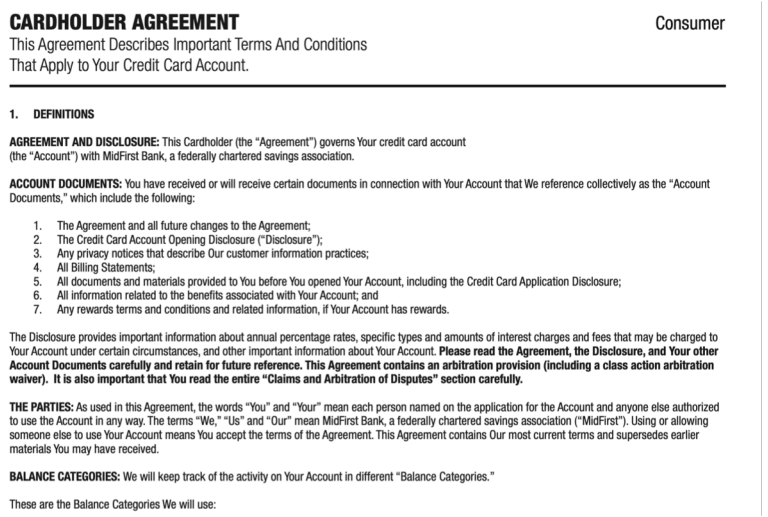
Analyzing Credit Card Agreements

Patrick Cullinane
Project repo: https://github.com/cullinap/contract_reader

Problem Statement

- ➔ For the average person legal documents are hard to read and understand.
- ➔ According to one study it would take 76 days each year to read all the privacy statements you sign up for.[1]
- ➔ Most people do not have the time to read terms & conditions
- ➔ Are there ways to apply natural language processing to group and return relevant results?



CARDHOLDER AGREEMENT Consumer

This Agreement Describes Important Terms And Conditions
That Apply to Your Credit Card Account.

1. DEFINITIONS

AGREEMENT AND DISCLOSURE: This Cardholder (the "Agreement") governs Your credit card account (the "Account") with MidFirst Bank, a federally chartered savings association.

ACCOUNT DOCUMENTS: You have received or will receive certain documents in connection with Your Account that We reference collectively as the "Account Documents," which include the following:

1. The Agreement and all future changes to the Agreement;
2. The Credit Card Account Opening Disclosure ("Disclosure");
3. Any privacy notices that describe Our customer information practices;
4. All Billing Statements;
5. All documents and materials provided to You before You opened Your Account, including the Credit Card Application Disclosure;
6. All information related to the benefits associated with Your Account; and
7. Any rewards terms and conditions and related information, if Your Account has rewards.

The Disclosure provides important information about annual percentage rates, specific types and amounts of interest charges and fees that may be charged to Your Account under certain circumstances, and other important information about Your Account. **Please read the Agreement, the Disclosure, and Your other Account Documents carefully and retain for future reference. This Agreement contains an arbitration provision (including a class action arbitration waiver). It is also important that You read the entire "Claims and Arbitration of Disputes" section carefully.**

THE PARTIES: As used in this Agreement, the words "You" and "Your" mean each person named on the application for the Account and anyone else authorized to use the Account in any way. The terms "We," "Us" and "Our" mean MidFirst Bank, a federally chartered savings association ("MidFirst"). Using or allowing someone else to use Your Account means You accept the terms of the Agreement. This Agreement contains Our most current terms and supersedes earlier materials You may have received.

BALANCE CATEGORIES: We will keep track of the activity on Your Account in different "Balance Categories."

These are the Balance Categories We will use:

Figure 1.: Example cardholder agreement for a credit card

1. I/S: A Journal of Law and Policy for the Information Society, vol. 4, no. 3 (2008), 543-568.

Process Overview

- ➔ Credit card agreements from the Consumer Finance Protection Bureau
- ➔ Convert PDF's to text data, combine into a giant corpus, preprocess, and apply weighting algorithms.
- ➔ "Break" document into "terms" & "documents"
- ➔ Find the optimal number of topics (k) based on a mean topic coherence metric.

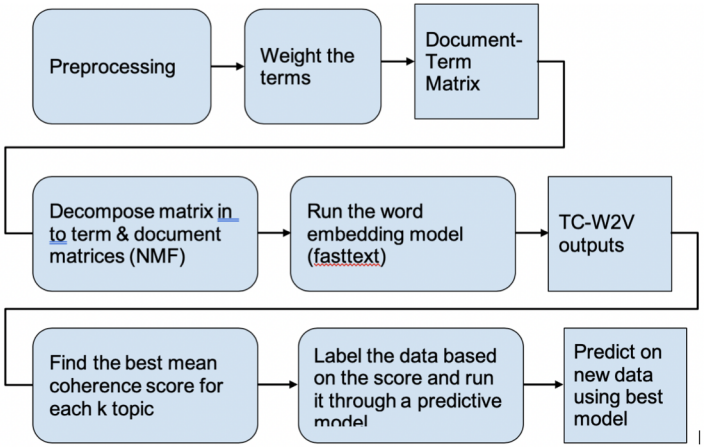


Figure 2.: General overview of the process

Dataset from : https://www.consumerfinance.gov/credit-cards/agreements/

Relevant Documents and Terms

- After taking the text from PDFs, doing some simple preprocessing, and weighting important words we decompose the data set using a technique called non-negative matrix factorization.
- We determine an optimal topic number (k)
- The input is a matrix with documents on one axes and terms on the other.
- The output is two matrices: terms x k and documents x k.

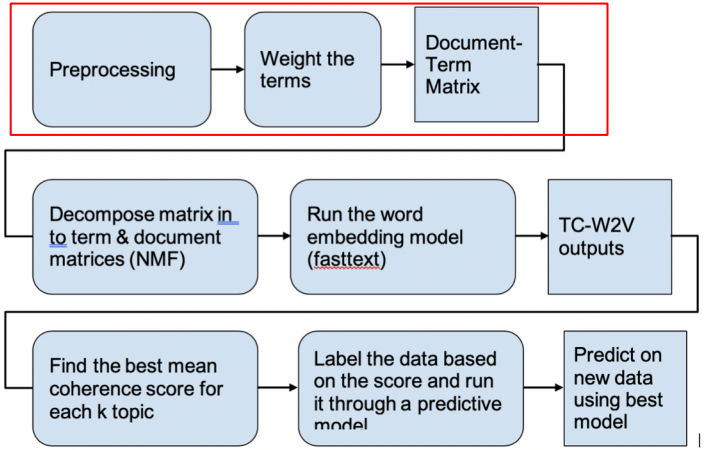
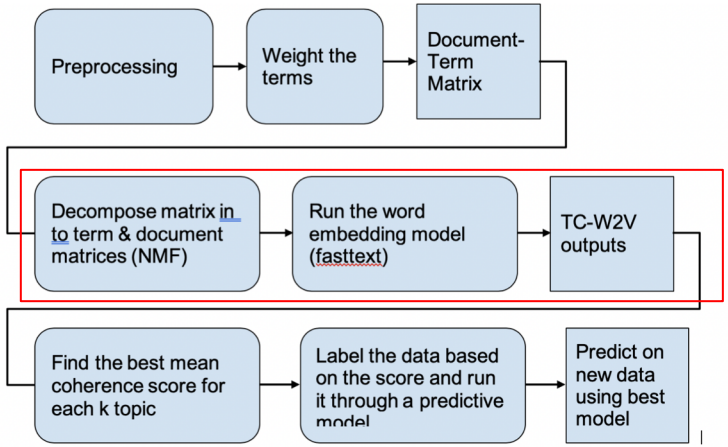


Figure 2.: General overview of the process

Determine the optimal number of topics

- ➔ Run the data through a word embedding model.
- ➔ Returns a score of how “close” a word will be to other words. Ex. Soviet is likely to to be found near Russia and Stalin as opposed to watermelon.
- ➔ Find the top terms for each topic by comparing our NMF model v. TFIDF model for each topic
- ➔ Calculates similarity score for each word pair in topic and returns the average score for each topic.



Examine the Output

➔ Return the top terms and documents for each of our topics based on the optimal topic number.

```
#=#=#=#=#=#=#=#=#=#=#=#=#=#=#=#=#=#=#=#=#=#=#=#=#=#=#=#=#=#=
top terms for topic 1
['apr', 'rate', 'prime', 'prime_rate', 'base', 'market_base', 'vari_market', 'apr_vari', 'base_prime', 'vari']
==================================================
top ground truth label: This APR will vary with the market based on the Prime Rate.
==================================================
top documents for each topic 1
                                         ground_truth
1   This APR will not vary with the Market based \...
2   Annual Percentage Rate (APR) for \nPurchases 1...
3   20.24% \nThis APR will vary with the market ba...
4   This APR will vary with the market based on th...
5   Your APR will vary with the market based on th...
6   18.49%\nThis APR will vary with the market bas...
7   14.99%\nThis APR will vary with the market bas...
8   5.50%    \n     This APR will vary with the ma...
9   19.99% - 30.99% Your APR will vary with the ma...
10  These APRs will vary with the market based on ...
11  These APRs will vary with the market based on ...
12  This APR will vary with the market based on th...
13  26.24% \nThis APR will vary with the market ba...
14  26.24% \nThis APR will vary with the market ba...
15  26.24% \nThis APR will vary with the market ba...
16  This APR will vary with the market based on th...
17  27.99%  This APR will vary with the market bas...
18  This APR will vary with the Market based on \n...
19  APR will vary with the market based on the Pri...
```

Figure 2.: General overview of the process

Label the Dataset and Create a Model

- ➔ Return 50 “documents” for each topic, and label them with a name.
  - ◆ We chose: 'payment', 'rate', 'agreement',' error', 'fee', 'purchases', 'interest', 'credit_tips'
- ➔ Run 450 examples (50x9) through a classification model.
- ➔ Best performing model was Support Vector Machines.

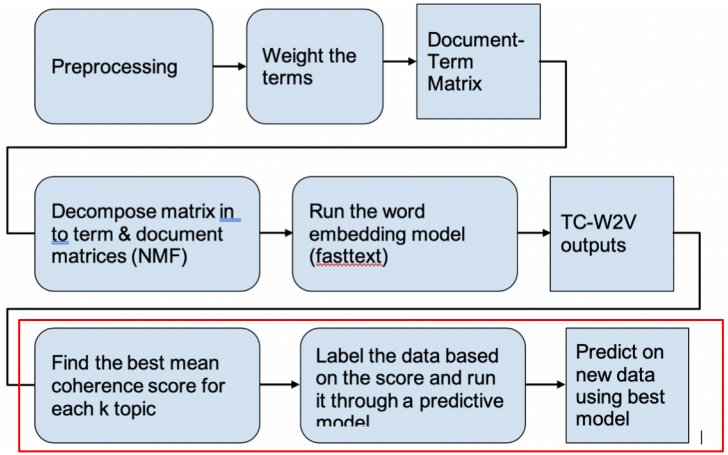


Figure 2.: General overview of the process

Feed a Credit Card Agreement into the Model

- In this example we feed a document into our model.
  - It returns the categories that it found in the document.
  - The label ‘rate’ gives us the APR, when you open your account, the APR for balance transfers, and the APR for Cash Advances.
- Now we know quickly that our APR will be higher if we take a cash advance. Good to know if you are shopping around for credit cards.

```
#create a Read class and
from utils.reader import Read

df = Read()
```

```
your document has the following categories
agreement
purchases
payment
error
rate
credit_tips
fee
```

```
df.section(label='rate')
```

```
Your APR will be 20.24% when you open your account.
This APR will vary with the market based on the Prime Rate.
APR for Balance Transfers Your APR will be 20.24% when you open your account.
This APR will vary with the market based on the Prime Rate.
APR for Cash Advances 24.99%. This APR will vary with the market based on the Prime Rate.
```

Conclusion and Future Work

- ➔ It appears this model did a good job classifying snippets within the text that are labelled as credit tips, purchases, errors, and agreement.
- ➔ There is some crossover between rate and interest as these categories possibly explain the same information.
- ➔ The next step in this project is twofold:
  - ◆ Further refine the topics to be more specific to the theme provided (i.e. adjust k value, rename labels)
  - ◆ Subdivide topics further to develop more granularity on specific topics.
    - For example: “Rate” subtopics can include: APR, cash advance, prime rate, etc..