10/26/2020

Prepare abstract for UG research symposium; apply by Nov. 1; final presentation ready by Dec. 2

<https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0#:~:text=Focussing%20on%20the%20log%2Dlikelihood,model%20that%20was%20learned%20earlier.&text=The%20concept%20of%20topic%20coherence,topics%20inferred%20by%20a%20model>.

MALLET – other LDA? More

Abstract – 4-5 sentnces

10/19/2020

* Fine-tune preprocessing steps
* Use bigrams / trigrams to interpret the topics from LDA
  + Get just the list of bigrams / trigrams
* Perplexity? Coherence? What are they, what are good values?
  + “how to use perplexity / coherence to find optimal numbers of topics”

10/12/2020

Keep working on relevant code …

10/5/2020

Figure out Github sharing… (analyticsvidhya site; medium: towardsdatascience)

Q2: Do textual complaints reflect existing Issues?

Easiest low hanging fruits: **wordcloud**; dig out **bigrams/trigrams**; table/**visualization** for patterns

* Can use **topic modeling, LDA**, GenSim, (NLTK?) : Needs a bit more research/learning on our part

Q2a, Q2b refinement here…

Q1: What is distribution of complaints? (how many against what company?)

Tables/visualize aggregate numbers for companies… (Navient is big…)

Q1a: Is the pattern in complaints for handling private/federal loans same/different?

Q1b: Is there a change in pattern over time?

9/23/2020

Get Github! Account – send it to you Dan.

Research questions on student loans

* + ~3 narrow research Qs (‘HW’ for next week)
    - Does Navient has the most complaints?
      * Private/federal separately?
    - Do complaints change over time?
      * Words/phrases
      * Categories/issues

Data back to 2012; 2015 on: complaints data; 2017 Apr. 24: stopped taking in consumers disputed data;

To do: keep data set munging / preprocessing, look at some text analytics results and charts to map out further steps.

9/16/2020

* Looking at final product: poster (emailed example from XX);
  + Structure: like academic research paper: sent 3 links for review
* Get some early thoughts on research questions
  + Initial poking around in text data: complaint texts from CFPB; Python pandas/numpy/nltk

9/8/2020

* Pick up where we were. Interest: text analytics. Agreed: let’s keep working on CFPB student load data. (Download till 8/31/2020)

4/16/2020

* We have excel file: ‘complaints each month all’: monthly complaint counts per company and federal/private loans, for 2015-2019; (huge jump in January of 2016)
  + **Needs to look at chart, trend**
* Have sample file to start text analysis
  + Daniel will start looking/understanding it

4/9/2020

* Clean up txt analysis code + send to DL
* Table code is in ‘Exploring data’ py file
* We’ll need more months breakouts for complaint numbers
  + So far narratives only complaints
  + **Add columns for all complaint counts (the whole time period)**
  + **Add columns for private/federal loan (flag: private/non-private) each company that does both have a separate row for them**

3/24/2020

* **Let’s pull out all Navient complaint into a separate .csv file.**
  + See how **category, issue names might change over time**.
* Zip codes: 555XX format; how to visualize / correlate with other info?  John Johnson
* Navient complaint jump: 1/2017; CFPB sues Navient on 1/18/2017;
  + **Why large number of complaints only posted from that month on?**
* Pick early Dec 2016 – May 2017; aggregate complaints per company per months. Sort it by number of complaints in January.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| * Company | * Dec | * **Jan** |  |  |  | * May |
| * Navient |  | * **#** |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

3/18/2020

* **Remove duplicates for all complaints (find code segment + send to)**
* Non-federal and Private categories: one replaces the other, but practically the same
  + **Did CFPB say anything about category coding change?**
* Multiple issues: one could replace the other
  + Let’s see time chart for the different issues to see if this is true.
* Navient: do they handle both federal/private loans? –> check it!!!
* DL: complaint number graphing by state
  + Add zip codes; (tables; sorted)
  + Complaints per zip code (relative to population, demographics)

Work:

* Tried to remove duplicates based on column ‘Consumer complaint narrative’ but nothing dropped.
* Combined Non-federal and Private sub-products
  + No duplicates by ‘Consumer complaint narrative’
  + Graphed the data (‘updated\_private\_data’)
* Sorted by count of complaints per zip code in excel file
* Navient has both federal/private loans

3/11/2020

* All of the following descriptives below: 1st: ALL Student loan complaints; 2nd: just the ones with narrative (subset of ALL)
* Complaint chart by month; big jump (01/2017**: why?; line chart by sub-product?)**
* Sub-product: federal (greatest), not federal, private
* **Chart by issues (11) (too many sub-issues(48) to chart, but see numbers in table)**
* Companies (Private loan: 159; Federal: 202; non-federal: 182) **Created sorted table by number of complaints per company (per category)**
* Table of number of complaints for [sub-product / issue (3 x 11)], [sub-product / sub-issue (3 x 48)],
* Research question possibilities:
  + Is there a difference between the 3 sub-product types in:
    - Number of complaints

Work:

(need help with strategies of how to find cause of influx of complaints)

Need to clean my code next time

* + Not all sub-product data has the same start date (private: 4/2017, federal: 3/2016, non-federal: 1/2015)
  + Each sub-product has a large increase towards the beginning of 2017
  + Created a table of the overall data count of each issue (also created tables for each sub-product)
  + Created a table of each company showing the number of complaints it has
  + Navient has the most issues by far. I created a table with a count of each issue for Navient alone. The issues *Dealing with my lender or service* and *Dealing with your lender or servicer* has close to the same counts of just under 4,000. Also, there are lots of issues with *Struggling to repay your loan* and *Can’t repay my loan.* This tells us that with Navient people were having issues paying off their loans and were having issues working with the lenders or servicers.

2/26/2020

Student Loan data from CFPB

Packages: nltk; genism (glove: alg; avoid for now!)

* 25195 rows, all have narratives from 1/1/15
* 3 sub-products
* 11 issues
* 48 sub-issues
*  data table, chart?
*  time dimension for complaints (types)
*  per company complaints (Navient?)

<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/>

Text analytics

* Lower case
* Remove extra characters
* Remove stop words (probably other strings as well; ‘xxxxx’
* Tokenize / stemming

2/12/2020

Look at:

<https://www.consumerfinance.gov/>

<https://www.consumerfinance.gov/data-research/consumer-complaints/>

getting acquainted:

* Download data (chunk?)
* Chunking it? (different dimension, variable, category)
* Python code / excel

See if there is interesting bits directions…