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 hosts archives 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.

Data Pre-processing - Gensim provides 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.
- Words are **stemmed —** words are reduced to their root form.

Below is the display of some Tokens from the first Document after Preprocessing a nd ready to apply Machine Learning Algorithm.

['russ', 'talk', 'minute', 'feedback', 'loop', 'problems', 'welfare', 'state', 'get', 'bigger', 'private', 'way s', 'help', 'get', 'smaller', 'charity', 'family', 'incentives', 'marry', 'smaller']

- 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

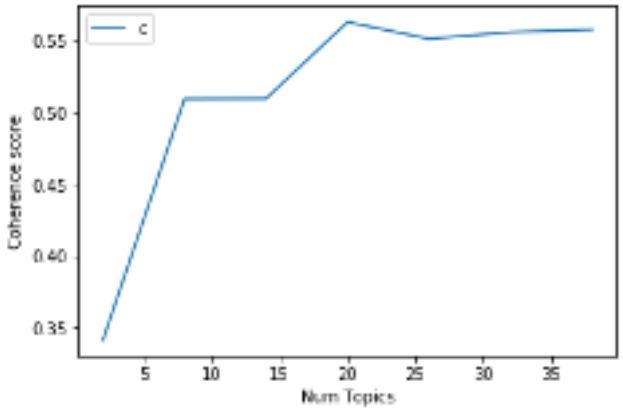
Identify similarity and dissimilarity of Podcasts from the vast collection of archives in Econtalk based on the Semantic Structure of Topics covered in the Text documents.

- **LDA with 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 low with 0. 28.
- **LDA with TFIDF Transformation** – LDAMulticore Model is trained after transformation of Words using TFIDF vectorization, the Coherence is much better than previous score with c_v of 0.40

Model	Iterations	Perplexity	Coherence Score
LDA (Bag of Words)	50	-8.8225	0.2893
LDA (Bag of Words)	1000	-8.6904	0.2853
LDA (TFIDF)	50	-17.0504	0.3824
LDA (TFIDF)	1000	-16.2631	0.4068
LDA Mallet (Bag of Words)	50	-11.6712	0.5077

- **LDA Mallet with 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 evaluation of LDA mallet over multiple Topics.

Model	Number of Topics	Iterations	Coherence Score
LDA Mallet	2	1000	0.3415
LDA Mallet	8	1000	0.5096
LDA Mallet	14	1000	0.5098
LDA Mallet	20	1000	0.5632
LDA Mallet	26	1000	0.5515
LDA Mallet	32	1000	0.5561
LDA Mallet	40	1000	0.558



Conclusion

LDA Mallet performs well in determining the Semantic Structure of the documents when the Number of Topics is 20. We can determine the most dominant Topic in each of the Document from the same Model.

Document No	Dominant Topic	Topic Percentage Contribution	Keywords	Text
0	2	0.3756	hayek, society, ideas, social, theory, keynes,...	[intro, record, date, june, russ, ambitious, b...
1	2	0.3776	hayek, society, ideas, social, theory, keynes,...	[russ, talk, minute, feedback, loop, problems,...
2	14	0.3465	city, cities, drive, cars, york, service, char...	[intro, record, date, june, russ, topic, today...
3	13	0.2576	income, labor, workers, job, growth, wage, dat...	[russ, actually, make, clear, try, away, littl...
4	11	0.497	technology, company, machine, internet, google...	[intro, record, date, june, russ, inevitable, ...
5	11	0.483	technology, company, machine, internet, google...	[russ, okay, point, people, book, human, be, p...
6	11	0.4324	technology, company, machine, internet, google...	[intro, record, date, russ, practical, work, q...
7	11	0.3515	technology, company, machine, internet, google...	[russ, want, twitter, question, mention, book,...
8	18	0.2403	love, listen, experience, remember, attention,...	[intro, record, date, russ, ambrose, bierce, b...
9	18	0.2127	love, listen, experience, remember, attention,...	[russ, want, read, definition, read, definitio..

Most Representative Document for each topic

Topic Number	Topic Percentage Contrib	Keywords	Text
0	0.6306	food, store, farm, farmers, produce, corn, pla...	[intro, record, date, june, russ, roberts, go,...
1	0.6608	sell, profit, license, drug, prison, card, sho...	[russ, start, want, remind, listeners, econtal...
2	0.6233	hayek, society, ideas, social, theory, keynes,...	[intro, record, date, january, hayek, have, go...
3	0.6357	game, sport, team, baseball, players, football...	[intro, record, date, march, topic, find, inte...
4	0.6074	smith, moral, bitcoin, adam, theory, wealth, s...	[russ, roberts, aspect, dennis, rasmussen, int...