 0.013*"innovation" + ''0.013*"license" + 0.012*"private" + 0.011*"industry" + 0.008*"sector" + ''0.008*"patent" + 0.007*"service" + 0.007*"regulatory"'
8	'0.015*"body" + 0.013*"sugar" + 0.012*"kidney" + 0.011*"diet" + ' '0.010*"disease" + 0.009*"heart" + 0.009*"weight" + 0.008*"obesity" + ''0.008*"exercise" + 0.007*"eat"'
18	0.037*"smith" + 0.023*"moral" + 0.013*"adam" + 0.010*"society" + ''0.010*"theory" + 0.009*"wealth" + 0.008*"social" + 0.007*"hume" + ''0.006*"sentiments" + 0.006*"nature"'
23	0.056*"bank" + 0.020*"financial" + 0.018*"risk" + 0.014*"crisis" + ' '0.013*"loan" + 0.012*"fund" + 0.012*"capital" + 0.010*"debt" + ' '0.010*"mortgage" + 0.010*"assets"

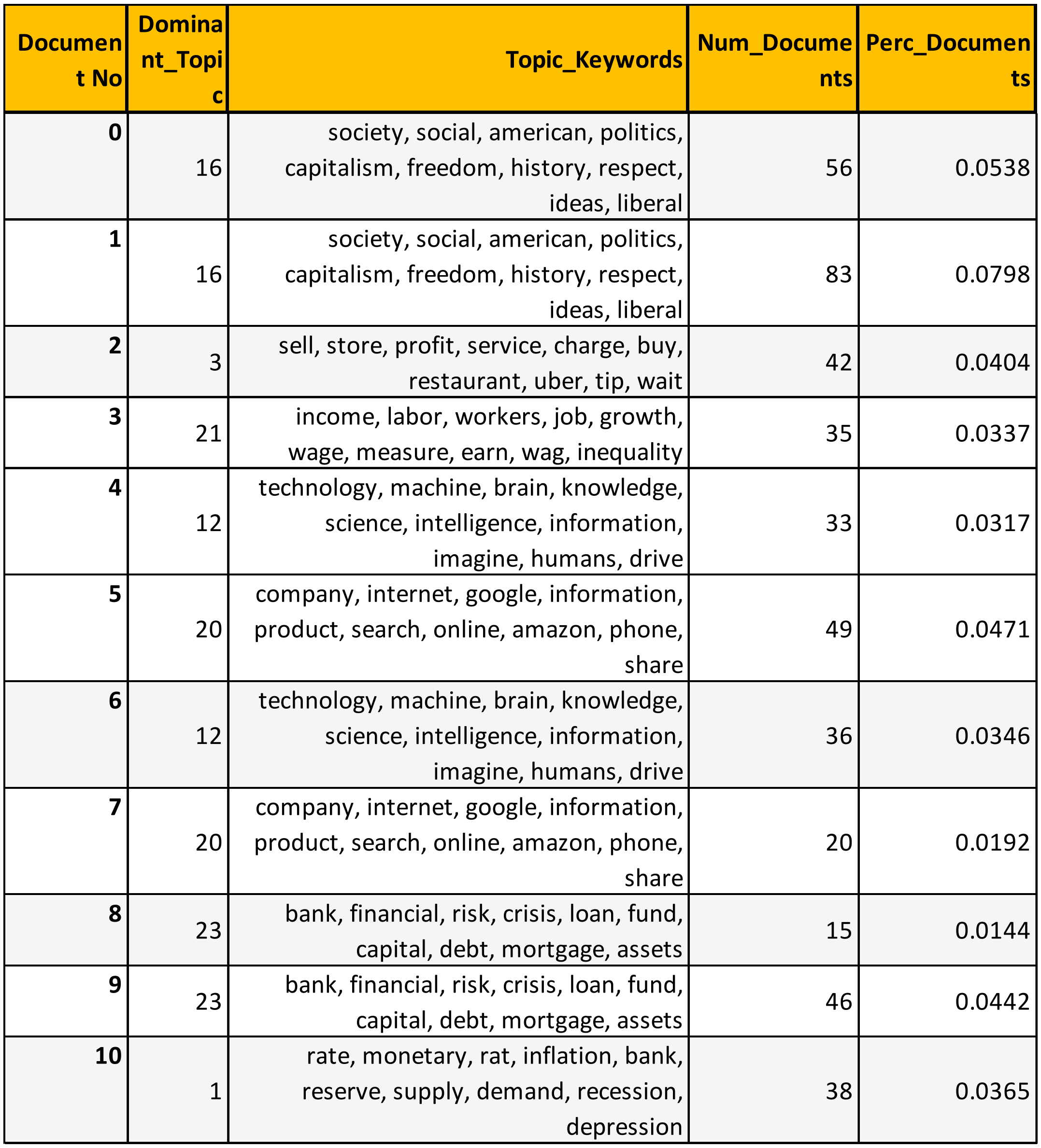
Conclusion

Topic Modeling is performed on multiple documents to identify similarity based on the Semantic Structure of Documents by using TF and TFIDF after preprocessing the Documents by removing Stopwords and Lemmatizing the vocabulary to maintain uniform terms across documents. Optimal Model has been determined by exploring Hyperparameters and different Iterations and evaluating the Coherence Score of C_V . LDA Mallet Model performs relatively better than other Models by classifying the Terms into 26 Topics.

Here we can relate the Dominant Topic Associated with each of the Document by gazing through the terms in Transcripts and evaluate how well the Documents are classified by Mallet model.

Document No	Dominant Topic	Topic Percentage Contribution	Keywords	Document Terms
0	16	0.513	society, social, american, politics, capitalism, freedom, history, respect, ideas, liberal	intro, record, date, june, russ, ambitious, book, try, analyze, america, go, think, provoke, start, america, argue, republicans, democrats, suffer, nostalgia, past, explain, guest, right, book, begin, intense, frustration, overwhelm, public, life, america, evident, politics, ways, things, want, understand, kind, frustration, source, listen, express, politics, listen, political, life, drench, kind, nostalgia, widely, share, sense, america, express, slogans, hear, politics, lose, grind, fast, peak, americans, remember, exactly, peak, vary, republicans, democrats, people, leave, peak, century, america, moment, enormous, confidence, large, institutions, government, labor, business, solve, kinds, challenge, alongside, liberalize, culture, people, right, miss, time, york, time, april, ask, donald, trump, america, great
1	16	0.5307	society, social, american, politics, capitalism, freedom, history, respect, ideas, liberal	russ, talk, minute, feedback, loop, problems, welfare, state, get, bigger, private, ways, help, get, smaller, charity, family, incentives, marry, smaller, need, spouse, thrive, economically, extent, problems, self, create, economist, look, underlie, economic, force, market, force, cultural, economic, dynamism, economy, mainly, good, thing, unleash, policy, responses, think, problem, worse, create, demand, think, need, education, failure, failure, aren, root, problem, symptom, guest, think, root, maybe, slightly, different, force, operate, society, virtue, free, society, force, encourage, think, isolate, individuals, discourage, think, understand, human, flourish, kind, social, order, think, instead, flourish, individual, truth, society, wealthy, successful, dangerous, especially, come, people, live, go, people, need
2	3	0.1932	sell, store, profit, service, charge, buy, restaurant, uber, tip, wait	intro, record, date, june, russ, topic, today, inequality, base, thoughtful, piece, write, reaction, article, york, time, york, time, article, bemoan, fact, norwegian, cruise, line, create, separate, status, cruise, elite, guests, turn, better, treatment, passengers, fancier, cabin, better, amenities, extent, physically, isolate, mass, know, right, word, cruise, ship, passengers, ironies, article, article, criticize, physical, isolation, different, class, interest, question, concern, socially, point, author, misunderstand, economics, situation, miss, guest, thing, account, seriously, notion, reveal, preferences, huge, indignity, respect, people, cruise, ship, expect, basic, demand, evaporate, elite, group, get, pamper, like, mean, people, boat, guess, plus, regard, modest, amenity, people, little, star, glitter
3	21	0.2115	income, labor, workers, job, growth, wage, measure, earn, wag, inequality	russ, actually, make, clear, try, away, little, fix, cost, point, make, go, rephrase, go, push, general, feel, resentment, point, make, class, versus, coach, easily, see, automobile, fix, phenomenon, little, research, development, know, drive, honda, accord, drive, lexus, drive, ferrari, know, richard, know, guest, love, ferrari, russ, know, live, urban, guest, york, russ, poor, fellow, go, honda, fabulous, line, lexus, remarkably, great, fraction, price, think, remark, market, force, give, lower, market, dregs, incredibly, high, quality, products, cellphone, trip, country, airplane, fact, seat, little, cramp, major, issue, want, complainers, go, push, earlier, point, think, motivate, feel, confession, people, earn, money, robert, frank
4	12	0.362	technology, machine, brain, knowledge, science, intelligence, information, imagine, humans, drive	intro, record, date, june, russ, inevitable, great, book, book, read, single, word, succeed, host, time, able, despite, short, time, read, book, savor, vivid, snapshot, present, beautifully, write, speculation, head, provocative, mind, blow, true, go, touch, trend, go, start, chapter, call, argue, look, today, years, feel, like, look, great, hang, fruit, easy, come, stuff, certain, malaise, feel, look, technology, today, people, think, know, fly, cars, twitter, bunch, social, media, optimistic, present, potential, future, guest, read, history, look, go, optimistic, improvements, tend, overcome, problems, introduce, technology, introduce, host, problems, haven, general, think, look, history, optimistic, allow, look, future, russ, kind, internet, thing
5	20	0.3243	company, internet, google, information, product, search, online, amazon, phone, share	russ, okay, point, people, book, human, be, people, screen, course, mean, death, read, point, like, point, people, read, probably, human, history, order, magnitude, read, blow, novels, history, book, read, kinds, things, wonderful, point, observation, differ, look, solve, problems, people, book, versus, people, screen, versus, technology, talk, guest, yeah, general, thesis, people, book, book, kind, fix, finish, precise, immutable, monumental, case, text, foundation, western, civilization, extent, eastern, civilization, thousand, document, country, know, constitution, bible, book, author, root, authority, authority, come, author, kind, orientation, have, cheap, access, book, public, libraries, literacy, read, write, produce, incredible, know, year, explosion, civilization, mean, sort, ways

The below table explains how many documents are related to each other based on the dominant topics , which gives a very good notion to the audience of Podcast to stay in tune with their topics of interest and catch up with the latest trends in Econtalk.



Document No	Dominant_Topic	Topic_Keywords	Num_Documents	Perc_Documents
0	16	society, social, american, politics, capitalism, freedom, history, respect, ideas, liberal	56	0.0538
1	16	society, social, american, politics, capitalism, freedom, history, respect, ideas, liberal	83	0.0798
2	3	sell, store, profit, service, charge, buy, restaurant, uber, tip, wait	42	0.0404
3	21	income, labor, workers, job, growth, wage, measure, earn, wag, inequality	35	0.0337
4	12	technology, machine, brain, knowledge, science, intelligence, information, imagine, humans, drive	33	0.0317
5	20	company, internet, google, information, product, search, online, amazon, phone, share	49	0.0471
6	12	technology, machine, brain, knowledge, science, intelligence, information, imagine, humans, drive	36	0.0346
7	20	company, internet, google, information, product, search, online, amazon, phone, share	20	0.0192
8	23	bank, financial, risk, crisis, loan, fund, capital, debt, mortgage, assets	15	0.0144
9	23	bank, financial, risk, crisis, loan, fund, capital, debt, mortgage, assets	46	0.0442
10	1	rate, monetary, rat, inflation, bank, reserve, supply, demand, recession, depression	38	0.0365