Algorithmic Comment Processing

Members*

Gayani Perera, Liliana Cruz-Lopez, Minsu Yeom, Pranjal Bajaj

* In alphabetical order

Industry Mentors

Junghoon Woo, Director Data Scientist,
Data & Analytics (The Lighthouse), KPMG LLP, US

Viral Chawda, Principal, Innovation & Enterprise Solutions (I&ES), Lighthouse and Global lead,
AI & Analytics for Government & Infrastructure, KPMG LLP, US

Data Science Institute Mentor

Sining Chen, Lecturer, Columbia University

OUR GOAL

Automate the Identification and Summarisation of Sections in PDF Documents

Roadmap

1. Problem Statement
2. Module 1: PDF Ingestion
3. Module 1: Data Preparation
4. Module 1: Modelling
5. Module 2: Section Summarization

Problem Statement: Background

Client: Regulations.gov

Pre-rule

Final Ruling

Problem Statement: Business Impact

Prior to Automation

12-20 Weeks

30 People

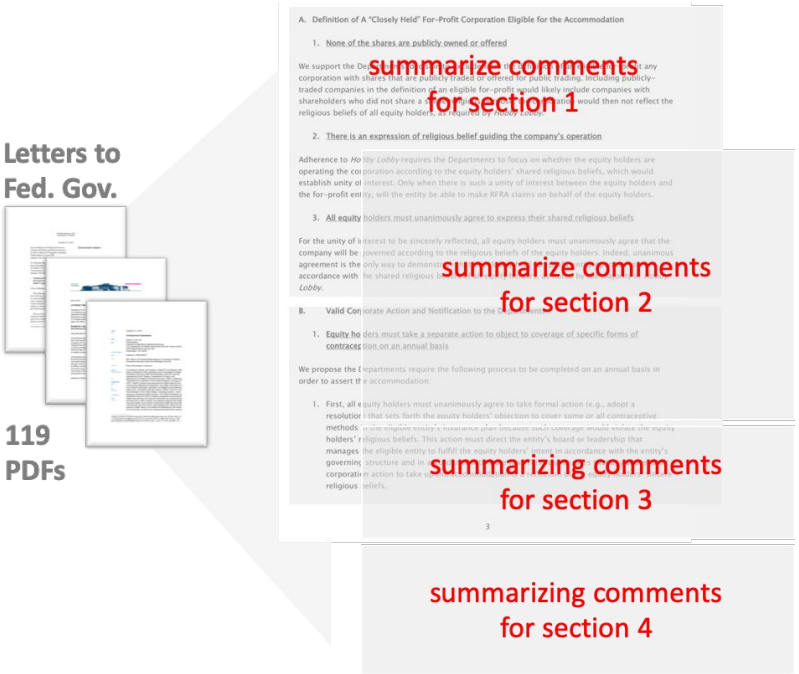
Post Automation

2 People

2 Weeks

Source: KPMG

Problem Statement: Our Solution

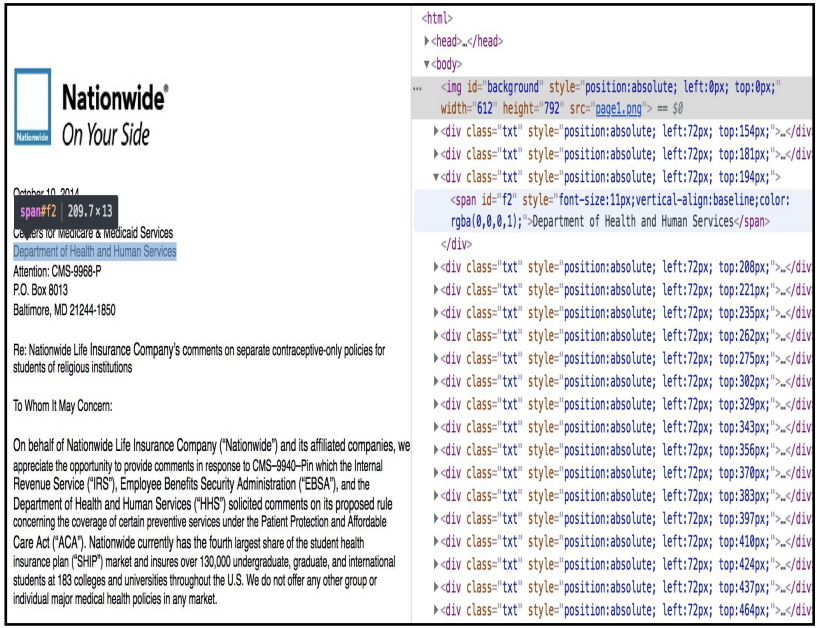


Filename	SectionID	Summary
CMS_2014_0115_0059.pdf	1.0	[students of religious institutions, To Whom It May Concern:, On behalf of Nationwide Life Insurance Company ("Nationwide") and its affiliated companies, we, appreciate the opportunity to provide ...
CMS_2014_0115_0059.pdf	2.0	[significant administrative burden upon all parties., The Department's basic premise that "issuers generally would find that providing such contraceptive, coverage is cost neutral" is in error (Fe...
CMS_2014_0115_0059.pdf	3.0	[automatically enrolled in a contraceptive-only health plan., Students who choose to attend a religious institution of higher learning do so for a reason, and most, of the time, these students str...
CMS_2014_0115_0059.pdf	4.0	[partaking in providing contraceptive coverage to its students if the same SHIP it contracts, with for general student health must also provide contraceptive coverage via individual, policies., In...
CMS_2014_0115_0059.pdf	5.0	[contrary to basic contract law., An issued health insurance policy is a contract between an insurance company and the insured., Contracts are binding and enforceable only when one party extends a...

Module 1

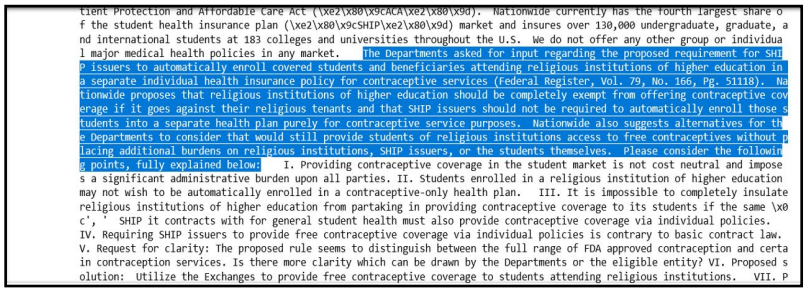
PDF Ingestion: PDF to ?

PDFs to HTMLs

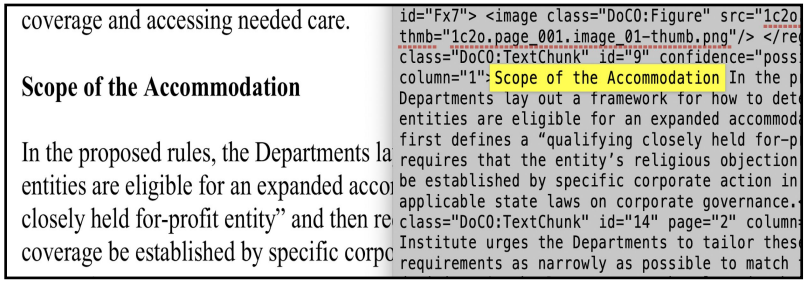


Information extracted from HTMLs led us to build extra features used in our models.

Two other attempts



PDFs to Text
Issue: White spaces only between the paragraphs



PDFs to XML
Issue: A section title appears within a paragraph

Data Science Institute
COLUMBIA UNIVERSITY

KPMG

PDF Ingestion: Can you tell which one is an original PDF?

December 5, 2017

Centers for Medicare & Medicaid Services
Department of Health and Human Services
P.O. Box 8016
Baltimore, MD 21244-8016
Attention: CMS-9940-IFC

national partnership for women & families

Jacobs Institute of Women's Health
THE GEORGE WASHINGTON UNIVERSITY

Submitted electronically at www.regulations.gov

Subject: Interim Final Rule on Religious Exemptions and Accommodations for Coverage of Certain Preventive Services Under the Affordable Care Act [CMS-9940-IFC]

Union of Concerned Scientists
Science for a Healthy Planet and Safer World

The National Partnership for Women & Families, Jacobs Institute of Women's Health, and Union of Concerned Scientists submit the following comments in response to the Interim Final Rules ("the Rules") titled "Moral Exemptions and Accommodations for Coverage of Certain Preventive Services Under the Affordable Care Act"[1] and "Religious Exemptions and Accommodations for Coverage of Certain Preventive Services Under the Affordable Care Act,"[2] published in the Federal Register on October 13, 2017, by the Department of the Treasury, the Department of Labor, and the Department of Health and Human Services ("the Departments").

Our organizations work to ensure that U.S. policy decision-making is fully informed by scientific evidence and the best available data, and that the public has reliable access to independent scientific information and analysis produced and acquired by the federal government. The role of scientific evidence in public health decision-making is imperative, and we oppose any efforts to diminish the role of science in federal policymaking.

Unfortunately, the Rules are a prime example of regulatory decision-making that ignores scientific evidence and the best available data. The Departments' summary of the evidence is arbitrary and cherry-picked. The Departments understate the efficacy and health benefits of contraceptives and overstate the health risks of contraceptives by selectively interpreting data, overlooking well-established evidence, and promoting unfounded doubt. Further, both Rules falsely assert certain types of FDA-approved contraceptive methods to abortifacients.

The Rules thus cause dual harm by undermining women's access to essential preventive health care and undermining the integrity of science in governance. Public health policy should be informed by the best available scientific evidence. Instead, the Departments use false claims about contraception that are contrary to medical and public health evidence, misstate or ignore research, and undermine the agencies' role as a source of accurate health information.

The Departments serve a critical role in collecting and managing important information and data on issues that are vital to the public. In making policy, it is essential that the Departments enhance their credibility on issues of science and evidence, not undermine it. Thus, the

1

December 5, 2017

Centers for Medicare & Medicaid Services
Department of Health and Human Services
P.O. Box 8016
Baltimore, MD 21244-8016
Attention: CMS-9940-IFC

national partnership for women & families

Jacobs Institute of Women's Health
THE GEORGE WASHINGTON UNIVERSITY

Submitted electronically at www.regulations.gov

Subject: Interim Final Rule on Religious Exemptions and Accommodations for Coverage of Certain Preventive Services Under the Affordable Care Act [CMS-9940-IFC]

Union of Concerned Scientists
Science for a Healthy Planet and Safer World

The National Partnership for Women & Families, Jacobs Institute of Women's Health, and Union of Concerned Scientists submit the following comments in response to the Interim Final Rules ("the Rules") titled "Moral Exemptions and Accommodations for Coverage of Certain Preventive Services Under the Affordable Care Act"[1] and "Religious Exemptions and Accommodations for Coverage of Certain Preventive Services Under the Affordable Care Act,"[2] published in the Federal Register on October 13, 2017, by the Department of the Treasury, the Department of Labor, and the Department of Health and Human Services ("the Departments").

Our organizations work to ensure that U.S. policy decision-making is fully informed by scientific evidence and the best available data, and that the public has reliable access to independent scientific information and analysis produced and acquired by the federal government. The role of scientific evidence in public health decision-making is imperative, and we oppose any efforts to diminish the role of science in federal policymaking.

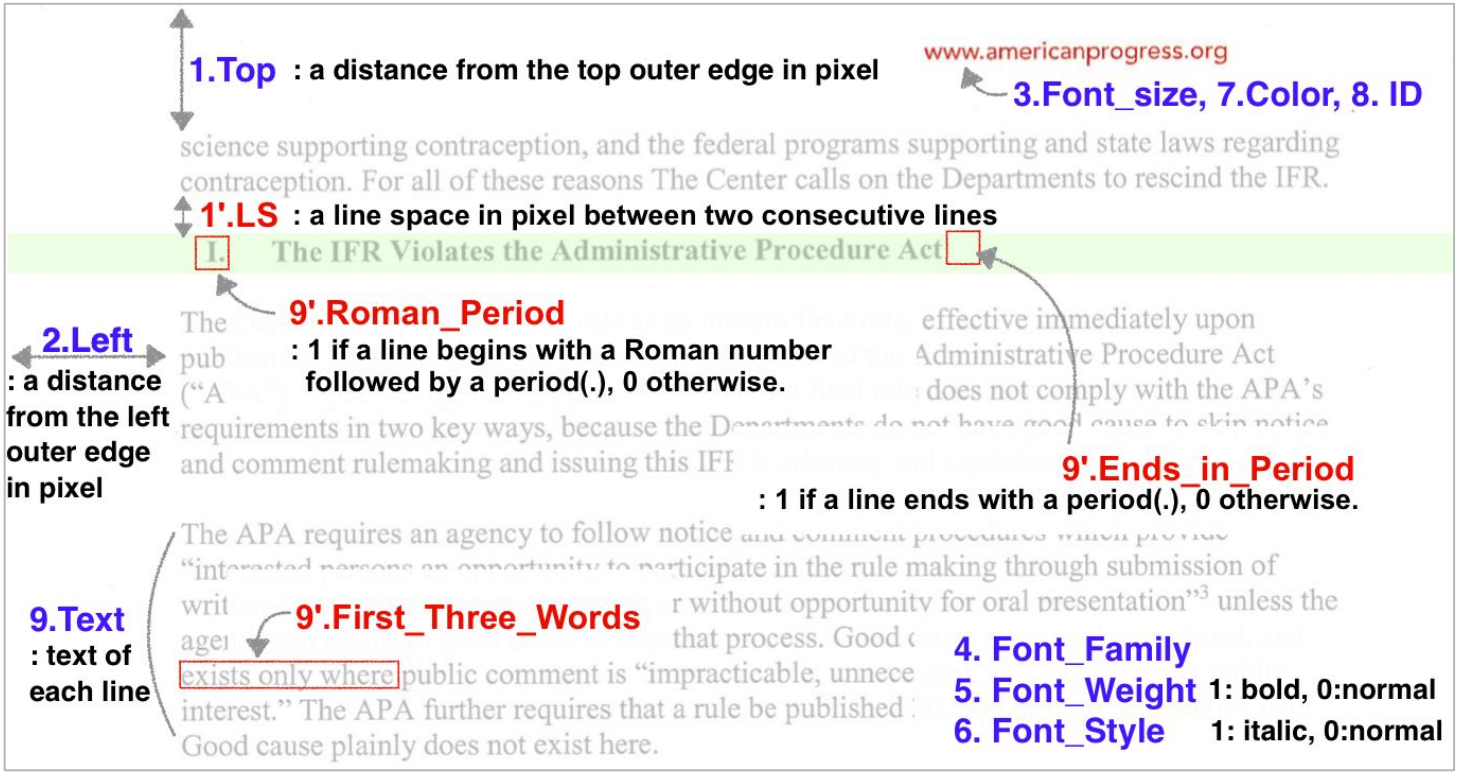
Unfortunately, the Rules are a prime example of regulatory decision-making that ignores scientific evidence and the best available data. The Departments' summary of the evidence is arbitrary and cherry-picked. The Departments understate the efficacy and health benefits of contraceptives and overstate the health risks of contraceptives by selectively interpreting data, overlooking well-established evidence, and promoting unfounded doubt. Further, both Rules falsely assert certain types of FDA-approved contraceptive methods to abortifacients.

The Rules thus cause dual harm by undermining women's access to essential preventive health care and undermining the integrity of science in governance. Public health policy should be informed by the best available scientific evidence. Instead, the Departments use false claims about contraception that are contrary to medical and public health evidence, misstate or ignore research, and undermine the agencies' role as a source of accurate health information.

The Departments serve a critical role in collecting and managing important information and data on issues that are vital to the public. In making policy, it is essential that the Departments enhance their credibility on issues of science and evidence, not undermine it. Thus, the

1

Data Preparation: Feature Engineering

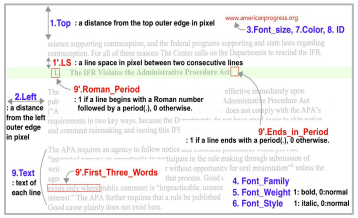


HTML-based features ("raw") in blue. Engineered features from the raw in red.

Data Preparation: Feature Engineering

Category	Feature Name	Description
Binary	Leading_Char_Upper	A line start with a uppercase character
	Leading_Numeral	A line start with Arabic or Roman numeral
	Ends_in_Period	A line ends with a period
	Leading_Number_Period	A line starts with any numeral combination followed by period
	Leading_Char_Period	A line start with any uppercase or lowercase character followed by period
	Leading_Roman_Numeral	A line start with any Roman numeral
	Roman_Period	A line start with Roman numeral followed by period
Numerical	Num_Word	Number of words in the text line
	Num_of_Spec_Char	Number of special characters in the text line
	LS	A line space between previous and current lines.
	Punctuation_Count	Number of punctuations in the text line
	Title_Word_Count	Number of title word counts in the text line
	Upper_Case_Word_Count	Number of uppercase word counts in the text line
	Ratio_of_Title_Word_To_Total	Ratio of the number of title words to all words in the line
Categorical	Document	File Name
Textural	Last_Word	Last word of the text line
	First_Three_Words	First three words of the text line

Data Preparation: HTML to Data frame



LS	font-weight	ratio_of_title_word_to_total	first_3_words	Leading_Char_Period	Num_Words	Class
17.0	0.0	0.000000	belief in publicly	0.0	15.0	0.0
28.0	0.0	0.833333	Re Patient Protection	0.0	12.0	1.0
14.0	0.0	0.090909	assistance in languages	0.0	11.0	0.0
14.0	0.0	0.600000	CVS Health Head	0.0	5.0	0.0
13.0	0.0	0.000000	concerning that the	0.0	18.0	0.0
12.0	0.0	0.071429	cheaper than the	0.0	14.0	0.0
13.0	0.0	0.625000	Supreme Court ruling	0.0	16.0	0.0
13.0	0.0	0.000000	a separate contraceptiveonly	0.0	5.0	0.0
13.0	1.0	0.000000	more clarity regarding	0.0	10.0	0.0
21.0	0.0	0.066667	Finally in the	0.0	15.0	0.0

PDF

Features

Data frame

Data Preparation: Getting Modelling-ready

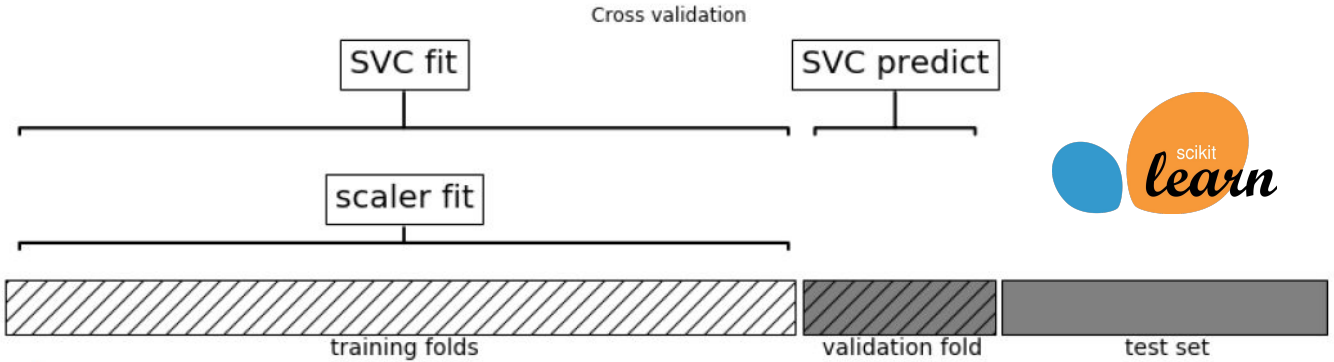
- Treating Missing Data
- OneHotEncoding Categorical Data
- Scaling Continuous Features
- Transforming Text Data: Last Word and First 3 Words
  - One Hot Encoded Representation: CountVectorizer and TfidfVectorizer
  - n_grams: (e.g. “not happy”, “deeply sad”)
  - stop_words (e.g. “a”, “in”)

Modelling: Test-Train splits and Pipelines

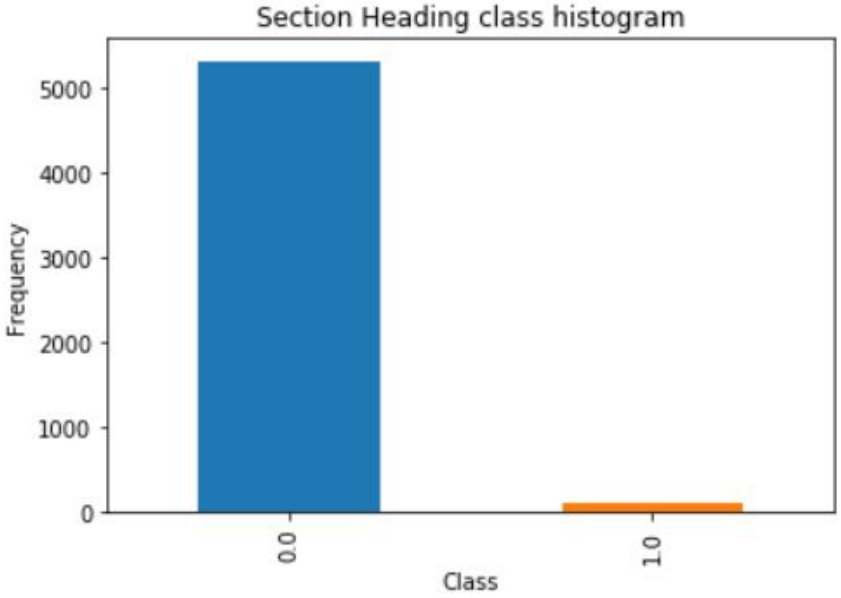
7744 lines coming from 19 documents

- 70% Train
- 30% Test

Scikit-learn pipeline prevents leakage by chaining transformations with cross-validation



Modelling: Class Imbalance and Evaluation Metrics



2.08% of the lines are section titles

- False Negative: Section titles incorrectly identified as a in-text line
- False Positive: in-text line incorrectly identified as a section header
- In our scenario, we cared slightly more about False Negatives.

Modelling: Algorithms

Classification Algorithms:

1. Baseline Model: Logistic Regression
2. Random Forest Classifier
3. XGBoost Classifier

Outlier Detection Algorithms:

1. Isolation Forest: Picks outliers by randomly selecting features
2. Elliptic Envelope: Assume Gaussian Covariance to isolate outliers

Parameter Tuning and Cross-validation

- Grid-search over parameters
- Using a 5-fold cross-validation: Stratified Shuffle Split
- Embedded in a scikit-learn Pipeline

Modelling: Best Results on an Independent Test Set

Random Forest

Max Depth: 50
Number of Trees: 100

Oversampling Minority Class

Empirical Rule
Any line that begins with "RE:" is labelled as a section title

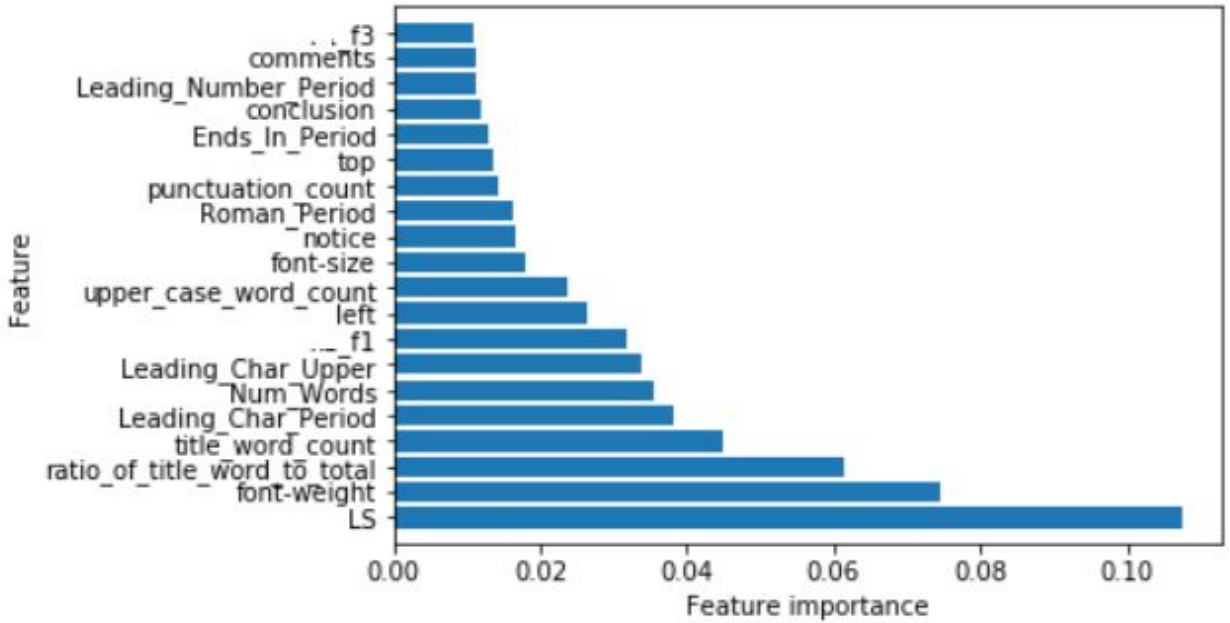
Confusion Matrix

True Negatives: 2,273	False Positives: 10
False Negatives: 4	True Positives: 36

Results Table

Threshold	Precision	Recall	F1 Score	ROC AUC	Accuracy
0.31	0.78	0.90	0.84	0.95	0.99

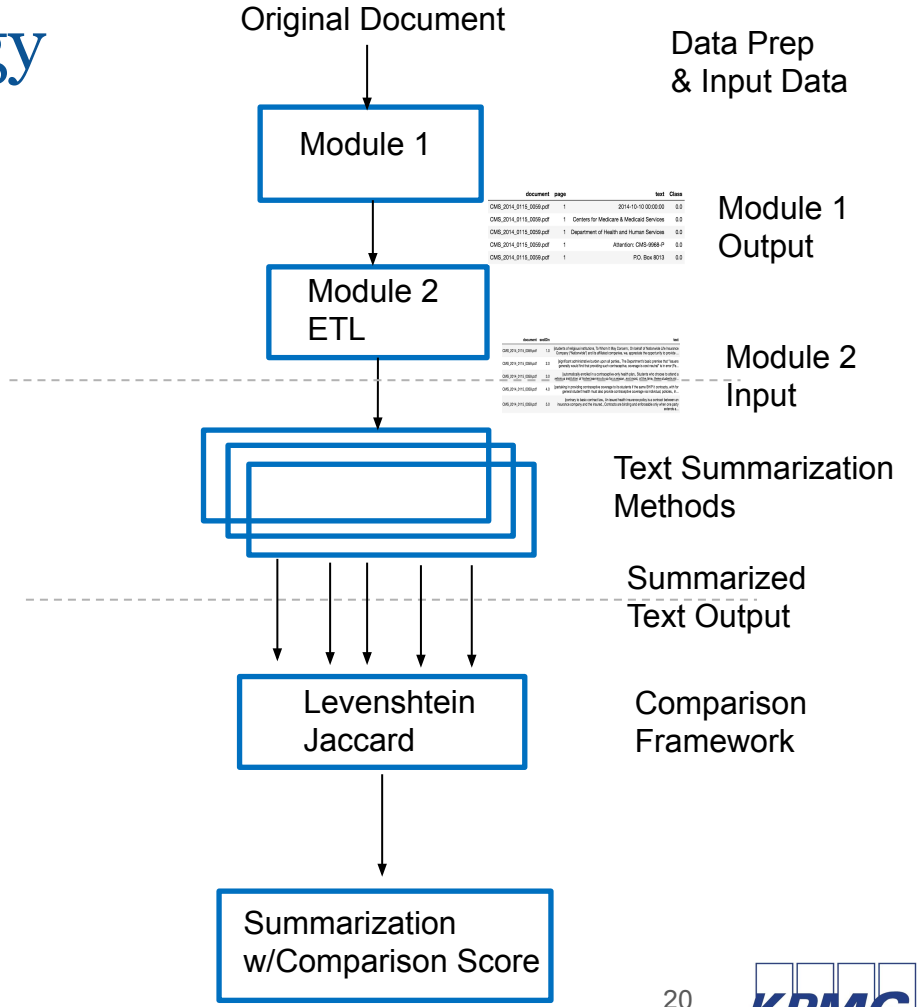
Modelling: Important Features



Module 2

Background and Methodology

- Module 2 Objective: take the intermediate output generated by Module 1 and produce good quality text summarization
- We consider 5 different text summarization techniques that range from simple frequency based to semantic based analysis
- We consider two metrics (Levenshtein distance, Jaccard distance) to compare the output generated by these 5 methods
- Experimental evaluation and comparison of summarization output
- Lessons learned from summarization exploration



How Text Summarization Works?

Broadly two categories of Text Summarization: Extractive and Abstractive

Extractive Summarization: Sentences are ranked based on important part of the sentences. Summarization method chooses top ranked sentences.

Different algorithm and techniques are used to define weights for the sentences and further rank them based on importance and similarity among each other.

Input document → sentences similarity → weight sentences → select sentences with higher rank.

Abstractive Summarization: This method produces summarization that is more human like where important concepts are produced.

This method selects words based on semantic understanding and tries to summarize based on important concepts. Most methods interpret and examine the text using advanced natural language techniques in order to generate a new shorter text that conveys the most critical information.

Input document → understand context → semantics → create own summary.

Extractive Summarization returns top-N sentences as summarized output whereas Abstractive Summarization produces a key set of concepts as summarization based on semantic analysis. The latter is often hard and more complex but more human-like.

Data Preparation for Summarization Step

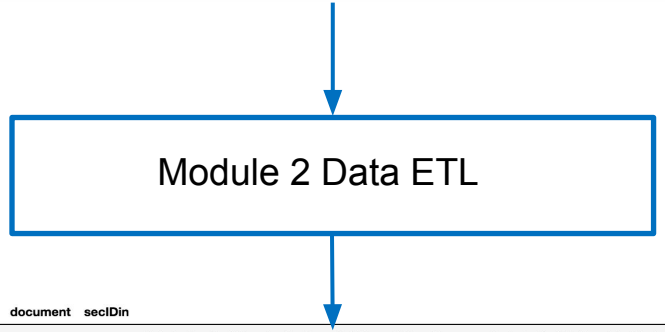
Data Preparation (Module 2 ETL)

- Original document is processed by Module 1 to generate a set of meta tags
- Module 2 ETL utilizes Module 1 Output to generate input data with appropriate features for Text Summarization Methods

Module 1 Output schema

1. **document:** name of the document
2. **page :** page number where each text belongs to
3. **text:** the text from each line is store in this column
4. **Class:** the classification of each line text line

document	page	text	Class
CMS_2014_0115_0059.pdf	1	2014-10-10 00:00:00	0.0
CMS_2014_0115_0059.pdf	1	Centers for Medicare & Medicaid Services	0.0
CMS_2014_0115_0059.pdf	1	Department of Health and Human Services	0.0
CMS_2014_0115_0059.pdf	1	Attention: CMS-9968-P	0.0
CMS_2014_0115_0059.pdf	1	P.O. Box 8013	0.0



Module 2 Input Schema

1. **document :** document name
2. **secIDin:** the section id of a particular text
3. **text:** the text for each section

document	secIDin	text
CMS_2014_0115_0059.pdf	1.0	[students of religious institutions, To Whom It May Concern:, On behalf of Nationwide Life Insurance Company ("Nationwide") and its affiliated companies, we, appreciate the opportunity to provide ...
CMS_2014_0115_0059.pdf	2.0	[significant administrative burden upon all parties., The Department's basic premise that "issuers generally would find that providing such contraceptive, coverage is cost neutral" is in error (Fe...
CMS_2014_0115_0059.pdf	3.0	[automatically enrolled in a contraceptive-only health plan., Students who choose to attend a religious institution of higher learning do so for a reason, and most, of the time, these students str...
CMS_2014_0115_0059.pdf	4.0	[partaking in providing contraceptive coverage to its students if the same SHIP it contracts, with for general student health must also provide contraceptive coverage via individual, policies., In...
CMS_2014_0115_0059.pdf	5.0	[contrary to basic contract law., An issued health insurance policy is a contract between an insurance company and the insured., Contracts are binding and enforceable only when one party extends a...

Summarization Models

	Luhn Model	Lex Rank Model	Tex Rank Model	LSA Model	NLTK
Core Idea	Each sentence is assigned a score based on frequency of occurrence and distance among significant words; next is to extract top-N sentences with top scores.	Sentences are assigned a score based on TF-IDF and creating a graph with edges between similar sentences; PageRank based approach is used to compute rank of each sentence; top-N ranked sentences are extracted.	Similar to LexRank; While LexRank uses cosine similarity of TF-IDF vectors, TextRank uses a measure based on the number of words two sentences have in common.	LSA projects data into a lower dimensional space using SVD; singular vectors can capture and represent word combination patterns; magnitude of singular value indicates importance of the pattern in a document.	Simple text based approach summarization using basic NLP techniques such as word tokenization.
Category	Extractive	Extractive	Extractive	Close to abtractive	Extractive
Frequency based ranking	✓				✓
Graph based ranking		✓	✓		
ML Unsupervised				✓	
Semantic				✓	

Comparing Summarization Quality with Similarity Metrics … cont

How do we know whether summarization is good quality?

- Our hypothesis: If summarization output produced by these methods are ***”very similar”*** to each other, this consensus is an indicator that ***summarization quality may be good***. Conversely, if the output are “highly dissimilar”, the summarization quality is at least is non conclusive.

- We want to experimentally validate if ***“maximal consensus”*** is a good policy of picking good summarization.

- Automated hypothesis testing: We choose two metrics to measure similarity between two strings
    - Levenshtein distance: measures similarity at character level
    - Jaccard distance: measures dissimilarity at word level

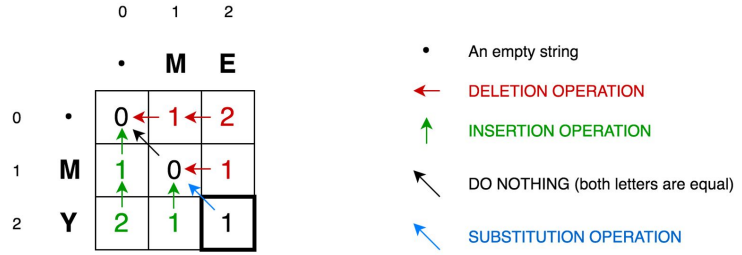
Comparing Summarization Quality with Similarity Metrics

Levenshtein distance: similarity between two strings

Mathematically, the Levenshtein distance between two strings a, b (of length $|a|$ and $|b|$ respectively) is given by $lev_{a,b}(|a|, |b|)$

$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j) + 1 \\ \operatorname{lev}_{a,b}(i,j-1) + 1 \\ \operatorname{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

where $1_{(a_i \neq b_j)}$ is the indicator function equal to 0 when $a_i = b_j$ and equal to 1 otherwise, and $lev_{a,b}(i,j)$ is the distance between the first i characters of a and the first j characters of b



Jaccard distance: dissimilarity between two strings

M_{11} represents the total number of attributes where *A* and *B* both have a value of 1.

M_{01} represents the total number of attributes where the attribute of *A* is 0 and the attribute of *B* is 1.

M_{10} represents the total number of attributes where the attribute of *A* is 1 and the attribute of *B* is 0.

M_{00} represents the total number of attributes where *A* and *B* both have a value of 0.

$$\frac{M_{01} + M_{10}}{M_{01} + M_{10} + M_{11}}$$

		A	
		0	1
B	0	M_{00}	M_{10}
	1	M_{01}	M_{11}

Experiments

Similarity	Dissimilarity	Similarity	Dissimilarity	Similarity	Similarity	Dissimilarity	Similarity	Similarity	Dissimilarity	Similarity	Dissimilarity	Similarity	Dissimilarity	Similarity	Dissimilarity	Similarity	Similarity	Dissimilarity	Similarity	Dissimilarity
Levin score for Lex_Rank and LSA	Jaccard score for Lex_Rank and LSA	Levin score for Lex_Rank and Luhn	Jaccard score for Lex_Rank and Luhn	Levin score for LSA and Luhn	Jaccard score for LSA and Luhn	Jaccard score for LSA and Luhn	Levin score for TextRank and Luhn	Jaccard score for TextRank and Luhn	Jaccard score for TextRank and Luhn	Levin score for LSA and TextRank	Jaccard score for LSA and TextRank	Levin score for TextRank and Lex_Rank	Jaccard score for TextRank and Lex_Rank	Levin score for NLTK and Luhn	Jaccard score for NLTK and Luhn	Levin score for NLTK and Lex_Rank	Jaccard score for NLTK and Lex_Rank	Levin score for NLTK and TextRank	Jaccard score for NLTK and TextRank
0.52231405	0.721153846	1	0	0.52231405	0.278846154	0.721153846	0.465495609	0.401709402	0.598290598	0.462887989	0.792307692	0.465495609	0.598290598	0.510080645	0.776470588	0.510080645	0.776470588	0.411392405	0.825688073
0.45807771	0.892857143	0.801639344	0.455445545	0.453488372	0.105263158	0.894736842	0.738880918	0.540540541	0.459459459	0.435763889	0.879032258	0.732835821	0.530434783	0.47639485	0.8125	0.498861048	0.857142857	0.408745247	0.855855856
0.586166471	0.421875	0.476595745	0.813186813	0.483135825	0.426966292	0.573033708	0.789915966	0.505617978	0.494382022	0.712601995	0.573033708	0.528541226	0.457142857	0.488322718	0.842696629	0.538681948	0.777777778	0.512658228	0.829545455
0.476095618	0.827956989	0.496927129	0.838095238	0.798086124	0.608695652	0.391304348	0.749140893	0.448275862	0.551724138	0.511175899	0.79787234	0.512911843	0.872727273	0.52722063	0.787234043	0.69582505	0.697674419	0.545101843	0.804123711
0.493464052	0.828125	0.843205575	0.333333333	0.530973451	0.208955224	0.791044776	0.55027933	0.517857143	0.482142857	0.538461538	0.846153846	0.566153846	0.480769231	0.5572843	0.568965517	0.555382215	0.574074074	0.988505747	0.196078431
0.488771466	0.802631579	0.499445061	0.773809524	0.738947368	0.506849315	0.493150685	1	1	0	0.738947368	0.493150685	0.499445061	0.773809524	0.461883408	0.784810127	0.655221745	0.361702128	0.461883408	0.784810127
0.517836594	0.880434783	0.615062762	0.463768116	0.72361809	0.409638554	0.590361446	1	1	0	0.72361809	0.590361446	0.615062762	0.463768116	0.49787234	0.826086957	0.484029484	0.839506173	0.49787234	0.826086957
0.795555556	0.535211268	0.491155047	0.620253165	0.464921466	0.132653061	0.867346939	1	1	0	0.464921466	0.867346939	0.491155047	0.620253165	0.494736842	0.670731707	0.987709497	0.24137931	0.494736842	0.670731707
0.634032634	0.671641791	0.458452722	0.854166667	0.595443833	0.528735632	0.471264368	1	1	0	0.595443833	0.471264368	0.458452722	0.854166667	0.432432432	0.852631579	0.570048309	0.333333333	0.432432432	0.852631579
1	0	1	0	1	1	0	1	1	0	1	0	1	0	0.421940928	0.704545455	0.421940928	0.704545455	0.421940928	0.704545455
0.45804541	0.866666667	0.491909385	0.595959596	0.44648318	0.136842105	0.863157895	0.490166415	0.15037594	0.84962406	0.447679709	0.900826446	0.750369276	0.522522523	0.47429171	0.875	0.469035533	0.855769231	0.455648926	0.882352941
0.438650307	0.87704918	0.503026968	0.5	0.413793103	0.128571429	0.871428571	0.606753813	0.552845528	0.447154472	0.43537415	0.853448276	0.479101684	0.819444444	0.434246575	0.848484848	0.451507742	0.88034188	0.638844302	0.704081633
0.468305304	0.845238095	0.445364238	0.875	0.436632747	0.169811321	0.830188679	0.753941056	0.425925926	0.574074074	0.443359375	0.851485149	0.455256299	0.876106195	0.450586265	0.843478261	0.489841986	0.675	0.471416007	0.831775701
1	0	0.407286923	0.897959184	0.407286923	0.102040816	0.897959184	1	1	0	0.407286923	0.897959184	0.407286923	0.897959184	0.370116518	0.902097902	0.4925	0.716216216	0.370116518	0.902097902
0.49382716	0.823529412	0.524590164	0.815533981	0.498207885	0.138211382	0.861788618	1	1	0	0.498207885	0.861788618	0.524590164	0.815533981	0.500541712	0.804123711	0.541720154	0.567164179	0.500541712	0.804123711
0.443266172	0.854166667	1	0	0.443266172	0.145833333	0.854166667	1	1	0	0.443266172	0.854166667	1	0	0.468784228	0.811764706	0.468784228	0.811764706	0.468784228	0.811764706
0.491295938	0.834862385	0.751918159	0.583333333	0.50364204	0.163461538	0.836538462	0.751918159	0.416666667	0.583333333	0.491295938	0.834862385	1	0	0.541448059	0.845360825	0.536062378	0.831683168	0.536062378	0.831683168
0.48940678	0.831683168	0.51044505	0.761904762	0.489099526	0.177570093	0.822429907	0.745158792	0.37962963	0.62037037	0.458553792	0.867768595	0.471186441	0.81512605	0.493975904	0.775510204	0.531073446	0.808510638	0.48	0.81981982
0.469333333	0.550561798	0.511124474	0.448275862	0.430835735	0.198412698	0.801587302	0.769311613	0.46969697	0.53030303	0.44981685	0.788135593	0.464634146	0.813793103	0.372934697	0.85	0.446428571	0.828282828	0.378205128	0.888888889
0.42020202	0.858585859	1	0	0.42020202	0.141414141	0.858585859	0.816466552	0.555555556	0.444444444	0.418467583	0.885714286	0.816466552	0.444444444	0.648711944	0.569230769	0.648711944	0.569230769	0.430839002	0.845238095
0.435013263	0.831325301	0.699588477	0.438596491	0.458204334	0.18	0.82	0.693404635	0.397959184	0.602040816	0.435919791	0.875968992	0.379272326	0.84	0.367479675	0.882352941	0.475	0.87755102	0.302648172	0.927083333
0.490280778	0.835164835	0.710323575	0.602040816	0.425496689	0.134453782	0.865546218	1	1	0	0.425496689	0.865546218	0.710323575	0.602040816	0.453740949	0.865546218	0.501560874	0.847826087	0.453740949	0.865546218
0.420091324	0.919117647	1	0	0.420091324	0.080882353	0.919117647	1	1	0	0.420091324	0.919117647	1	0	0.415627598	0.92	0.415627598	0.92	0.415627598	0.92
0.437012263	0.887573964	0.733627963	0.539325843	0.399258916	0.107843137	0.892156863	1	1	0	0.399258916	0.892156863	0.733627963	0.539325843	0.343977591	0.877192982	0.394366197	0.867647059	0.343977591	0.877192982
0.477348066	0.873684211	0.444610778	0.903703704	0.399639964	0.138888889	0.861111111	0.706125258	0.440366972	0.559633028	0.442270059	0.828282828	0.482758621	0.862903226	0.436960276	0.851851852	0.504201681	0.886597938	0.505144995	0.83
0.434782609	0.890322581	0.65919511	0.62745098	0.406964091	0.094972067	0.905027933	0.606312292	0.5125	0.4875	0.392783505	0.914893617	0.456174334	0.887755102	0.368591823	0.868613139	0.444073456	0.899159664	0.37492392	0.875862069
0.460595447	0.877192982	0.44589309	0.672131148	0.422680412	0.102564103	0.897435897	1	1	0	0.422680412	0.897435897	0.44589309	0.672131148	0.152091255	0.933333333	0.249221184	0.903225806	0.152091255	0.933333333
1	0	0.771929825	0.396551724	0.771929825	0.603448276	0.396551724	1	1	0	0.771929825	0.396551724	0.771929825	0.396551724	0.140339635	0.925373134	0.186351706	0.898058252	0.140339635	0.925373134
0.465152702	0.893939394	0.367634855	0.901345291	0.34410407	0.077981651	0.922018349	1	1	0	0.34410407	0.922018349	0.367634855	0.901345291	0.262801932	0.904255319	0.402489627	0.906542056	0.262801932	0.904255319
1	0	1	0	1	1	0	1	1	0	1	0	1	0	0.227272727	0.869565217	0.227272727	0.869565217	0.227272727	0.869565217
0.442548921	0.710691824	0.848448687	0.385	0.313968958	0.090909091	0.909090909	1	1	0	0.313968958	0.909090909	0.848448687	0.385	0.270833333	0.916666667	0.277837838	0.918128655	0.270833333	0.916666667
0.394321767	0.901408451	0.621434079	0.669421488	0.522439586	0.12745098	0.87254902	0.621434079	0.330578512	0.669421488	0.394321767	0.901408451	1	0	0.674318508	0.545454545	0.350364964	0.904	0.350364964	0.904
0.374269006	0.824427481	1	0	0.374269006	0.175572519	0.824427481	1	1	0	0.374269006	0.824427481	1	0	0.401691332	0.845588235	0.401691332	0.845588235	0.401691332	0.845588235
0.562330623	0.647540984	0.785087719	0.381294964	0.340963855	0.102040816	0.897959184	0.785087719	0.618705036	0.381294964	0.562330623	0.647540984	1	0	0.346428571	0.898648649	0.423796791	0.847222222	0.423796791	0.847222222
0.406660824	0.869918699	0.43620178	0.826815642	0.317915309	0.096551724	0.903448276	0.707278481	0.653333333	0.346666667	0.326654523	0.907407407	0.431115276	0.601226994	0.330944625	0.866666667	0.424189308	0.843478261	0.310868245	0.875
0.463672391	0.601694915	0.525161933	0.538461538	0.419121734	0.119047619	0.880952381	0.754867257	0.483221477	0.516778523	0.41992077	0.907407407	0.458734177	0.875	0.395465995	0.849315068	0.445126631	0.791666667	0.394601542	0.917241379
0.304822566	0.885572139	0.672566372	0.420408163	0.284873022	0.094117647	0.905882353	0.629287911	0.503546099	0.496453901	0.292307692	0.903361345	0.844660194	0.367256637	0.274408284	0.921487603	0.289244851	0.927083333	0.268620269	0.934210526
0.55613577	0.815789474	0.506887052	0.641304348	0.447284345	0.126315789	0.873684211	0.854673998	0.565217391	0.434782609	0.444210526	0.87628866	0.68	0.619565217	0.498069498	0.848484848	0.539976825	0.802469136	0.460362942	0.862745098
0.556359876	0.446153846	1	0	0.556359876	0.553846154	0.446153846	1	1	0	0.556359876	0.446153846	1	0	0.357771261	0.878787879	0.357771261	0.878787879	0.357771261	0.878787879
0.539419087	0.793103448	0.4585126	0.845070423	0.393527508	0.123188406	0.876811594	0.772858518	0.555555556	0.444444444	0.41130742	0.897810219	0.448897796	0.848920863	0.440179143	0.868613139	0.737270876	0.636363636	0.429867411	0.864661654
0.423307626	0.879310345	0.654024052	0.610062893	0.319210679	0.083832335	0.916167665	1	1	0	0.319210679	0.916167665	0.654024052	0.610062893	0.362527716	0.875739645	0.442307692	0.879032258	0.362527716	0.875739645

Experiments and Results

Key Results

- L-score is ***more optimistic*** compared to J-score
- All methods have ***lowest similarity J-score with LSA***
- Luhn and Text Rank seem to have ***highest similarity*** J-score
- LexRank and Text Rank summarization ***differs significantly*** although both use PageRanking/Graph based model!
- Maximal Consensus (highest number of methods with similar summarization) provided good summarization and validates our hypothesis
- Associativity of similarity does not hold with summarization!

Jaccard Score (word level)					
	LSA	Luhn	LexRank	Text Rank	NLTK
LSA	100%	29%	29%	28%	24%
Luhn	71%	100%	52%	74%	18%
Lex Rank	71%	48%	100%	51%	23%
Text Rank	72%	26%	49%	100%	17%
NLTK	76%	82%	77%	83%	100%
Simmilarity average: 35%					
Dissimilarity average: 65%					

Leveinshtein Score (character level)					
	LSA	Luhn	LexRank	Text Rank	NLTK
LSA	100%	47%	47%	48%	50%
Luhn	53%	100%	31%	16%	58%
Lex Rank	53%	69%	100%	31%	53%
Text Rank	52%	84%	69%	100%	58%
NLTK	50%	42%	47%	42%	100%
Dissimilarity average: 43%					
Simmilarity average: 57%					

Summarization Output

		Input	Comparison of Output				
document	sectDin	text	NLTK	Lex_Rank	LSA	TextRank	Luhn
CMS_2014_0115_0059.pdf	6	['FDA approved contraception as prescribed,Äù and ,Äúcertain contraception services,Äù. Is there', 'more clarity regarding which methods of contraception must be excluded?', 'The proposed rule references methods of contraception in two seemly distinct ways. The first', 'reference is regarding the coverage requirement for non-eligible entities which is referenced', 'numerous times as ,ÄúFood and Drug Administration (FDA) approved contraception as prescribed by', 'a health care provider,Äù. These references are made both directly and through reference to the', 'Health Resources and Services (HRSA) guidelines. A second and distinct reference to contraceptive', 'methods, specifically in connection with those methods to which the eligible entity may object is', 'repeatedly referred as ,Äúcertain contraceptive services,Äù. Nationwide requests clarity on these', 'references. Is there a method of contraception that is both required under Section 2713 of the', 'Public Health Services Act (PHS) and which would be allowably included in the coverage offered by', 'the eligible entity? The eligible entity filing the notice EBSA 700, the prescribed method of notice', 'found in the interim rule, to request exclusion from the coverage requirement is in the best position', 'to specifically name the types of contraceptive services which would be allowable, if any. Including', 'this information on the notice would be a valuable tool in crafting an acceptable coverage agreement', 'for both parties.', 'Nationwide also proposes two alternate solutions for students; please see below.']	The commenter stated that a second and distinct reference to contraceptive methods, specifically in connection with those methods to which the eligible entity may object is repeatedly referred as ,Äúcertain contraceptive services,Äù. The commented further stated that fda approved contraception as prescribed,Äù and ,Äúcertain contraception services,Äù.	The commenter stated that is there', 'more clarity regarding which methods of contraception must be excluded? The commented further stated that a second and distinct reference to contraceptive', 'methods, specifically in connection with those methods to which the eligible entity may object is', 'repeatedly referred as ,Äúcertain contraceptive services,Äù.	The commenter stated that is there a method of contraception that is both required under section 2713 of the', 'public health services act (phs) and which would be allowably included in the coverage offered by', 'the eligible entity? The commented further stated that including', 'this information on the notice would be a valuable tool in crafting an acceptable coverage agreement', 'for both parties.	The commenter stated that is there a method of contraception that is both required under section 2713 of the', 'public health services act (phs) and which would be allowably included in the coverage offered by', 'the eligible entity? The commented further stated that the eligible entity filing the notice ebsa 700, the prescribed method of notice', 'found in the interim rule, to request exclusion from the coverage requirement is in the best position', 'to specifically name the types of contraceptive services which would be allowable, if any.	The commenter stated that is there a method of contraception that is both required under section 2713 of the', 'public health services act (phs) and which would be allowably included in the coverage offered by', 'the eligible entity? The commented further stated that the eligible entity filing the notice ebsa 700, the prescribed method of notice', 'found in the interim rule, to request exclusion from the coverage requirement is in the best position', 'to specifically name the types of contraceptive services which would be allowable, if any.

Best Model:

- Maximal consensus on summarization seems to be a good choice
- Luhn and Text Rank have highest similarity score in our analysis
- Jaccard score is a better candidate for text summarization comparison

Lesson Learned & Future Work

Lessons

- Check the integrity of your dataset until the last moment
- Make sure to manually inspect where your model is making mistakes
- ML is not a panacea to all ills, so be flexible about other ways of supporting it
- NLTK based summarization are counterintuitive as was shown in metrics table
- Jaccard score is a better metric for comparison
- Maximal consensus based summarization gives better quality results

Future Work

- Evaluate abstractive summarization
- Explore CNN vector representations
- Evaluate models using other metrics such as Rouge, Blue, and Meteor

Thanks!
Questions?

Main contribution from team members

Gayani Perera

- PDF Ingestion, Feature engineering, model implementation : Random Forest
- Extractive text summarization

Liliana Cruz-Lopez

- Module 1: Converted PDFs to HTMLs, extracted raw features from HTMLs and contributed to engineered features
- Module 2: completed end-to-end text summarization

Minsu Yeom

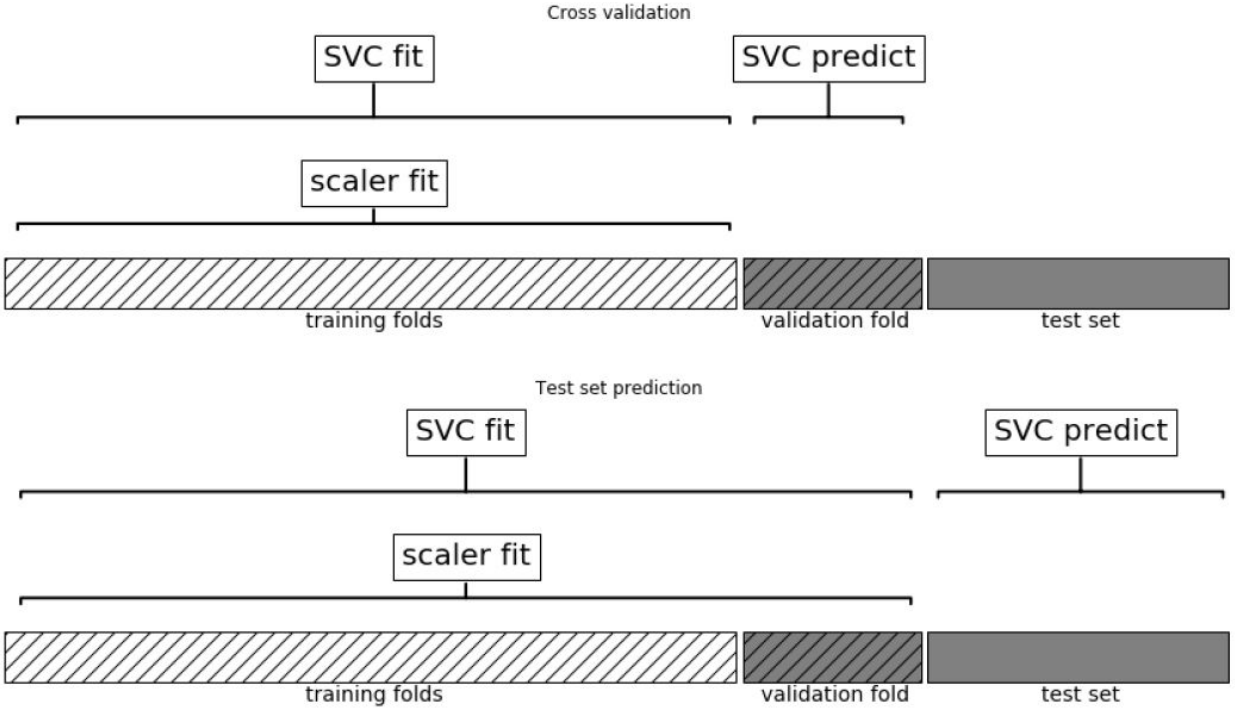
- Preprocessing: Feature engineering (Line space(LS), Ratio of title word to total), Converted PDFs to XMLs
- Model implementation: XGBoost

Pranjal Bajaj

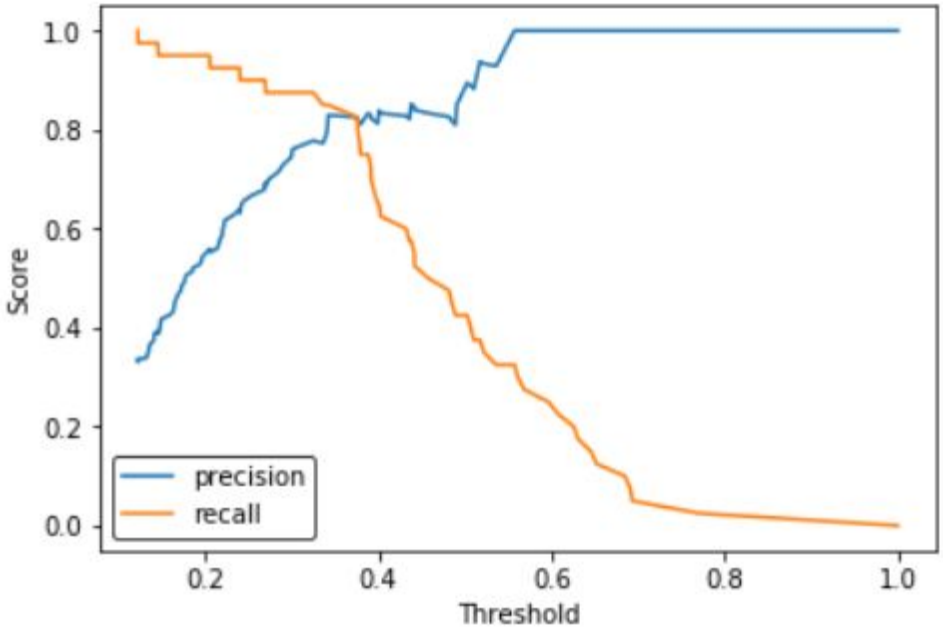
- Model concept and development
- Model implementation: Choosing Metrics and Implementing best practices using scikit-learn

Appendix

Scikit-learn Pipeline



Precision - Recall vs Threshold for Best Model: Random Forest



HTML-based features (“raw”)

Category	Feature Name	Description
Numerical	Font_Size	Size of the font in each line
	Left	A distance from left outer edge in pixel
	Top	A distance from top outer edge in pixel
Categorical	Font_Family	Font style of the text
	Font_Weight	Line of the text bold or normal
	Font_Style	Line of the text italic or normal
	Color	Color of the text
	id	Class id in HTML containing all the information of font
Text	Text	Text of each line

Table of Best Results

Method	Threshold	Preision	Recall	F1-Score	Accuracy	ROC		TN	FN	TP	FP
RF with new 4 features no SMOTE	0.27	0.71	0.88	0.79	0.99	0.93		2269	5	35	14
	0.28	0.74	0.88	0.8	0.99	0.93		2271	5	35	12
RF with new 4 features no SMOTE with RF association rule	0.27	0.72	0.9	0.8	0.99	0.95		2269	4	36	14
	0.28	0.75	0.9	0.82	0.99	0.95		2271	4	36	12
RF with new 4 features with SMOTE	0.27	0.7	0.88	0.78	0.99	0.93		2268	5	35	15
	0.28	0.71	0.88	0.79	0.99	0.93		2269	5	35	14
	0.29	0.74	0.88	0.8	0.99	0.93		2271	5	35	12
	0.3	0.76	0.88	0.81	0.99	0.94		2272	5	35	11
	0.31	0.78	0.88	0.82	0.99	0.94		2273	5	35	10
	0.32	0.78	0.88	0.82	0.99	0.94		2273	5	35	10
RF with new 4 features with SMOTE with RF association rule	0.27	0.71	0.9	0.79	0.99	0.95		2268	4	36	15
	0.28	0.72	0.9	0.8	0.99	0.95		2269	4	36	14
	0.29	0.75	0.9	0.82	0.99	0.95		2271	4	36	12
	0.3	0.77	0.9	0.83	0.99	0.95		2272	4	36	11
BEST RESULTS	0.31	0.78	0.9	0.84	0.99	0.95		2273	4	36	10
	0.32	0.78	0.9	0.84	0.99	0.95		2273	4	36	10