Team 2
Eric Waltenburg Political Science Department,

Greg Chininis Stephen Everett
Jeremy Frank Jim He
Josh Lefton Spencer Prentiss

CS490-DSC, Instructor: Jean Honorio, Spring 2022

Business Goals and Data Mining Goals

- Business goals:
  - Given that the sponsor for our project is a professor (Prof. Waltenburg) and not a business or corporate partner, our “business units” are a potential gain in knowledge regarding the U.S. Court of Appeals. Information gained can then be used in future studies about the United States judiciary system to promote change and/or reform.
- Data mining goals:
  - We are tasked with, given an input of written court decisions in pdf format, creating metrics for both "polarity" and "complexity" that can be applied to each decision as well as aggregated across many decisions. Unlike other groups, it is more difficult to verify/score our model given the subjectivity of the metrics created. Thus, this project will require close collaboration with Prof. Waltenburg to ensure the results make sense.

Data Preparation

- The main attribute used was the "Opinion" attribute, which was a string comprised of the entire written opinion in question. In addition, each court opinion also has CaseID and CourtID attributes that were not used in modeling but were quite helpful for subsetting and aggregating results.
- We cleaned the data by removing stopwords, punctuation, and numerics. We also lemmatized the words and created a document term matrix based on the cleaned text.
- The final dataset has 6,000 samples, 2 numeric features CaseID and CourtID, 1 text feature Opinions, and after completing the modeling stage also have 3 “score” columns. Those being complexity, polarity, and subjectivity.

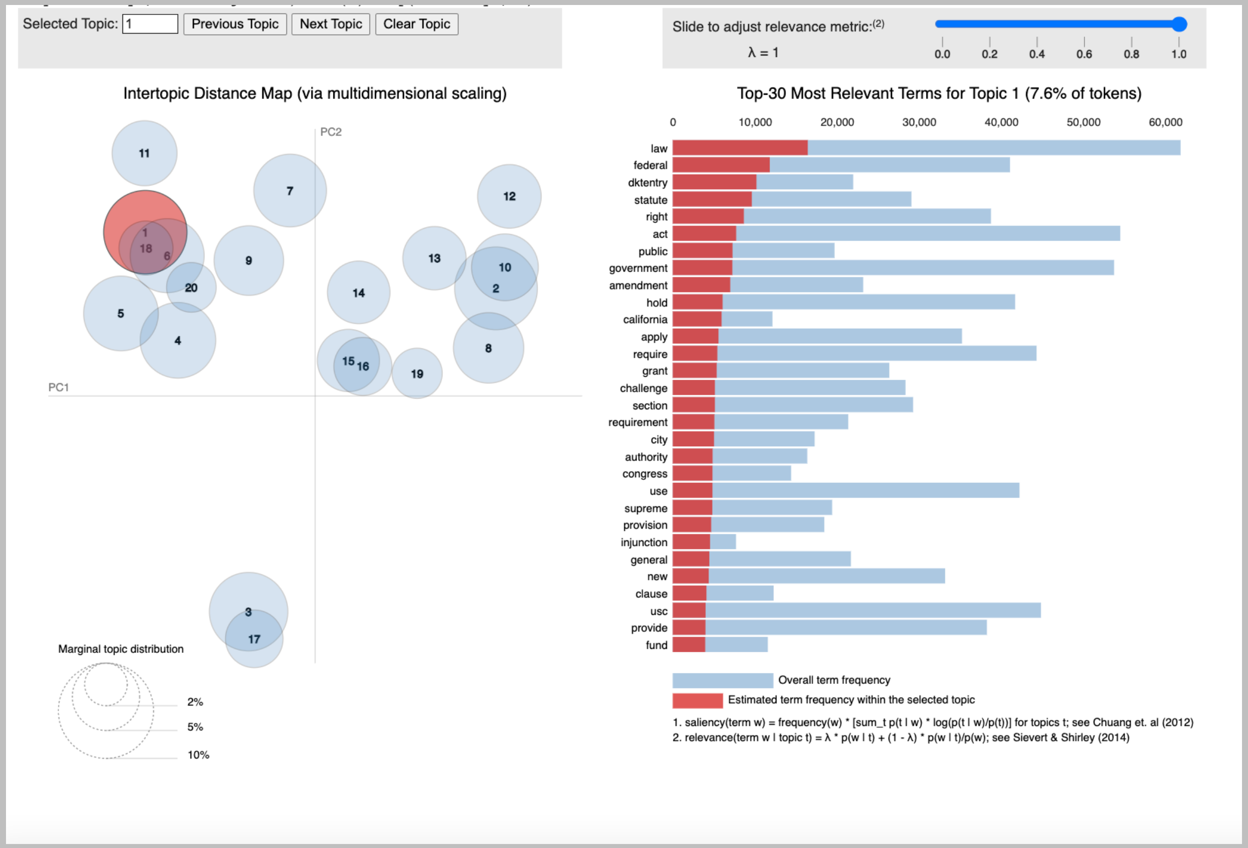
Model Description

- Machine learning algorithms:
  - To model complexity, we needed to use a technique that can determine which topic(s) are covered in the document. The model we chose for this was Latent Dirichlet Allocation (LDA).
  - To model polarity/subjectivity, we needed to use a sentiment analysis modeling technique that would determine how positive/negative the sentiment of the text is. We decided to use TextBlob for this.

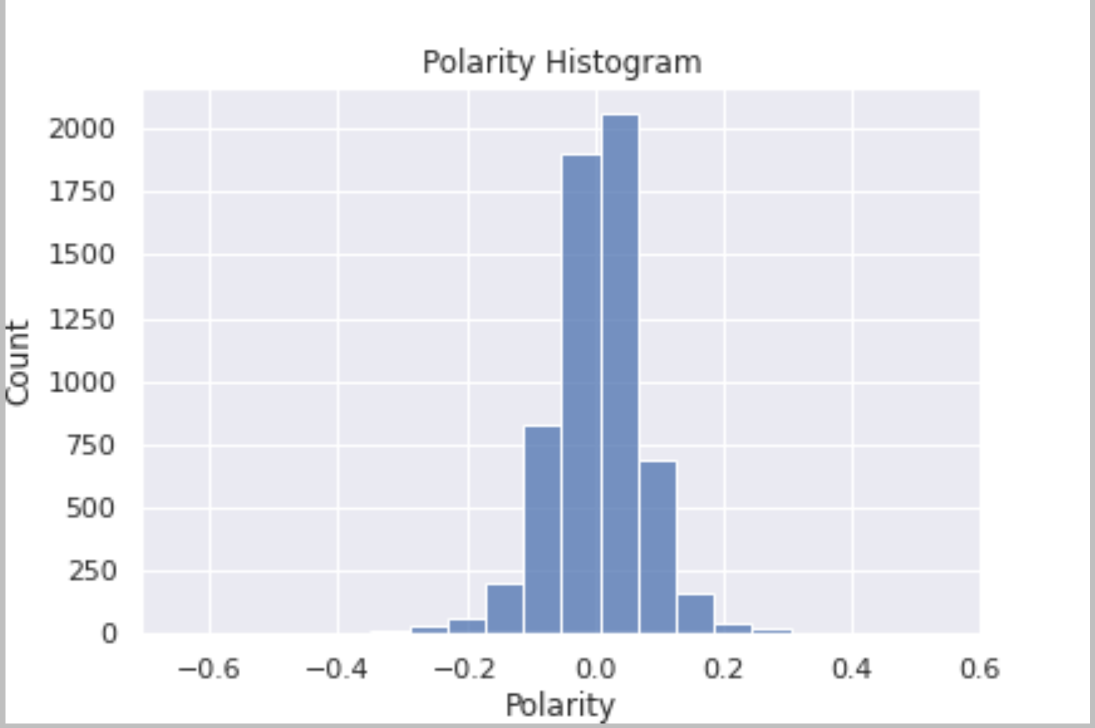
Cross-validation:

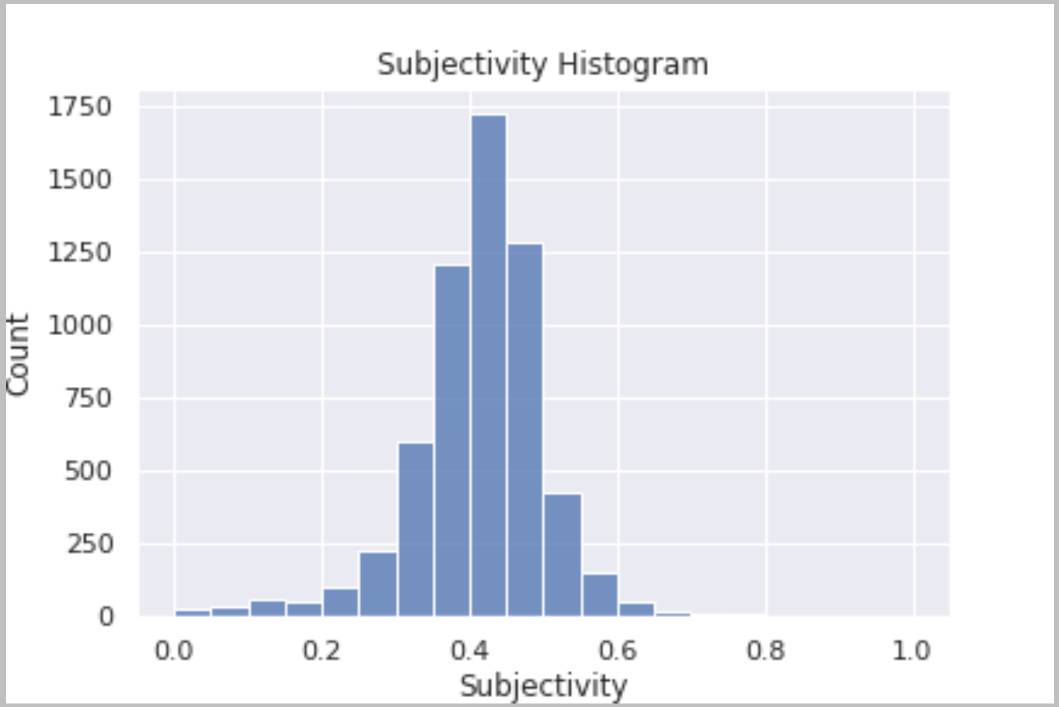
- For each hyperparameter value, we randomly shuffled the data and ran it 5 times. Then we aggregated the results to determine performance and the best model parameters.

Experimental Results



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)





Impact of the Model to the Business Goals

- In terms of what we wanted to achieve in creating reliable metrics for polarity and complexity, we believe we have met our data mining goal. Additionally, after meeting with Professor Waltenburg and taking him through our results, he was very pleased with the metrics that we came up with. We were successful in creating metrics that are understandable to those without much data science experience, as Professor Waltenburg was able to grasp what we did easily.
- We really will not know, however, what the full effect of our metrics will be until Professor Waltenburg completes his final research. Professor Waltenburg plans to use our metrics to see if they stem from more diverse courts as well as some other possible relationships dealing with ideology of the judges.

Submit this document on Brightspace, as a Word document (PDF or other formats are not allowed). Write only on the blue areas. Do not change or remove any text outside of the blue areas.

Report the amount of effort that each team member put into this assignment. List the names of each of the team members (including those who did not work, if any) and their percentage of effort (from 0% to 100%). In a team where everybody made roughly the same effort, I expect to see 100% for all.

Spencer – 100%
Greg – 100%
Josh – 100%
Jim – 100%
Stephen – 100%
Jeremy – 100%

1. DETERMINE BUSINESS OBJECTIVES

The first objective of the analyst is to thoroughly understand, from a business perspective, what the customer really wants to accomplish. Often the customer has many competing objectives and constraints that must be properly balanced. The analyst's goal is to uncover important factors at the beginning of the project that can influence the final outcome. A likely consequence of neglecting this step would be to expend a great deal of effort producing the correct answers to the wrong questions.

1.1. Background

Collate the information that is known about the organization's business situation at the start of the project. These details not only serve to more closely identify the business goals to be achieved but also serve to identify resources, both human and material, that may be used or needed during the course of the project.

Organization

- [Optional] Develop organizational charts identifying divisions, departments, and project groups. The chart should also identify managers' names and responsibilities
- [Optional] Identify key persons in the business and their roles
- Identify an internal sponsor (financial sponsor and primary user/domain expert)
- [Optional] Indicate if there is a steering committee and list members
- Identify the business units which are affected by the data mining project (e.g., Marketing, Sales, Finance)

The internal sponsor of this project is Professor Eric Waltenburg. Contrary to many of the other project opportunities, our sponsor is not necessarily interested in business units, but rather than gain of information in a certain field. Our “business units” is potential knowledge regarding the U.S. Court of Appeals. Information gained from this task can be used in future studies about the United States judiciary system and potentially used for changes/reform.

Problem area

- Identify the problem area (e.g., marketing, customer care, business development, etc.)

- Describe the problem in general terms
- [Optional] Check the current status of the project (e.g., Check if it is already clear within the business unit that a data mining project is to be performed, or whether data mining needs to be promoted as a key technology in the business)
- Clarify prerequisites of the project (e.g., What is the motivation of the project? [Optional] Does the business already use data mining?)
- [Optional] If necessary, prepare presentations and present data mining to the business
- [Optional] Identify target groups for the project result (e.g., Are we expected to deliver a report for top management or an operational system to be used by naive end users?)
- Identify the users' needs and expectations

The issue is at the crossroads of judicial political science and small group dynamics. The question being posed is how the configuration of a panel and small group forces might influence the complexity and polarity of a court judgment.

Another problem area is innovating new metrics of court ruling complexity and polarity, as well as adding data on judge archetypes, which will reveal trends in how diverse court panel compositions affect court rulings.

Professor Waltenburg hopes to obtain a better understanding of how the configuration of a panel and small group forces might influence the complexity and polarity of a court judgement.

Current solution

- Describe any solution currently used to address the problem
- Describe the advantages and disadvantages of the current solution and the level to which it is accepted by the users

It appears that most of the current knowledge on the topic is based more on psychological beliefs and less on empirical research. “The Law of Group Polarization” (Sunstein, 2002) addresses how group-think can lead to different decisions, and while it’s possible to hypothesize that courts would act in a similar manner, there is nothing currently that can quantify its impact (if at all). While relevant parties currently will use their beliefs to make decisions, it would be much more advantageous to have evidence to back it up.

1.2. Business objectives

Describe the customer's primary objective, from a business perspective. In addition to the primary business objective, there are typically a large number of related business questions that the customer would like to address. For example, the primary business goal might be to keep current customers by predicting when they are prone to move to a competitor, while a secondary business objective might be to determine whether lower fees affect only one particular segment of customers.

- Informally describe the problem to be solved

- Specify all business questions as precisely as possible
- [Optional] Specify any other business requirements (e.g., the business does not want to lose any customers)
- Specify expected benefits in business terms

Beware of setting unattainable goals—make them as realistic as possible.

Our goal is to create new metrics of a court ruling's complexity and polarity, as well as potentially adding data on judge archetypes. Additionally, to better understand the nature of the court. And in a wider view, it helps people understand how a small group affects the final decision.

The overarching research objective is to show that more complex and polarizing decisions stem from more diverse panels. However, Professor Waltenburg is not making any statements on the value of more diverse panels. This will give us a fuller and better understanding of how small group forces affect output of collective decision making, providing the first step towards drawing conclusions.

More specifically, our first goal is to create accurate measurements for the complexity and polarity of court decisions. Once we complete this, we can work towards finding meaningful relationships between these measurements and the diversity of the panels. With this information we can gain a fuller and deeper understanding of the decisions that are handed down, and a better understanding of how small group forces affect the output of collective decisions. In addition, if we can show that the relationships between our variables are meaningful, we can try to predict the outcomes of future decisions.

1.3. Business success criteria

Describe the criteria for a successful or useful outcome to the project from the business point of view. This might be quite specific and readily measurable, such as reduction of customer churn to a certain level, or general and subjective, such as "give useful insights into the relationships." In the latter case, be sure to indicate who would make the subjective judgment.

- Specify business success criteria (e.g., Improve response rate in a mailing campaign by 10 percent and sign-up rate by 20 percent)
- Identify who assesses the success criteria

Each of the success criteria should relate to at least one of the specified business objectives.

There are no specific criteria for which we would expect to result from a successful project. Ideally, this research would be useful to those who care about court decisions and rulings, but that is not necessarily the goal.

Generally, we want to come up with and decide on the best ways to measure the polarity and complexity of court decisions. Ideally, we would like to be fairly accurate in determining these measurements. Since these are highly subjective, we want Prof. Waltenburg to assess if we are accurate in these measurements. In addition, we want to determine if there is a relationship between the polarity/complexity and the diversity of the panel. Since Prof. Waltenburg is providing us with data concerning the diversity of the panels, success should be slightly more objective, and will hinge on our ability to accurately determine meaningful relationships between the variables. Ideally, we would like to find relationships with high correlation.

2. ASSESS SITUATION

This task involves more detailed fact-finding about all of the resources, constraints, assumptions, and other factors that should be considered in determining the data analysis goal and in developing the project plan.

2.1. Inventory of resources

List the resources available to the project, including personnel (business and data experts, technical support, data mining experts), data (fixed extracts, access to live warehoused or operational data), computing resources (hardware platforms), and software (data mining tools, other relevant software).

Hardware resources

- Identify the base hardware
- Establish the availability of the base hardware for the data mining project
- [Optional] Check if the hardware maintenance schedule conflicts with the availability of the hardware for the data mining project
- [Optional] Identify the hardware available for the data mining tool to be used (if the tool is known at this stage)

As of right now, the software we have at our disposal includes the basic programming languages (Python), Jupyter notebooks, and any data mining packages that are readily available for these programming languages. Currently, the hardware we have at our disposal includes our personal computers as well as Purdue computers and clusters.

Sources of data and knowledge

- Identify data sources
- Identify type of data sources (online sources, experts, written documentation, etc.)
- [Optional] Identify knowledge sources
- [Optional] Identify type of knowledge sources (online sources, experts, written documentation, etc.)
- Check available tools and techniques
- Describe the relevant background knowledge (informally or formally)

Our data sources include PDFs of written court decisions.

The main resource we are using is the 12 US circuit opinions in the past 20 years. Those opinions are published on the gov info page and were collected by the US Government Publishing Office. Each report includes the names of judges that were hearing the case also explanation and analysis of their decision.

We should also be receiving data from some of Professor Waltenburg's graduate students who will be creating archetypes for court judges. This data will be used to judge the makeup of the panels. (We are still unaware how the data will be structured)

Personnel sources

- Identify project sponsor (if different from internal sponsor as in Section 1.1)
- [Optional] Identify system administrator, database administrator, and technical support staff for further questions
- Identify market analysts, data mining experts, and statisticians, and check their availability
- Check availability of domain experts for later phases

The internal project sponsor, Professor Eric Waltenburg, is an expert in judicial political and state politics. He is also leading a research course that is looking into the effect of group polarization on law. Professor Eric Waltenburg said that we could exchange opinions via email or schedule meetings.

We also have access to our TA, Hasan, who should be able to help us in the data mining aspect of our project. Prof. Waltenburg also suggested that we look into existing methods to measure the polarity of text, so our domain experts will end up including the internet to some extent.

2.2. Requirements, assumptions, and constraints

List all requirements of the project, including schedule of completion, comprehensibility, and quality of results and security, as well as legal issues. As part of this output, make sure that you are allowed to use the data.
List the assumptions made by the project. These may be assumptions about the data, which can be verified during data mining, but may also include non-verifiable assumptions related to the project. It is particularly important to list the latter if they will affect the validity of the results.
List the constraints made on the project. These constraints might involve lack of resources to carry out some of the tasks in the project in the time required, or there may be legal or ethical constraints on the use of the data or the solution needed to carry out the data mining task.

Requirements

- [Optional] Specify target group profile
- [Optional] Capture all requirements on scheduling
- Capture requirements on comprehensibility, accuracy, deploy ability, maintainability, and repeatability of the data mining project and the resulting model(s)
- Capture requirements on security, legal restrictions, privacy, reporting, and project schedule

We will require circuit court results in a tabular format. The data was delivered in pdf format, but we have been informed that a course TA is working to put it into a tabular format. To make the results of our study more comprehensible, we will likely sacrifice some accuracy to favor a smaller number of classifications. Prof. Waltenburg's team has researched judge archetypes and estimated that there may be around five types of judge archetypes, so we will not choose a much larger number even if our analysis shows that it may be more accurate. Accuracy of classification can be performed using inter- and intracluster distance metrics if Prof. Waltenburg would like us to evaluate his team's classification, in which case we would just need his team's results. Accuracy of decision complexity and polarization and panel diversity will require some input from the domain expert, Prof. Waltenburg, to establish a ground truth. The model will be deployed by providing the code with instructions on how to use it to Prof. Waltenburg's team. The model will be unit tested to ensure that the team can use it for future data if they would like. The results should be repeatable given a new set of decisions or judges, although exact results may vary if we use a randomized algorithm such as K-means clustering.

Assumptions

- Clarify all assumptions (including implicit ones) and make them explicit (e.g., to address the business question, a minimum number of customers with age above 50 is necessary)
- List assumptions on data quality (e.g., accuracy, availability)
- [Optional] List assumptions on external factors (e.g., economic issues, competitive products, technical advances)
- [Optional] Clarify assumptions that lead to any of the estimates (e.g., the price of a specific tool is assumed to be lower than $1,000)
- List all assumptions regarding whether it is necessary to understand and describe or explain the model (e.g., how should the model and results be presented to senior management/sponsor)

Our project assumes that there are around five types of judge archetypes that can be determined from looking at the judges' background. Additionally, once we actually get the format of our data, we might need to make some assumptions about what is considered polarizing and complex. Another assumption we are making is that the decisions from each panel are made up of 3 judges. Some cases can be heard in an en banc session, that is where all the judges in a court of appeals hear the case, rather than a panel of judges. However, we are assuming each case has a panel of 3 judges.

We are assuming that the data provided to us is large enough and covers a wide variety of cases to be able to determine a reasonable quantifiable score that measures how polarizing and complex a decision will be from a Court of Appeals case. Additionally, if we are given the judge archetype data, we assume that all the data provides us with a large enough historical database to be able to create a classification model.

In terms of explaining the model, we are also assuming that we will need to describe what factors go into our scoring of polarity and complexity. If we are provided with enough data to perform classification, we will ensure said model is easily understood for Professor Waltenburg and his research team.

Constraints

- Check general constraints (e.g., legal issues, budget, timescales, and resources)
- Check access rights to data sources (e.g., access restrictions, password required)
- Check technical accessibility of data (operating systems, data management system, file or database format)
- Check whether relevant knowledge is accessible
- [Optional] Check budget constraints (fixed costs, implementation costs, etc.)

There are no legal constraints with this project or budget constraints on our end. The main constraint right now is still obtaining the data in the desired format. There are no constraints to access the data as it is freely available on a government website. However, as of now, the data is in the form of multiple, lengthy PDF files that also contain information that is irrelevant to our project. We have been told a course TA will provide us with the data in the desired format but have not received it yet. Another constraint is whether we will be receiving data from Professor Waltenburg's other classes on the judge archetypes that will help us determine the makeup of the panels. Additionally, the time constraint is the remaining length of the semester (about 3 months).

Other constraints will be if we are able to use tools and hardware that have the capability to handle large amounts of text data.

Professor Waltenburg was already able to give us a crash course in the relevant knowledge and said we could set up more meetings or email him if we had any further questions about the research topic.

2.3. Risks and contingencies

List the risks, that is, the events that might occur, impacting schedule, cost, or result. List the corresponding contingency plans: what action will be taken to avoid or minimize the impact or recover from the occurrence of the foreseen risks.

Identify risks

- [Optional] Identify business risks (e.g., competitor comes up with better results first)
- [Optional] Identify organizational risks (e.g., department requesting project doesn't have funding for the project)
- [Optional] Identify financial risks (e.g., further funding depends on initial data mining results)
- Identify technical risks
- Identify risks that depend on data and data sources (e.g., poor quality and coverage)

If a course TA does not provide it for us, one risk is not being able to quickly convert the PDFs into readable formats and being forced to spend a lot of our time just converting them into something useable.

We may also run into problems if we do not receive the judge archetype data from Professor Waltenburg's other research students as it would prevent us from completing some of our analysis.

We will be analyzing and building our models based on a very large amount of data, so we may run into issues concerning the runtime of our code.

We will probably run into issues with misspelled words in the decisions, as well.

Develop contingency plans

- Determine conditions under which each risk may occur

- Develop contingency plans

We will remain in contact with the course TAs to monitor progress on the pdf conversion. Until we receive the data, we can work on building our domain knowledge in the subject so that we will be ready to select relevant features and perform meaningful analysis once we receive the data. If this starts taking many weeks, we can offer help parsing the data.

If we are not given judge archetype data or Prof. Waltenburg requests that we form our own classifications, we can build a separate model to classify the judges. We could meet with the team to discuss their approach and determine ways we might be able to improve it. If Prof. Waltenburg would like us to use their classification as is, this won't be necessary.

We might run into runtime issues due to the large amount of data we are using. If this occurs, the simplest solution will be to cut down the number of decisions we use to build the model. There may also be ways to simplify the models we are using.

We will most likely have issues with misspelled words in the decisions. Since it will be nearly impossible to identify all the misspelled words and convert them to their correct spelling, we may need to ignore words that fall below a certain frequency threshold. This will also improve the runtime of the model.

2.4. Terminology

Compile a glossary of terminology relevant to the project. This should include at least two components: (1) A glossary of relevant business terminology, which forms part of the business understanding available to the project (2) A glossary of data mining terminology, illustrated with examples relevant to the business problem in question

- [Optional] Check prior availability of glossaries; otherwise begin to draft glossaries
- Talk to domain experts to understand their terminology
- Become familiar with the business terminology

Complexity: We still have an undefined concept of complexity at this point as we could go several different ways with it, but it basically could boil down to how many different topics are discussed/ how hard is the text to read, etc.

Polarity: In terms of this project, we are looking at whether an opinion falls on the left or right side of the political ideology spectrum.

Diversity: A measure of how different the judges and their backgrounds are on a given panel.

Panel: A group of 3-5 justices (sometimes more) that presides and rules on a given case.

Court of Appeals: The appellate court's task is not to rule whether something or someone was right or wrong, rather was the law applied to the case correctly/constitutionally.

2.5. Costs and benefits

Prepare a cost-benefit analysis for the project, comparing the costs of the project with the potential benefits to the business if it is successful

- [Optional] Estimate costs for data collection
- [Optional] Estimate costs of developing and implementing a solution
- Identify benefits (e.g., improved customer satisfaction, ROI, and increase in revenue)
- [Optional] Estimate operating costs

[Optional] Remember to identify hidden costs, such as repeated data extraction and preparation, changes in workflows, and time required for training.

A main benefit is the increased scope of research into this subject. Specifically, the benefit is to gain a deeper understanding of the makeup of the decisions being handed down by the Circuit Court of Appeals and how small group forces affect the output of collective decision making. This will provide a basis to form further hypotheses about this subject later.

In addition, if we can show that the diversity of the panel is related to the complexity/polarity of the decision handed down, we can help predict the outcomes of future trials.

3. DETERMINE DATA MINING GOALS

A business goal states objectives in business terminology; a data mining goal states project objectives in technical terms. For example, the business goal might be, "Increase catalog sales to existing customers," while a data mining goal might be, "Predict how many widgets a customer will buy, given their purchases over the past three years, relevant demographic information, and the price of the item."

3.1. Data mining goals

Describe the intended outputs of the project that enable the achievement of the business objectives. Note that these are normally technical outputs.

- Translate the business questions to data mining goals (e.g., a marketing campaign requires segmentation of customers in order to decide whom to approach in this campaign; the level/size of the segments should be specified).
- Specify data mining problem type (e.g., classification, description, prediction, and clustering).

This data mining project has several different problem types. The key variables are attempting to discern are complexity and polarity of court opinions as well as diversity of panels. Complexity and polarity are tricky problems to solve but through the use of sentiment analysis we can classify each opinion as liberal or conservative. Complexity is very hard to determine/measure so several options will be looked at. Some include topic modeling or simply creating a basic metric. For diversity we should be provided panel makeup from Professor Waltenburg, which would allow us to create a metric quantifying how diverse a panel is.

3.2. Data mining success criteria

Define the criteria for a successful outcome to the project in technical terms, for example a certain level of predictive accuracy or a propensity-to-purchase profile with a given degree of "lift." As with business success criteria, it may be necessary to describe these in subjective terms, in which case the person or persons making the subjective judgment should be identified.

- Specify criteria for model assessment (e.g., model accuracy, performance and complexity)
- Define benchmarks for evaluation criteria
- Specify criteria which address subjective assessment criteria (e.g., model explain ability and data and marketing insight provided by the model)

Remember that the data mining success criteria are different than the business success criteria defined earlier.

Model accuracy and performance evaluation will require input from Prof. Waltenburg's team since they are domain experts. We can give them random samples of input and ask for the output they would expect and compare it to the model’s results. Low complexity will be favored since the results of the project are intended to measure and explain a phenomenon. It is important that the results are understandable to non-data mining experts. So, we will favor NLP models with high levels of comprehensibility that we are able to explain to the team. We will confirm that it is sufficiently understandable with the team before implementation.

4. PRODUCE PROJECT PLAN

Describe the intended plan for achieving the data mining goals and thereby achieving the business goals.

4.1. Project plan

List the stages to be executed in the project, together with their duration, resources required, inputs, outputs, and dependencies. Wherever possible, make explicit the large-scale iterations in the data mining process—for example, repetitions of the modeling and evaluation phases. As part of the project plan, it is also important to analyze dependencies between time schedule and risks. Mark results of these analyses explicitly in the project plan, ideally with actions and recommendations for actions if the risks are manifested.

Although this is the only task in which the project plan is directly named, it nevertheless should be consulted continually and reviewed throughout the project. The project plan should be consulted at minimum whenever a new task is started or a further iteration of a task or activity is begun.

- Define the initial project plan [Optional] and discuss the feasibility with all involved personnel
- Combine all identified goals and selected techniques in a coherent procedure that solves the business questions and meets the business success criteria

- [Optional] Estimate the effort and resources needed to achieve and deploy the solution. (It is useful to consider other people's experience when estimating timescales for data mining projects. For example, it is often postulated that 50-70 percent of the time and effort in a data mining project is used in the Data Preparation Phase and 20-30 percent in the Data Understanding Phase, while only 10-20 percent is spent in each of the Modeling, Evaluation, and Business Understanding Phases and 5-10 percent in the Deployment Phase.)
- Identify critical steps
- [Optional] Mark decision points
- [Optional] Mark review points
- [Optional] Identify major iterations

We have about 3 months to finish the project. This is our rough initial plan that will most likely change once we have access to data:

Our first step is to conduct our data analysis report and ensure that we have properly stored the text along with data about the cases in a structured manner.

The second step will be to define the measurements for complexity and polarity and determine the best techniques for creating these scoring metrics, as well as implementing these measurements. Additionally, if we do not receive the judge archetype data, we can create a model to classify the judges.

Then, we will evaluate the performance of our metrics as well as the accuracy of the classification model, analyze the results, and make the according adjustments and improvements to the models.

Finally, we will prepare our final reports and presentation.

4.2. Initial assessment of tools and techniques

At the end of the first phase, the project team performs an initial assessment of tools and techniques. Here, it is important to select a data mining tool that supports various methods for different stages of the process, since the selection of tools and techniques may influence the entire project.

- Create a list of selection criteria for tools and techniques (or use an existing one if available)
- Choose potential tools and techniques
- Evaluate appropriateness of techniques
- Review and prioritize applicable techniques according to the evaluation of alternative solutions

The selection criteria would include whether it could handle large amounts of data, in this case text data. Is it flexible enough to handle/manipulate different data structures? (i.e., Pdf's, text files, json file, ...). Also, how easy is it to use or do we have any prior knowledge of the tool?

The main tools we are looking at are cluster based, like Apache Spark, or Hadoop, or tools that can be run locally. We would primarily use python for the usage of pandas and several NLP libraries like NLTK.

We potentially could be working with a large amount of text data. If the amount is too large cluster-based methods would be our safest bet. If the data is not too large, we could use more local based computing, however this could be very slow depending on what types of data mining techniques are being used.

The priority for rankings is as follows:

Cluster-Based tools > local > (if forced to use online/ cloud computing methods)

Submit this document on Brightspace, as a Word document (PDF or other formats are not allowed). Write only on the blue areas. Do not change or remove any text outside of the blue areas.

Report the amount of effort that each team member put into this assignment. List the names of each of the team members (including those who did not work, if any) and their percentage of effort (from 0% to 100%). In a team where everybody made roughly the same effort, I expect to see 100% for all.

Spencer - 100%
Greg – 100%
Josh – 100%
Jim – 100%
Stephen – 100%
Jeremy – 100%

1. COLLECT INITIAL DATA

Acquire the data (or access to the data) listed in the project resources. This initial collection includes data loading, if necessary for data understanding. For example, if you intend to use a specific tool for data understanding, it is logical to load your data into this tool.
Describe all the various data used for the project, and include any selection requirements for more detailed data. The data collection report should also define whether some attributes are relatively more important than others.
Remember that any assessment of data quality should be made not just of the individual data sources but also of any data that results from merging data sources. Because of inconsistencies between the sources, merged data may present problems that do not exist in the individual data sources.

Data requirements planning

- Plan which information is needed (e.g., only for given attributes, or specific additional information)
- Check if all the information needed (to solve the data mining goals) is actually available

Most of the data provided in the PDFs will be needed. The pdfs contain court opinions with detailed information on how the opinions were reached. These cases had previous rulings from lower-level courts, but were appealed. The PDFs summarize the legal justification the circuit court used for its appeal decision.

More specifically we will need all the text in each file, the court's district, the case id, and the justices that presided over the case (needed if we work with the judicial archetype data.

We will not necessarily need the identities of the plaintiffs and defendants from the cases.

All this data is available in the PDFs we were given. However, the files will need to be scraped to access all of it.

Selection criteria

- Specify selection criteria (e.g., Which attributes are necessary for the specified data mining goals? Which attributes have been identified as being irrelevant? How many attributes can we handle with the chosen techniques?)
- Select tables/files of interest
- Select data within a table/file
- Think about how long a history one should use (e.g., even if 18 months of data are available, only 12 months may be needed for the exercise)

We will primarily be dealing with the Opinion attribute and performing analysis on it to come up with complexity and polarity scores. We most likely will not be using the CaseID, CourtID, and Year for these scores, although we may use them to organize the data in different ways. The modeling techniques we are using can easily handle the attributes we have.

The critical table is just the court case opinions.

As mentioned earlier, within each PDF file, we will be selecting the data of the Court ID, Case ID, Year, and the Opinion.

We will be using three years (2022, 2021, and 2020) of history to construct these metrics and to create, validate, and test our model.

Over 20 years of data was available for each court. These 3 years were selected for a couple of reasons. First, we and Professor Waltenburg are interested in more recent behavior within the Court of Appeals. Second, Professor Waltenburg himself suggested using 3 years of data. Lastly, we did some quick calculations and found that if all 20 years were used, we would have around 250,000 total PDFs we would need to handle. This would push the limits of our data storage methods and cause computation time to increase significantly. If given access to much better hardware and more time one could expand this project to encompass all the data.

Be aware that data collected from different sources may give rise to quality problems when merged (e.g., address files merged with a customer database may show inconsistencies of format, invalidity of data, etc.).

Insertion of data

- If the data contain free text entries, do we need to encode them for modeling or do we want to group specific entries?
- How can missing attributes be acquired?
- How can we best extract the data?

All the data are free text entries—several pages for each court case. After using batch downloader to download all the pdfs from www.govinfo.gov/app/collection/uscourts/appellate/ ,

We can utilize the python library “pdfplumber” to extract text from a pdf. We then can loop through each file and convert the text into a string. This will allow for much faster manipulation and eliminate the need to extract data every single time. We can then store each of the strings in a pandas data frame which will have four attributes, Court_Id, Case_Id, Year, and Opinion. We will then have our data in a more manageable format, and we can easily take subsets to check for model stability.

We will need to encode each entry for modeling while performing our text analysis. There are several ways to encode our text data. One of the most popular methods is TF-IDF encoding. However, since the number of unique words does not necessarily correlate to greater complexity and we do not want to potentially skew our data in any one direction, we are looking at other options. These other options include Word2vec and Lda2vec and Bag of Words (BOW).

Remember that some knowledge about the data may be available from non-electronic sources (e.g., from people, printed text, etc.).
Remember that it may be necessary to preprocess the data (time-series data, weighted averages, etc.).

2. DESCRIBE DATA

Examine the "gross" properties of the acquired data and report on the results.
Describe the data that has been acquired, including the format of the data, the quantity of the data (e.g., the number of records and fields within each table), the identities of the fields, and any other surface features that have been discovered.

Volumetric analysis of data

- Identify data [Optional] and method of capture
- Access data sources
- Use statistical analyses if appropriate
- Report tables and their relations
- Check data volume [Optional], number of multiples, complexity
- Note if the data contain free text entries

We have one main data table:
1) Court Case table:
1. CourtID: The number corresponding to the Appellate Court (12 for D.C.)
2. CaseID: The number assigned to a given case in each court
3. Year: The year in which the case occurred
4. Opinion: The full text of the judge’s opinion on the case

There are three years of data available from each of the 11 Court of Appeals and the D.C. Court of Appeals.

Unique CaseID by CourtID:

Court 1:	985
Court 2:	834
Court 3:	2633
Court 4:	5777

Court 5: 8345
Court 6: 3410
Court 7: 2060
Court 8: 2488
Court 9: 1281
Court 10: 3432
Court 11: 4071
Court 12 (DC): 988

36,304 Unique cases in total

Attribute types and values

- Check accessibility and availability of attributes
- Check attribute types (numeric, symbolic, taxonomy, etc.)
- Check attribute value ranges
- Analyze attribute correlations
- Understand the meaning of each attribute and attribute value in business terms
- [Optional] For each attribute, compute basic statistics (e.g., compute distribution, average, max, min, standard deviation, variance, mode, skewness, etc.)
- [Optional] Analyze basic statistics and relate the results to their meaning in business terms
- Decide if the attribute is relevant for the specific data mining goal
- [Optional] Determine if the attribute meaning is used consistently
- Interview domain experts to obtain their opinion of attribute relevance
- Decide if it is necessary to balance the data (based on the modeling techniques to be used)

The type and range of all attributes are as follows:
1. CourtID: numeric, from 1-12.
2. CaseID: string, representing a unique ID.
3. Year: numeric, 2020-2022.
4. Opinion: string, no range

Keys

- Analyze key relationships
- Check amount of overlaps of key attribute values across tables

The CaseID is used as a primary key in our table.

Review assumptions/goals

- Update list of assumptions, if necessary

We are assuming that the opinions are written by panels of 3-5 judges.

Additionally, we are assuming that each PDF is structured in a similar manner.

We also assume that the entirety of the complexity of the decision of the case is present in the decision documents. Some are many pages long while others are only a few. We assume this is indicative of a more complex decision and not due to time constraints on the court/intentional omission of detail. While other measures of complexity will likely be more accurate, the number of pages or total words serves as a good proxy.

3. EXPLORE DATA

This task tackles the data mining questions that can be addressed using querying, visualization, and reporting techniques. These analyses may directly address the data mining goals. However, they may also contribute to or refine the data description and quality reports, and feed into the transformation and other data preparation steps needed before further analysis can occur.
Describe the results of this task, including first findings or initial hypotheses and their impact on the remainder of the project. The report may also include graphs and plots that indicate data characteristics or point to interesting data subsets worthy of further examination.

Data exploration

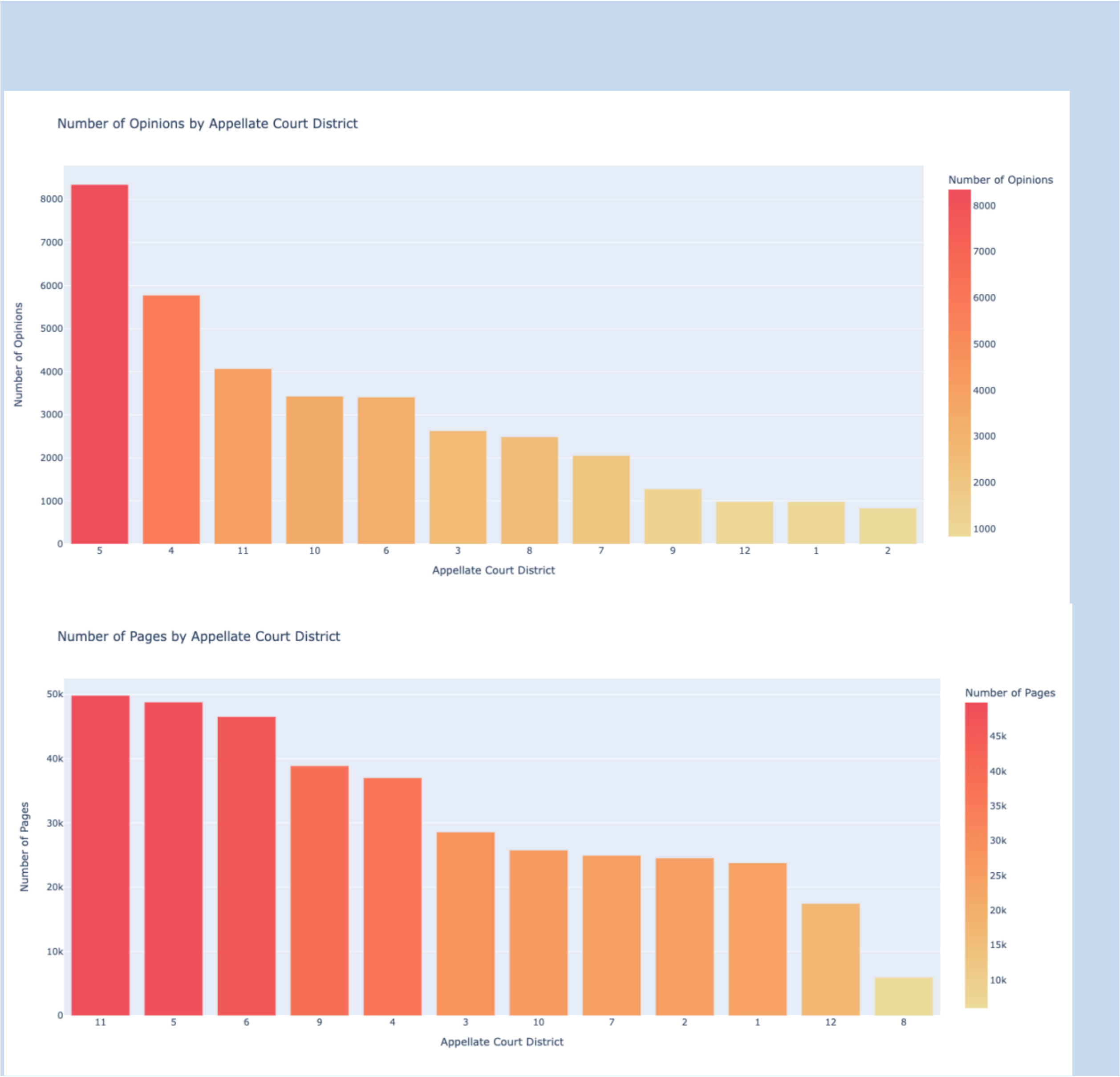
- Analyze properties of interesting attributes in detail (e.g., basic statistics, interesting sub-populations)
- [Optional] Identify characteristics of sub-populations

As you can see below the number of files per court varies significantly. You can also see that there is a significant difference in the number of pages per opinion by district. Showing that different courts behave differently by district.

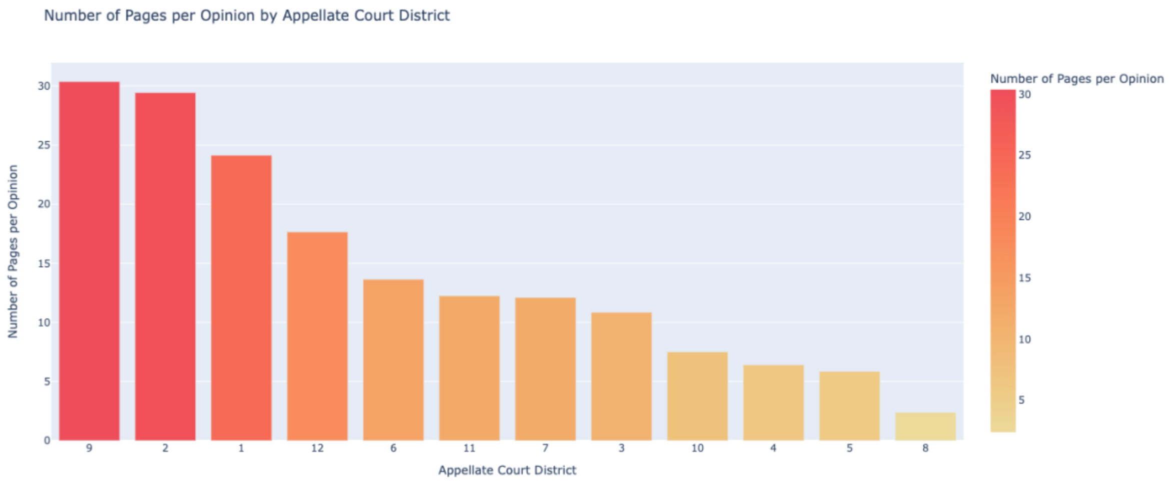
These differences could be due to different districts having different populations or having more justices which would allow for more cases to be decided on and more justices to preside over each case. We are planning to discuss this with Professor Waltenburg to see if he has any input for the discrepancies in the number of court filings and the length of them

Note: After digging further we discovered that some of the opinions are only a single page. A majority of these are simple listings that tell you an opinion was published. These provide zero relevant information and can be removed. ~ 60% of court 8’s data is made up of these files which caused a severe drop off in # of pages per opinion.

These simple files can be found in every court in varying amounts. The number for the other courts can be found in Sec. 4.2 as well as a further explanation of these irregular files.


Number of Opinions by Appellate Court District

Number of Pages by Appellate Court District



Form suppositions for future analysis

- Consider and evaluate information and findings in Section 2: Describe Data
- Form a hypothesis and identify actions
- [Optional] Transform the hypothesis into a data mining goal, if possible
- [Optional] Clarify data mining goals or make them more precise. A "blind" search is not necessarily useless, but a more directed search toward business objectives is preferable.
- Perform basic analysis to verify the hypothesis

Based on our initial data exploration and analysis of pages per opinion. It would be interesting to see if after our model has been applied if court 9 is the most complex and 8 is the least.

To test this, we would simply aggregate the complexity scores of all the opinions of each court and compare them.

4. VERIFY DATA QUALITY

Examine the quality of the data, addressing questions such as: Is the data complete (does it cover all the cases required)? Is it correct or does it contain errors? If there are errors, how common are they? Are there missing values in the data? If so, how are they represented, where do they occur, and how common are they?
List the results of the data quality verification; if there are quality problems, list possible solutions.

- Identify special values and catalog their meaning

Our special values are what we call “unnecessary court fillings” these are 1-page documents that state an opinion was published and that you can view it on a separate page.

These entries should be removed from our final data set as they provide no information and could skew results.

However, we will need to discuss the impact of removing them as court 8 is made up of 60% of these “unnecessary court fillings”, whereas the other courts may have one or two of these entries.

As stated earlier the specific number of “unnecessary court fillings” for the other courts can be found in the next section

There are also some 1–2-page documents that fix minor grammar issues in other rulings. These are fairly common in Court 1 and are identified with a “-1” at the end of the filename. However other courts may store these opinions differently or not store them at all. To fix this problem and the one above, we could filter out opinions that are below a page threshold (with caution to make sure we aren’t disregarding any relevant new information).

Review keys, attributes

- Check coverage (e.g., whether all possible values are represented)
- Check keys
- Verify that the meanings of attributes and contained values fit together
- Identify missing attributes and blank fields
- [Optional] Establish the meaning of missing data
- Check for attributes with different values that have similar meanings (e.g., low fat, diet)
- Check spelling and format of values (e.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter)
- Check for deviations, and decide whether a deviation is "noise" or may indicate an interesting phenomenon
- Check for plausibility of values, (e.g., all fields having the same or nearly the same values)

Many of the court documents contain limited information, such files are usually just one page file that contains details like: fix previous typo, reupload of previous document and denial of rehearing. We will confirm with our internal sponsor to see if they are useful to put into the final data. This is a very critical problem we should solve because some courts like Court 8 have 60% of the file is made up of this kind of file.

Court number	One page files percentage
1	15.04%
2	0.51%
3	0.37%
4	0.65%
5	0.93%
6	1.09%
7	0.98%
8	62.40%
9	0.47%
10	19.22%
11	1.01%

12	1.57%

Formatting and misspelling will be an issue in the next step where we perform sentiment analysis, since this data very heavily depends on text files. We may search for techniques that group similar words/phrases together to avoid misspellings.

There is no "noise" data that we have identified except the documents that are only one page.

Review any attributes that give answers that conflict with common sense (e.g., teenagers with high income levels).
Use visualization plots, histograms, etc. to reveal inconsistencies in the data.

Data quality in flat files

- If data are stored in flat files, check which delimiter is used and whether it is used consistently within all attributes
- If data are stored in flat files, check the number of fields in each record to see if they coincide

The format of the text of the document is different for each court. The beginning of the actual opinion in each document is usually identifiable by the phrase "Per Curium", however the punctuation and capitalization of the phrase varies across courts. In documents where this phrase isn't present, the start of the opinion is denoted by "JUDGE NAME, Circuit Judge:" at the start of the paragraph. Again, punctuation and capitalization vary.

The names of the judges on the panel are denoted by "Before Name, Name, and Name, Circuit Judges". This is consistent across all courts. Any suffixes that a judge may have in their name are also separated by commas, so we will need to write in exceptions for this in our code.

Noise and inconsistencies between sources

- Check consistencies and redundancies between different sources
- Plan for dealing with noise
- Detect the type of noise and which attributes are affected

One issue with the data is that some documents are exact duplicates of another. From examples in Court 1, it appears that whatever opinions are duplicates will also contain the case numbers of the other duplicates near the top of the document. To fix this issue, we will need to check for other case numbers within the text of each document, and then cross check that the opinions of the cases are indeed duplicates. This will probably take a lot of processing power, but it will be necessary as a "support vector" opinion with many duplicates could greatly skew our results.

Remember that it may be necessary to exclude some data since they do not exhibit either positive or negative behavior (e.g., to check on customers' loan behavior, exclude all those who have never borrowed, do not finance a home mortgage, those whose mortgage is nearing maturity, etc.).

Review whether assumptions are valid or not, given the current information on data and business knowledge.

Submit this document on Brightspace, as a Word document (PDF or other formats are not allowed). Write only on the blue areas. Do not change or remove any text outside of the blue areas.

Report the amount of effort that each team member put into this assignment. List the names of each of the team members (including those who did not work, if any) and their percentage of effort (from 0% to 100%). In a team where everybody made roughly the same effort, I expect to see 100% for all.

Spencer	100%
Greg	100%
Josh	100%
Jim	100%
Stephen	100%
Jeremy	100%

Here we consider dataset(s) produced by the data preparation phase, used for modeling or for the major analysis work of the project.

1. SELECT DATA

Decide on the data to be used for analysis. Criteria include relevance to the data mining goals, quality, and technical constraints such as limits on data volume or data types.

1.1. Rationale for inclusion/exclusion

List the data to be used/excluded and the reasons for these decisions.

- Collect appropriate additional data (from different sources—in-house as well as externally)
- Perform significance and correlation tests to decide if fields should be included
- Reconsider Data Selection Criteria (See "Data Understanding / Describe Data") in light of experiences of data quality and data exploration (i.e., may wish include/exclude other sets of data)
- Reconsider Data Selection Criteria (See "Data Understanding / Describe Data") in light of experience of modeling (i.e., model assessment may show that other datasets are needed)
- Select different data subsets (e.g., different attributes, only data which meet certain conditions)
- Consider the use of sampling techniques (e.g., A quick solution may involve splitting test and training datasets or reducing the size of the test dataset, if the tool cannot handle the full dataset. It may also be useful to have weighted samples to give different importance to different attributes or different values of the same attribute.)
- Document the rationale for inclusion/exclusion
- Check available techniques for sampling data

Just like in the Data Understanding Report, our main data we will be using will be the Opinion attribute. Again, the CaseID and CourtID attributes will not be directly involved in modeling, however, they may be used in subsetting the data. We are also adding binary attributes for each of the judges in each court. These attributes, however, will not be for our use, but to make the integration of our data with Professor Waltenburg's other research group a much easier task.

We decided to take a random sample of 500 cases from each of the 12 courts.

The sampling method used was to simply select 1000 files at random, without replacement, using NumPy's random.choice function. We then cleaned the files and threw out any that were deemed unnecessary until we reached a count of 500 cleaned and viable files. These were put into a panda's data frame and stored in a csv file.

The reason for the sampling came down to two primary factors. First, to save computational time as completing the scraping and processing would take more than 20 hours on the chosen hardware. Second, we have a very unbalanced data set with some courts containing 900 samples and others over 8,000. We worried that if we were to apply NLP models to the entire corpus the overabundance of data from the courts with more cases could skew results.

One could potentially increase the number of samples based on hardware and available time. And depending on how our model is handling the data or if we deem more data is necessary for modeling, we can add more samples to our total data set.

Based on Data Selection Criteria, decide if one or more attributes are more important than others and weight the attributes accordingly. Decide, based on the context (i.e., application, tool, etc.), how to handle the weighting.

2. CLEAN DATA

Raise the data quality to the level required by the selected analysis techniques. This may involve the selection of clean subsets of the data, the insertion of suitable defaults, or more ambitious techniques such as the estimation of missing data by modeling.
Describe the decisions and actions that were taken to address the data quality problems reported during the Verify Data Quality Task. If the data are to be used in the data mining exercise, the report should address outstanding data quality issues and what possible effect this could have on the results.

- Reconsider how to deal with any observed type of noise
- Correct, remove, or ignore noise
- Decide how to deal with special values and their meaning. The area of special values can give rise to many strange results and should be carefully examined. Examples of special values could arise through taking results of a survey where some questions were not asked or not answered. This might result in a value of 99 for unknown data. For example, 99 for marital status or political affiliation. Special values could also arise when data is truncated—e.g., 00 for 100-year-old people or all cars with 100,000 km on the odometer.
- Reconsider Data Selection Criteria (See "Data Understanding / Describe Data") in light of experiences of data cleaning (i.e., you may wish to include/exclude other sets of data).

As mentioned in Section 4 (Verify Data Quality) of the Data Understanding report, our data contained "unnecessary data fillings". It also contained some 1-2-page documents that fix minor grammar issues in other rulings. These were common in Court 1 and are identified with a "-1" at the end of the filename. To fix both problems we decided to have a page cutoff of one as well as remove files with a tag of one. We made sure to check to see if we would be removing any essential information. We decided, after consideration, to go ahead with this plan as the amount of data lost would be negligible. One or two files in total.

For the next step, we must further clean the data in the Opinion feature by removing stop words as well as frequently used words that appear in every document, like Appeal, Court, Judge, Decision, etc. These provide no value to the meaning of the text and can be removed. We have not removed them yet as we are still discussing several types of models and how many frequently used words should be left out. We will eventually remove them after we have decided on our modeling techniques and after having a discussion with Professor Waltenburg.

Remember that some fields may be irrelevant to the data mining goals and, therefore, noise in those fields has no significance. However, if noise is ignored for these reasons, it should be fully documented as the circumstances may change later.

3. CONSTRUCT DATA

This task includes constructive data preparation operations such as the production of derived attributes, complete new records, or transformed values for existing attributes.

- Check available construction mechanisms with the list of tools suggested for the project
- Decide whether it is best to perform the construction inside the tool or outside (i.e., which is more efficient, exact, repeatable)
- Reconsider Data Selection Criteria (See "Data Understanding / Describe Data") in light of experiences of data construction (i.e., you may wish include/exclude other sets of data)

Python is our preferred program for extraction and construction for two reasons. First, it has that library pdfplumber which allowed us to scrape the text out of our pdf files. Second, it has the pandas library which is robust and easy to use when it comes to data transformation and merging. Any language with similar libraries or functionalities would also be viable.

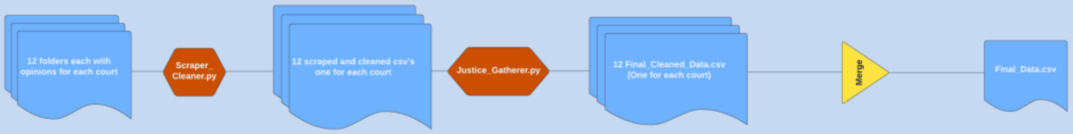
The following takes in 12 folders of pdfs each of variable size.

Scraper Cleaner.py takes in a file, scrapes the text data. It also removes data with a –1 file tag and any files under 2 pages.

The following takes in 12 csv files
Justice_Gatherer.py: Extract the justices that presided over the case and puts them into a new column.

Merge.py: Concatenates all the data into one pandas data frame that is then stored in the Final_data.csv file

The task diagram for preprocessing the data is the following:


12 folders each with opinions for each court
Scraper_ Cleaner.py
12 scraped and cleaned csv's one for each court
Justice_Gatherer.py
12 Final_Cleaned_Data.csv (One for each court)
Merge
Final_Data.csv

3.1. Derived attributes

Derived attributes are new attributes that are constructed from one or more existing attributes in the same record. An example might be: area = length * width.
Why should we need to construct derived attributes during the course of a data mining investigation? It should not be thought that only data from databases or other sources should be used in constructing a model. Derived attributes might be constructed because:

- Background knowledge convinces us that some fact is important and ought to be represented although we have no attribute currently to represent it
- The modeling algorithm in use handles only certain types of data—for example we are using linear regression and we suspect that there are certain non-linearities that will be not be included in the model
- The outcome of the modeling phase suggests that certain facts are not being covered

Derived attributes

- Decide if any attribute should be normalized (e.g., when using a clustering algorithm with age and income, in certain currencies, the income will dominate)
- [Optional] Consider adding new information on the relevant importance of attributes by adding new attributes (for example, attribute weights, weighted normalization)
- How can missing attributes be constructed or imputed? [Decide type of construction (e.g., aggregate, average, induction).]
- Add new attributes to the accessed data

We are creating a new attribute for each judge in the form of a binary attribute to determine which judges are involved in which cases. This is less for our use, but more so we will be able to integrate the results of the complexity/polarity of the opinions with the judge archetype data from Professor Waltenburg's other research group. This is known as "one-hot" encoding and is common in machine learning models.

We will use word2vec to transform the document text into vectors as input for LDA. We are considering using document length as a measure of complexity with the assumption that longer documents are more complex. We also may use lexile score or reading level as a measure of complexity. We are waiting on input from our domain expert, Prof. Waltenberg, to see if those are good indicators of complexity.

Before adding Derived Attributes, try to determine if and how they ease the model process or facilitate the modeling algorithm. Perhaps "income per person" is a better/easier attribute to use than "income per household." Do not derive attributes simply to reduce the number of input attributes.
Another type of derived attribute is the single-attribute transformation, usually performed to fit the needs of the modeling tools.

Single-attribute transformations

- Specify necessary transformation steps in terms of available transformation facilities (for example, change a binning of a numeric attribute)
- Perform transformation steps

We plan to use a pretrained model from a Python package such as gensim to vectorize the text. We plan to remove certain words that we expect are common across all documents and are not indicative of complexity, such as ‘court’ and ‘justice’. We can simply create a list and have our LDA model ignore those words.

Transformations may be necessary to change ranges to symbolic fields (e.g., ages to age ranges) or symbolic fields ("definitely yes," "yes," "don't know," "no") to numeric values. Modeling tools or algorithms often require them.

3.2. Generated records

Generated records are completely new records, which add new knowledge or represent new data that is not otherwise represented (e.g., having segmented the data, it may be useful to generate a record to represent the prototypical member of each segment for further processing).

- Check for available techniques if needed (e.g., mechanisms to construct prototypes for each segment of segmented data).

Not Applicable

4. INTEGRATE DATA

These are methods for combining information from multiple tables or other information sources to create new records or values.
Merging tables refers to joining together two or more tables that have different information about the same objects. At this stage, it may also be advisable to generate new records. It may also be recommended to generate aggregate values.
Aggregation refers to operations where new values are computed by summarizing information from multiple records and/or tables.

- Check if integration facilities are able to integrate the input sources as required

- Integrate sources and store results
- Reconsider Data Selection Criteria (See "Data Understanding / Describe Data") in light of experiences of data integration (i.e., you may wish to include/exclude other sets of data)

Because we only have one table, we don't need to do any joins among tables. Our data relies more on the single attribute of the court opinion for now; we will further process this attribute only to calculate the complexity score of each case. If/when our internal sponsor gives us the diversity score of the court panel then will find the correlation between both scores.

Remember that some knowledge may be contained in non-electronic format.

5. FORMAT DATA

Formatting transformations refers primarily to syntactic modifications made to the data that do not change its meaning, but might be required by the modeling tool.
Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict.

Rearranging attributes

- Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict.

No such attribute arrangement is required as the tools planned for use take the input and outcome fields as separate parameters.

Reordering records

- It might be important to change the order of the records in the dataset. Perhaps the modeling tool requires that the records be sorted according to the value of the outcome attribute.

No records reordering is required. Instead, it is recommended to shuffle the data randomly.

Reformatted within-value

- These are purely syntactic changes made to satisfy the requirements of the specific modeling tool
- Reconsider Data Selection Criteria (See "Data Understanding / Describe Data") in light of experiences of data cleaning (i.e., you may wish to include/exclude other sets of data)

The only non-numeric value we would need to change is the Opinion feature. This would be encoded into multidimensional numeric attributes. The exact encoding methodology has not been selected yet.

However, it may be necessary to tokenize the decision text. It depends on the pretrained model chosen for word2vec.

6. DATASET DESCRIPTION

Provide a general description of the final dataset (for instance, in terms of number of number of samples and number of features).

The final dataset has 6,000 samples, 2 numeric features CaseID and CourtID, 1 text feature Opinions, one-hot encoded justices, and after we complete the modeling stage will have 2 “score” columns. Those being complexity and polarity.

Submit a ZIP file on Brightspace containing:

- **This document, as a Word document (PDF or other formats are not allowed). Write only on the blue areas. Do not change or remove any text outside of the blue areas.**
- **The source code for the Data Preparation phase and for the Modeling phase. DO NOT include the dataset(s).**

Report the amount of effort that each team member put into this assignment. List the names of each of the team members (including those who did not work, if any) and their percentage of effort (from 0% to 100%). In a team where everybody made roughly the same effort, I expect to see 100% for all.

Jeremy 100%
Spencer 100%
Josh 100%
Greg 100%
Stephen 100%

1. SELECT MODELING TECHNIQUE

As the first step in modeling, select the actual initial modeling technique. If multiple techniques are to be applied, perform this task separately for each technique.
Remember that not all tools and techniques are applicable to each and every task. For certain problems, only some techniques are appropriate. It may be that only one tool or technique is available to solve the problem at hand—and that the tool may not be absolutely the best, from a technical standpoint.m

1.1. Modeling technique

Record the actual modeling technique that is used.

- Decide on appropriate technique for exercise, bearing in mind the tool selected.

To model complexity, we needed to use a technique that was able to input a document and determine which topic(s) are covered in the document. From that, we can compare the topics covered in the decision to topics that align with specific political/constitutional ideologies. We originally were going to use word2vec but realized that would not be due to our LDA model requiring a BOW style vector. However, after further research we found a model, lda2vec that combines both. This model is much newer/experimental, so we decided to stick with more established model. We also, saw in academic research that the results aren't significantly better in comparison to a normal LDA model with a BOW vector.

To model polarity/subjectivity, we needed to use a sentiment analysis modeling technique that would determine how positive/negative the sentiment of the text is, giving us polarity and subjectivity scores. We decided to use TextBlob for these scores. We chose TextBlob over some alternatives such as Vader and Flair because TextBlob scores polarity in a range [-1, 1], which we believe will be easiest for Prof. Waltenburg to use as a variable in his study and because we thought subjectivity might provide additional insight in the sense that more subjective decisions include more opinions and could therefore be more polarizing. Additionally, we decided against Vader because it ignores the context in which words are used. We decided against Flair because it is trained on IMDB data (movies) while TextBlob is trained on a more normal lexicon using everyday words so to speak.

Subjectivity can be defined as how opinionated the text is, which we thought may be an interesting parameter to explore in addition to polarity. We will provide Prof. Waltenburg with the scores if he would like. Calculating it had no effect on the polarity scores, so if he does not want those scores, we will not provide them, and our analysis was completed as if we hadn't calculated them.

1.2. Modeling assumptions

Many modeling techniques make specific assumptions about the data.

- Define any built-in assumptions made by the technique about the data (e.g., quality, format, distribution)
- Compare these assumptions with those in the Section 2 (Describe Data) of the Data Understanding report
- Make sure that these assumptions hold and go back to the Data Preparation report, if necessary

One assumption we made regarding complexity is that length of document is a proxy for complexity. It is not the only measure, but since we do not have a training data set that shows what types of documents are more and less complex, we must subtly introduce our own assumptions about the definition of the word. Professor Waltenburg confirmed that this assumption is valid.

Another assumption we made was that one-page documents are most likely "addendum" documents and are not actual decisions and were disregarded.

LDA ASSUMPTIONS:

1) The semantic content of a document is composed by combining one or more terms from one or more topics.

2) Certain terms are ambiguous, belonging to more than one topic, with different probability. However, in a document, the accompanying presence of specific neighboring terms (which belong to only one topic) will disambiguate their usage.

3) Most documents will contain only a relatively small number of topics, and individual topics will occur with differing frequencies. That is, they have a probability distribution, so that a given document is more likely to contain some topics than others.

4) The terms within a topic will have their own probability distributions, as within a topic some terms will be used much more frequently than others.

POLARITY/SUBJECTIVITY ASSUMPTIONS:

1) We assume that words in a legal context have a similar polarity as words in any other context. It is likely not 100% accurate since words like ‘appeal’ are more likely to have a neutral polarity in legal context, whereas it is more likely to have a positive polarity in other texts. We do not have a model trained on legal text since creating a corpus and training one would not have been feasible with the time restrictions of this project. However, we assume the contexts are close enough that it will work for our purposes.

2) We assume that the writers of the decisions convey the polarity of their decision using words that would be detected as positive or negative. If they are instructed to remain neutral in their writing, then the true polarity would not be picked up.

2) We assume that all the boilerplate text has been removed since this may add or deduct polarity from all the documents, affecting the variance. We assume this is true from our cleaning of the data.

2. GENERATE TEST DESIGN

Prior to building a model, it is necessary to define a procedure to test the model's quality and validity. For example, in supervised data mining tasks such as classification, it is common to use error rates as quality measures for data mining models. Therefore, the test design specifies that the dataset should be separated into training and test sets. The model is built on the training set and its quality estimated on the test set. Describe the intended plan for training, testing, and evaluating the models. A primary component of the plan is to decide how to divide the available dataset into training data, test data, and validation test sets.

- Check existing test designs for each data mining goal separately
- Decide on necessary steps (number of iterations, number of folds, etc.)
- Prepare data required for test

Complexity/LDA:

We were not provided with an existing test design by Professor Waltenburg, however through research we were able to find several sources creating a model similar to our own and we adapted their test designs to our model.

The following steps are taken for LDA model testing:

1) The dataset (specifically the ‘Opinion’ attribute) is transformed into a document term matrix by utilizing sklearn’s CountVectorizor.

2) The document term matrix is split into 2 parts: the “20% dataset” (for testing) and “80% dataset” (for parameter tuning and model training).

3) The parameter we are most concerned about is “n_components”. This is simply the number of topics that will be used in the model. There are several other parameters in the LDA model, however some only matter when using the ‘online’ learning_method. The batch learning_method will be used so these are not important. Lastly, the rest of the parameters are very sensitive and can cause a drastic change in results. We felt we did not have the requisite background knowledge to properly tune these parameters at this time, so we left them as their default values.

To tune the n_components parameter we will run each model 5 times with n_components = [10,15,20,25,30]. Each model needs to be run several times since LDA generates different probability distributions each time it runs. This leads to varying outputs each time. Running each parameter several times will better stabilize and ensure the results are more accurate.

3) The remaining "20% dataset" will be used to test and evaluate each model with a given n_components value by calculating the perplexity scores for each tested model. Since each n_components value is run 5 times its perplexity score is calculated for each run and then aggregated at the end. The aggregated perplexity scores will then be plotted and the n_components value with the lowest aggregated perplexity score will be selected.

POLARITY/SUBJECTIVITY:

The polarity/subjectivity is calculated using an out-of-the-box method from TextBlob. The TextBlob module (called sentiments) we are using has a default implementation called PatternAnalyzer that is already trained on a model based on the pattern Python library. Our data is unlabeled, so we are not able to inject any domain knowledge for which we would need to adjust parameters into the model used. Therefore, we did not need to do anything more to this model to train it.

3. BUILD MODEL

Run the modeling tool on the prepared dataset to create one or more models.

3.1. Parameter settings

With any modeling tool, there are often a large number of parameters that can be adjusted. List the parameters and their chosen values, along with the rationale for the choice.

- Set initial parameters
- Document reasons for choosing those values

LDA:

We cleaned the data by removing stop words, punctuation, extra spacing, numbers, converting all words to lowercase, and lemmatizing each word before modeling the data using LDA. We begin with using the default n_component: value of 10:

There is no particular reasoning for this choice of n_components. We will later try to do a proper parameter tuning as described in Section 2 (Generate test design) and fully performed in Section 4.2 (Revised parameter settings).

POLARITY:

Since the polarity/subjectivity model was already trained, we did not need to set parameters to find polarity and subjectivity. We simply ran TextBlob's sentiment method on the bags of words that represent the documents.

3.2. Models

Run the modeling tool on the prepared dataset to create one or more models.

- Run the selected technique on the input dataset to produce the model
- Post-process data mining results (e.g., edit rules, display trees)

LDA:

We run the following models from Scikit-Learn:
1) Latent Dirichlet Allocation (LDA)

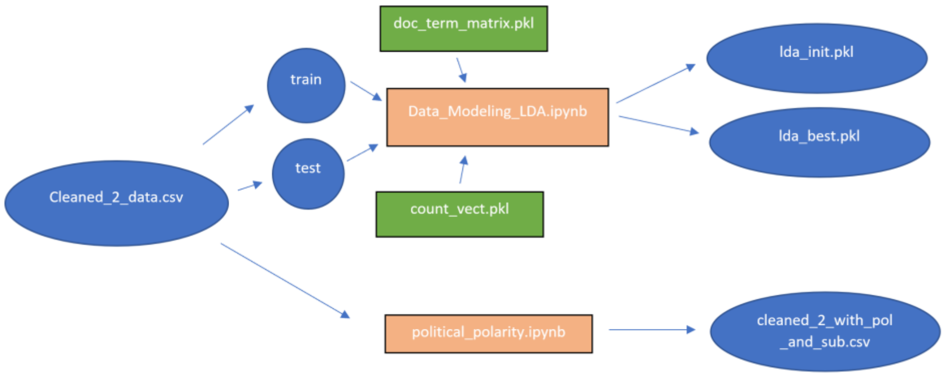
The scripts created for the above are:
Data_Modeling_LDA.ipynb

Each of the scripts above save the learnt models with the initial parameters (Section 3.1) and the best parameters found through a proper parameter tuning as described in Section 2 (Generate test design) and Section 4.2 (Revised parameter settings). We use the Pickle library for Python object serialization. Thus, the files corresponding to the saved learnt models have a .pkl extension.

POLARITY:

The code to produce the polarity/subjectivity is found in the political_polarity.ipynb file. The results are saved in the cleaned_2_with_pol_and_sub.csv file alongside the features extracted in the Cleaned_2_data.csv file.

DIAGRAM:



3.3. Model description

Describe the resulting model and assess its expected accuracy, robustness, and possible shortcomings. Report on the interpretation of the models and any difficulties encountered.

- Describe any characteristics of the current model that may be useful for the future
- Record parameter settings used to produce the model
- Give a detailed description of the model and any special features
- For rule-based models, list the rules produced, plus any assessment of per-rule or overall model accuracy and coverage
- For opaque models, list any technical information about the model (such as neural network topology) and any behavioral descriptions produced by the modeling process (such as accuracy or sensitivity)
- [Optional] Describe the model's behavior and interpretation
- [Optional] State conclusions regarding patterns in the data (if any); sometimes the model reveals important facts about the data without a separate assessment process (e.g., that the output or conclusion is duplicated in one of the inputs)

Latent Dirichlet Allocation (LDA):

LDA is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics.

Chosen (non-default) parameters:
1) LDA
    n_components = 10

In machine learning LDA is used to discover topics in a collection of documents, and then automatically classify any individual document within the collection in terms of how "relevant" it is to each of the discovered topics. A topic is considered to be a set of terms (i.e., individual words or phrases) that, taken together, suggest a shared theme.

If the document collection is sufficiently large, LDA will discover such sets of terms (i.e., topics) based upon the co-occurrence of individual terms, though the task of assigning a meaningful label to an individual topic is up to the user, and often requires specialized knowledge.

A list of modeling assumptions for LDA was discussed in section (1.2)

POLARITY:

The TextBlob sentiment analysis model we used for polarity/subjectivity by default uses the pattern Python library to calculate polarity and subjectivity for each word using a weighted averaging technique based on their own coded rules and lexicon. This lexicon doesn't contain stop words (since they inherently do not have sentiment), and each word in the lexicon has a defined part of speech, polarity, subjectivity, intensity, and confidence. The coded rules are as follows:

-The polarity has a range of [-1, 1] (-1 being negative and 1 being positive sentiment).

-For subjectivity, the score has a range of [0, 1] and the higher the score means the text is more concerned with an opinion over fact.

-The intensity parameter is defined as how a word modifies the words around it. For example, adverbs such as "very" would modify the word that follows.

-When modifiers are recognized, TextBlob ignores polarity and subjectivity scores for the modifier word, and simply uses the modifier's intensity score when calculating the polarity and subjectivity scores of the word that follows it.

-Negation words multiply polarity by –0.5, and do not affect subjectivity score at all. Additionally, when modifiers are used with negation (i.e., "not very good"), the inverse intensity score is included in the calculation for both polarity and subjectivity.

-TextBlob ignores one-letter words as well as words that it does not know anything about (i.e., not in its lexicon)

In summary, TextBlob goes along and looks for words/phrases in its own lexicon that it can assign polarity and subjectivity scores to and averages them all together for the entire text.

We rely on standard deviation, which was .074 on a range of scores [-1, 1], to assess whether there is enough variance in the polarity for Prof. Waltenburg to perform his statistical tests with meaning. We believe the variance is sufficient considering that the mean is <.004, indicating that there are a significant number of documents that lie in both sides of the positive/negative polarity. Subjectivity had an even greater standard deviation of .092 on a relatively smaller range of [0, 1] with a mean of .41. Again, the mean is close to the middle of the range of values and there is a sizable variance, indicating that there are documents on both sides of the "objective/subjective spectrum".

The fact that subjectivity scores are significantly greater than 0 gives support to the idea that polarized language is used throughout the documents. The subjectivity scores would have been much closer to 0 if the authors of the decisions had used completely neutral language.

4. ASSESS MODEL

The model should now be assessed to ensure that it meets the data mining success criteria and passes the desired test criteria. This is a purely technical assessment based on the outcome of the modeling tasks.

4.1. Model assessment

Summarize results of this task, list qualities of generated models (e.g., in terms of accuracy), and rank their quality in relation to each other.

- Evaluate results with respect to evaluation criteria
- Test result according to a test strategy (e.g.: Train and Test, Cross-validation, bootstrapping, etc.)
- Compare evaluation results and interpretation

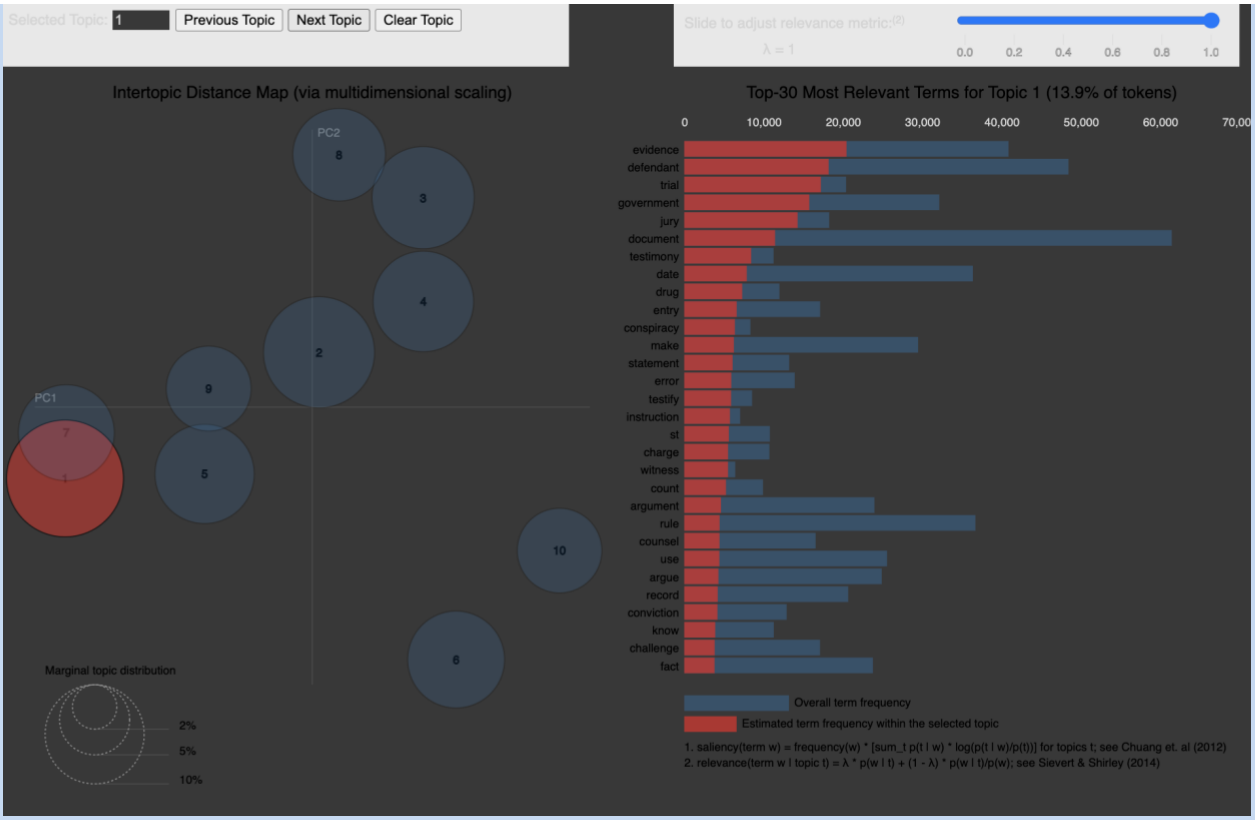
- Create ranking of results with respect to success and evaluation criteria
- Select best models
- Interpret results in business terms (as far as possible at this stage)
- [Optional] Get comments on models by domain or data experts
- Check plausibility of model
- Check effect on data mining goal
- Check model against given knowledge base to see if the discovered information is novel and useful
- Check reliability of result
- Analyze potential for deployment of each result
- If there is a verbal description of the generated model (e.g., via rules), assess the rules: Are they logical, are they feasible, are there too many or too few, do they offend common sense?
- Assess results
- Get insights into why a certain modeling technique and certain parameter settings lead to good/bad results

LDA:

We use perplexity to measure performance. Perplexity is a very weird and hard to grasp measure, but essentially perplexity is a statistical measure of how well a probability model predicts a sample. As applied to LDA, for a given value of k (number of topics), you estimate the LDA model. Then given the theoretical word distributions represented by the topics, compare that to the actual topic mixtures, or distribution of words in your documents.

As stated in section 2, the model will be run several times due to LDA generating different probability distributions each time it runs. In our case we will run it 5 times and aggregate the perplexity scores to get a better sense of how the model performed.

For LDA with n_components = 10 we got a perplexity score of 974.726. This is very hard for a human to interpret by itself, as it’s not as consumable as an accuracy score, but essentially the lower the perplexity the better. This will be used to compare models in section (4.2)



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

This is a visualization of our topic modeling. It shows the number of topics, the 30 most relevant terms in each topic, and their relevant distributions within a topic and within the entire corpus. The image above specifically shows the makeup of topic 1.

For creating our complexity measure we still have to convert these topic distributions into complexity scores. Our process is laid out in (4.2)

When looking through the generated topics everything seems to check out. Topic 7 for example seems to deal with financial cases as it is made up of terms like bank, fraud, tax, and loan.

The results are reliable since we are performing parameter tuning and have accounted for randomness by averaging our results. We could increase the sense of reliability by creating different subsets of training and testing data for each model to see how differences in training data effect the results.

The model is easily deployable, and results are quite fast. Can run locally and receive train in a few minutes. You get results in seconds.

Even though the given evaluation score is difficult to grasp. The visualization of the topics makes it far easier to grasp the end result.

PROS:

LDA is a generative model, but in text mining, it introduces a way to attach topical content to text documents. Each document is viewed as a mix of multiple distinct topics. An advantage of the LDA technique is that one does not have to know in advance what the topics will look like.

.

CONS:

Latent Dirichlet allocation when run on different datasets, LDA suffers from “order effects” i.e., different topics are generated if the order of training data is shuffled. Such order effects introduce a systematic error for any study. This error can relate to misleading results; specifically, inaccurate topic descriptions and a reduction in the efficacy of text mining classification results.

In future to balance out these cons we can introduce much more rigorous evaluation methods like creating several different training/testing subsets.

A thing to consider is that LDA doesn’t take into account context which is an issue. There is a model called lda2vec which combines LDA and the contextualization of text with word2vec. This model is still new and experimental, but future work could utilize this model for better results.

POLARITY/SUBJECTIVITY:

Since we did not build the polarity/subjectivity model, we are not assessing its validity. Our evaluation of its success will be initially based on a subjective assessment of the summary statistics, as described in 3.3. We will consult Prof. Waltenburg to see if the results match his expectations.

We can, however, assess the rules of the TextBlob default model. For the most part, they seem to be very logical and follow common sense since it basically calculates an average polarity/subjectivity score for the text by going through and looking at the scores for each word. Additionally, the scores for each word are affected by modifiers (in English, adverbs) as well as negation. There are not many rules. It could be argued that more rules could be added to better consider the context of each word.

This modeling technique should lead to good results as it gives a very simple view of whether an opinion is polarizing and subjective in the general sense. However, one could argue that the model could be better if we had a corpus/lexicon of political/judicial terms and how politically left/right each term is, although again we determined that was not feasible given the time restraints of this class.

"Lift Tables" and "Gain Tables" can be constructed to determine how well the model is predicting.

4.2. Revised parameter settings

According to the model assessment, revise parameter settings and tune them for the next run in the Build Model task. Iterate model building and assessment until you find the best model.

- Adjust parameters to produce better models.

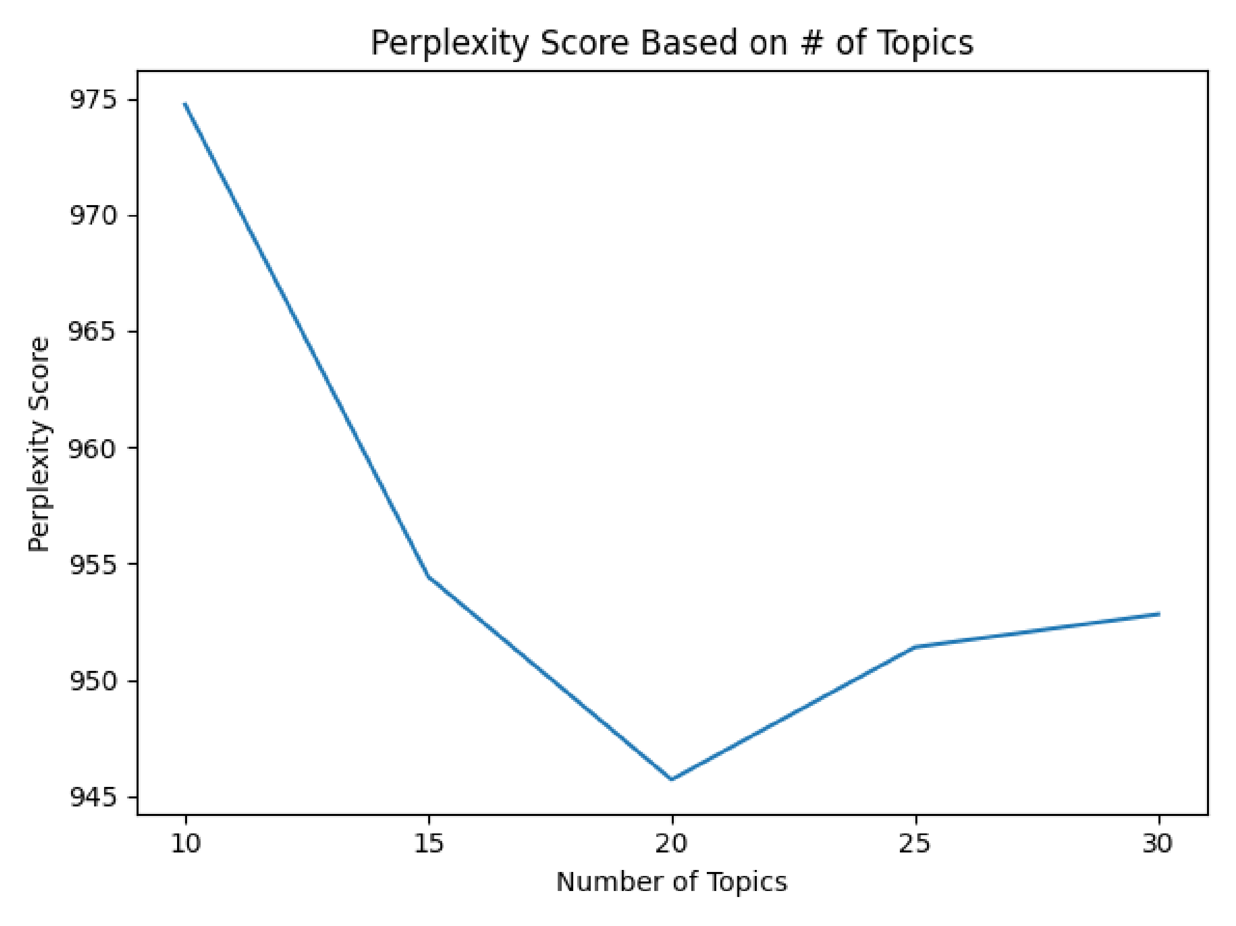
LDA:

To find the best n_components parameter:

The n_components parameter should be decided from the set = {10, 15, 20, 25, 30}. (Can be expanded, but we didn't find it necessary.)

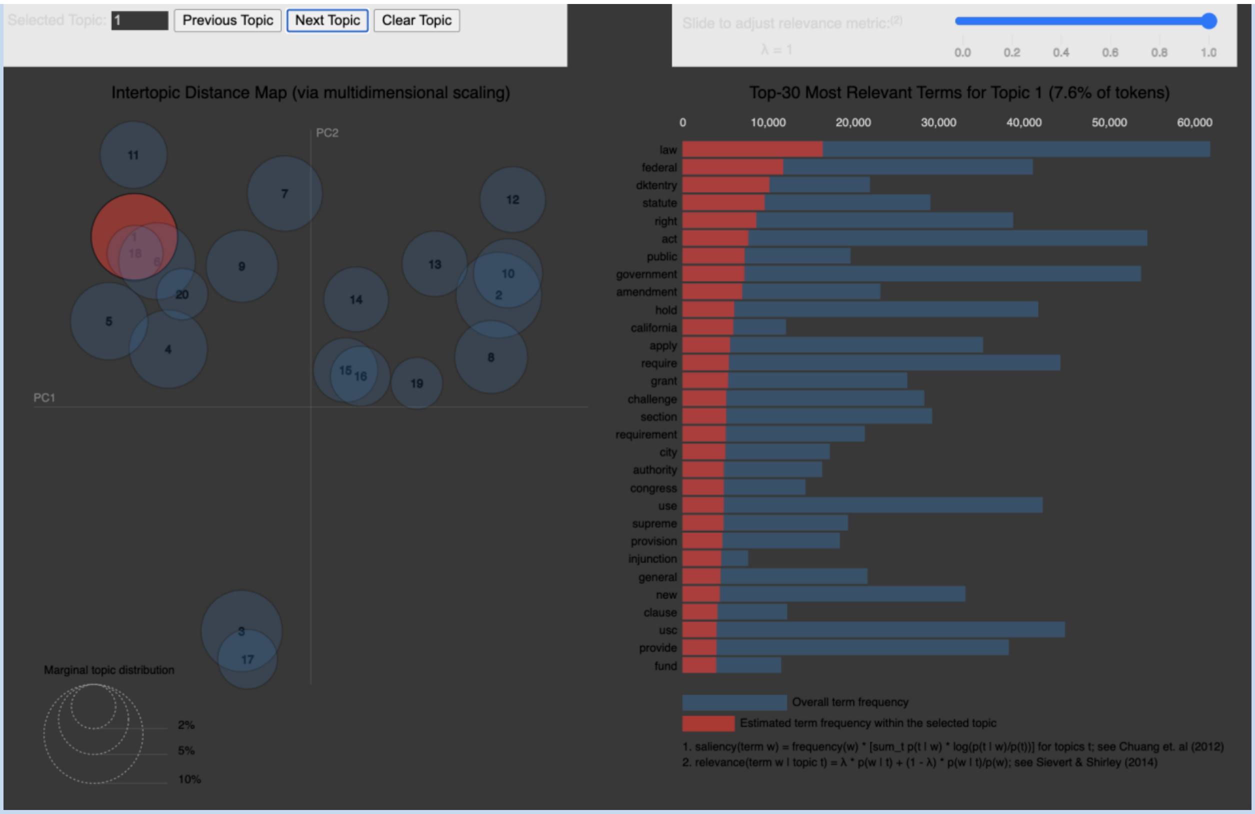
We run the same test for perplexity as laid out in both (2.1) & (4.1). This time we run each model 5 for each value in the n_components set.

After running through this test, we found that n_components = 20 is the best.



As we can see our best performing model in terms of perplexity is when n_components = 20

The topics generated by the best LDA model:



Selected Topic: 1

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

The next step which we have already begun implementing is to convert these topic modeling results to entropy scores. To do this we will gather the topic % for each document EX) doc1 {Topic 1: .2, Topic 2: .7, Topic 3: .1}. Then we would calculate entropy on this set to see how balanced the result is. The more balanced the more topics are covered. The log base will be equal to the number of topics. In the Example it would be 3. For our model it would be 20. Changing the log base ensures all values are between 0-1 and that a perfectly balanced set will be given a score of 1.

POLARITY/SUBJECTIVITY:

We do not have parameters to adjust for this part of the analysis.

Submit this document on Brightspace, as a Word document (PDF or other formats are not allowed). Write only on the blue areas. Do not change or remove any text outside of the blue areas.

Report the amount of effort that each team member put into this assignment. List the names of each of the team members (including those who did not work, if any) and their percentage of effort (from 0% to 100%). In a team where everybody made roughly the same effort, I expect to see 100% for all.

Spencer	100%
Greg	100%
Jeremy	100%
Jim	100%
Stephen	100%
Josh	100%

1. EVALUATE RESULTS

This step assesses the degree to which the model meets the business objectives, and seeks to determine if there is some business reason why this model is deficient. Another option is to test the model(s) on test applications in the real application, if time and budget constraints permit.
Moreover, evaluation also assesses other generated data mining results. Data mining results cover models that are related to the original business objectives and all other findings. Some are related to the original business objectives while others might unveil additional challenges, information, or hints for future directions.

For context, state the business objectives and data mining goals (Sections 1.2, 1.3, 3.1, 3.2 of the Business Understanding report).

Business Objectives:

Develop new criteria for assessing the complexity and polarity of a court ruling. Furthermore, to obtain a better grasp of the functions of the court. In a broader sense, it enables people to comprehend how a small group influences the ultimate verdict.

The primary focus of the research is to demonstrate that more diverse panels produce more complicated and divisive conclusions. Professor Waltenburg is reticent on the value of more varied panels and does not want this research to be any kind of normative statement, rather it's just a facet about judiciary systems he finds interesting. This research will hopefully offer us a better comprehension of how small group pressures influence collective decision-making output, which will be the first step toward drawing conclusions.

Our main objective is to develop reliable measures of the complexity and polarity of court rulings. Once this is completed, we will be able to concentrate on establishing meaningful correlations between these measures and the variety of the panels. With this knowledge, we may acquire a more complete and in-depth understanding of the decisions made, as well as a better grasp of how small group pressures impact the outcome of collective decisions. Furthermore, if we can demonstrate that the correlations between our variables are significant, we may be able to attempt to forecast the effects of future actions.

Business Success Criteria:

This project has no predefined measures of success. Ideally, this research would be useful to those interested in court decisions and ruling.

In general, we want to devise and decide on the best methods for measuring the polarity and complexity of court decisions. We'd like to be reasonably accurate in determining these measurements. We want Prof. Waltenburg to assess our accuracy in these measurements because they are highly subjective. Furthermore, we want to see if there is a link between the panel's polarity/complexity and diversity. Because Prof. Waltenburg is providing us with data on the diversity of the panels, our success should be slightly more objective, and will be dependent on our ability to accurately determine meaningful relationships between the variables. We would prefer to find relationships with a high correlation.

Data Mining Goals:

This data mining project includes a variety of problem types. The primary variables being investigated include the complexity and polarity of court opinions, as well as the diversity of panels. Complexity and polarity are difficult problems to solve, but by using NLP approaches such as topic modeling and sentiment analysis, we can characterize each opinion and try to determine its complexity and polarity. Professor Waltenburg should give us the panel makeup in order for us to construct a statistic that quantifies how varied a panel is.

Data Mining Success Criteria:

Because Prof. Waltenburg is a domain expert, model accuracy and performance evaluation will necessitate their input. We can offer them random samples of input and ask them to predict the outcome, which we can then compare to the model's results. Because the project's outputs are designed to quantify and describe a phenomenon, less complexity will be preferred. It is critical that the results are understandable to people who are not professionals in data mining. As a result, we will prefer NLP models with high degrees of comprehensibility that we can explain to the team. Before implementing it, we will validate with the team that it is properly understood.

1.1. Assessment of data mining results with respect to business success criteria

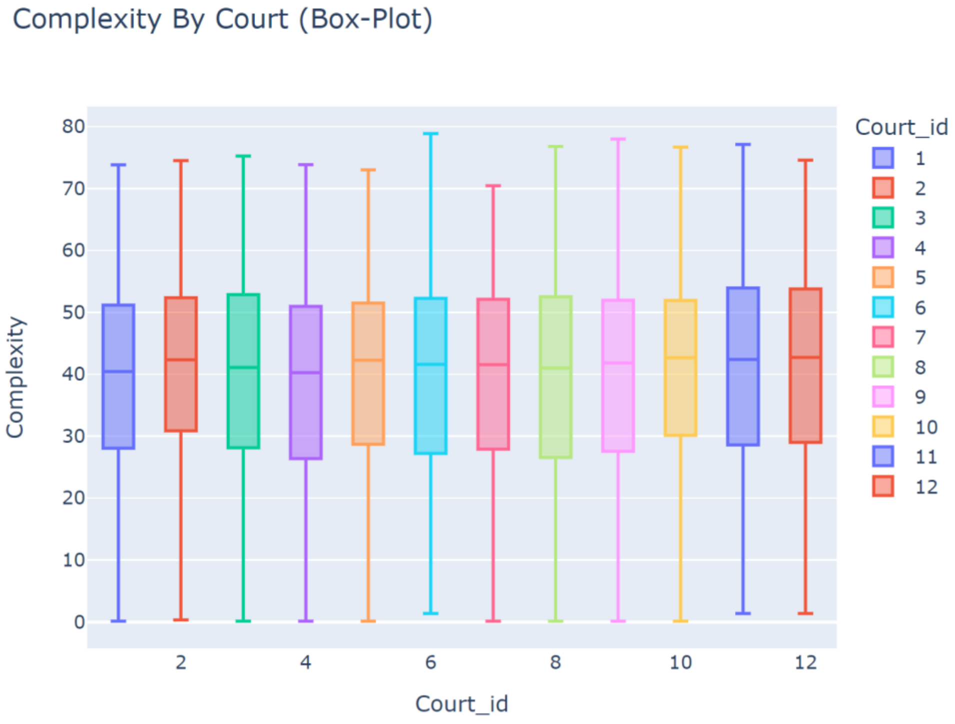
Summarize assessment results in terms of business success criteria, including a final statement related to whether the project already meets the initial business objectives.

- Understand the data mining results
- Interpret the results in terms of the application
- Check effect on data mining goal
- Check the data mining result against the given knowledge base to see if the discovered information is novel and useful

- Evaluate and assess results with respect to business success criteria (i.e., has the project achieved the original Business Objectives)
- [Optional] Compare evaluation results and interpretation
- Rank results with respect to business success criteria
- [Optional] Check effect of result on initial application goal
- [Optional] Determine if there are new business objectives to be addressed later in the project, or in new projects
- [Optional] State recommendations for future data mining projects

For our polarity measure we used the perplexity score to evaluate the best performing model and how many topics the model should create for each court decision. We were able to determine that the best performing model that had the lowest perplexity score of roughly 946 was the model that created 20 topics. The overall mean for the complexity scores across all documents and courts was 39.38 with a standard deviation of 17.01 showing that there is a wide range of levels of complexities among the different courts and decisions. Additionally, the maximum complexity score was 78.86 while the minimum complexity score was 0.11.

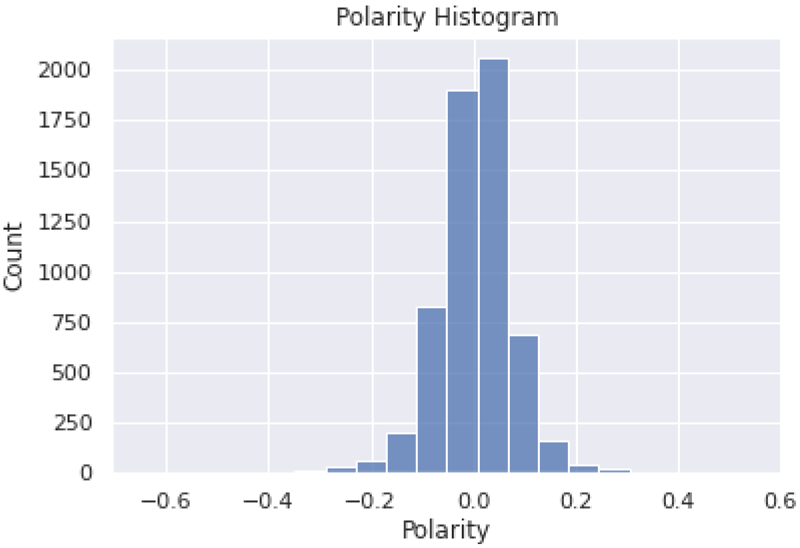
Here we can see the average complexity score among all decisions for each court:

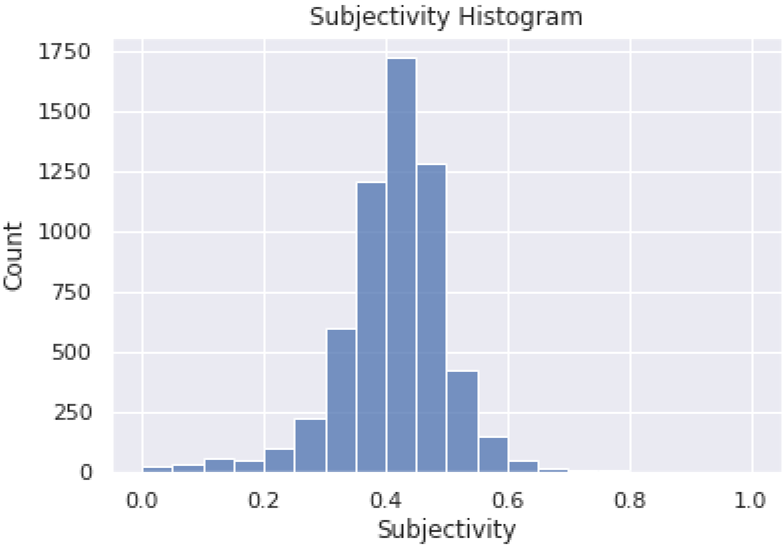


For polarity and subjectivity, our results also covered a wide range. For polarity, the mean was 0.0034 (being centered around zero shows there is a fairly even split of positively and negatively polar decisions) with a standard deviation of 0.074. The minimum polarity score was –0.647 and the maximum polarity score was 0.543. As for subjectivity, the mean was 0.413 (again showing the decisions were pretty evenly split between emotional language and neutral language) with a standard deviation of 0.092. Lastly, the minimum subjectivity score was 0 and the maximum subjectivity score was 1.

In terms of what we wanted to achieve in creating reliable metrics for polarity and complexity, we believe we have met our data mining goal. Additionally, after meeting with Professor Waltenburg and taking him through our results, he was very pleased with our metrics that we came up with. We were successful in creating metrics that are understandable to those without much data science experience, as Professor Waltenburg was able to grasp what we did easily.

We really will not know, however, what the full effect of our metrics will be until Professor Waltenburg completes his final research. Professor Waltenburg plans to use our metrics to see if they stem from more diverse courts as well as some other possible relationships detailed later.





1.2. Approved models

After accessing models with respect to business success criteria, select and approve the generated models that meet the selected criteria.

Given the discussion in the previous subsection, The LDA model with n_components = 20 will be chosen as our model for complexity. For polarity we will use the pre-built Textblob model outlined in the above section as well as in the data modeling report.

2. REVIEW PROCESS

At this point, the resulting model appears to be satisfactory and appears to satisfy business needs. It is now appropriate to make a more thorough review of the data mining engagement in order to determine if there is any important factor or task that has somehow been overlooked. At this stage of the data mining exercise, the Process Review takes the form of a Quality Assurance Review.
Summarize the process review and list activities that have been missed and/or should be repeated.

- Provide an overview of the data mining process used
- Analyze the data mining process. For each stage of the process ask:
    - Was it necessary?
    - Was it executed optimally?
    - In what ways could it be improved?
- Identify failures
- Identify misleading steps
- Identify possible alternative actions and/or unexpected paths in the process
- [Optional] Review data mining results with respect to business success criteria

Overview of the data mining process used:

1) Business Understanding:
We met with Professor Waltenburg to get an idea of what his overall project entails and what he wanted us to accomplish with our project in terms of creating new metrics for complexity and polarity. We also discussed what resources we had at our disposal, and he gave us some background knowledge on the topic of US Circuit Courts. We also read various articles detailing the subject provided by Professor Waltenburg and discussed the project that the group in his class is working on and how it is related to this one. This task was executed optimally as we felt much better about the subject after discussion.

2) Data Preparation and Processing:
In order to prepare our dataset, we had to scrape all the court rulings in the form of PDFs from the US Court of Appeals website and uploaded them to Google Drive. Once there, we were able to scrape the court rulings and other desired information from each document and put it into tabular form. Lastly, we were able to clean the data including removing stop words and lemmatizing the text. Using pdfplumber was an effective way of grabbing the pdfs and getting them to text and was quite necessary to move forward.

3) Predictive Model Training:
We built an LDA model to create complexity scores for each document and used a pre-built TextBlob model in order to create polarity scores for each document. In addition, we were able to include a

subjectivity feature for each document that indicates if the document contains more opinions rather than facts. It would be an interesting feature combining with the polarity score.

These three steps were all necessary and performed optimally. In terms of improvements, one major improvement would be to build a corpus of legal/political terms that also includes how these certain words fall on the left/right spectrum in the context of a court. This was not feasible given the time constraints of this class; however, it would have meant our polarity score would have been concerned with whether a document is more/less politically left/right rather than if a document has a positive/negative context. Basically, it would have provided us with a better view of the polarity of some words in the specific context of a court. For example, “appeal” would be likely to have a positive polarity in a normal context, whereas in a legal context it would most likely be more neutral.

One failure was that we were unable to perform the extra prediction regarding if higher complexity and more polar decisions stem from more diverse courts. While Professor Waltenburg did not need us to perform this step, we were going to attempt it if we had gotten the judge archetype data from Professor Waltenburg’s other research group with enough time.

In terms of misleading steps, one would be that every time LDA runs, it generates a different probability distribution each time. This means that some topics could be a lot more similar to others on different runs. While we did run LDA 5 times in order to stabilize the results, another possible improvement would be to try running the model more times for each n_component value if time had allowed.

A possible alternative to using TextBlob would be to use other pre-built models like Vader or Flair to see what results we would have gotten for polarity scores, or build our own model if time was not an issue. Lastly, an unexpected path that we took was that we were originally going to use word2vec (or doc2vec after being pointed toward it in a presentation) for our LDA model, however realized it would not be possible due to LDA needing a BOW style vector. We did find a model called lda2vec, however it is still quite experimental and wanted to go with a model that was more proven. Until there is a pre-built model that can decipher meanings of specific legal words and can differentiate them from their usual meanings, besides evaluating them in the context of just the court cases provided, we took the proper steps here.

3. DETERMINE NEXT STEPS

Based on the assessment results and the process review, the project team decides how to proceed. Decisions to be made include whether to finish this project and move on to deployment, to initiate further iterations, or to set up new data mining projects.

3.1. List of possible actions

List possible further actions along with the reasons for and against each option.

- Analyze the potential for deployment of each result
- Estimate potential for improvement of current process
- [Optional] Check remaining resources to determine if they allow additional process iterations (or whether additional resources can be made available)
- [Optional] Recommend alternative continuations
- Refine process plan

After meeting with Professor Waltenburg, he plans to go forward and combine our model and metrics with the results from his other research group that was tasked with creating judge archetypes. From this, he plans to see what the relationship is between the polarity and complexity of court decisions and how diverse the justice panels are.

We are sending him all our data, models, and processes to make it easy for him to continue his research. We could also continue with Professor Waltenburg and help him evaluate this relationship if he needs further help.

A possible improvement we could make to the process is if Professor Waltenburg determines that he wants polarity scores that are more focused on how politically left/right words in the decisions are (as opposed to how positive/negative) we would suggest creating a corpus of "political" words in a legal context and we could go back and revise our polarity model to be based on this corpus.

The data mining process can be refined into the following steps:
(1) Data collection
(2) Data Processing/cleaning
(3) Model Selection
(4) Model fine-tuning
(5) Model evaluation

3.2. Decision

Describe the decisions made, along with the rationale for them.

- Rank the possible actions
- Select one of the possible actions
- Document reasons for the choice

The ranking of possible actions are:

Data processing/cleaning > Data collection > Model selection > Model evaluation > Model fine-tuning
We choose Data processing/cleaning and Data collection selection/training as the most critical parts.

Due to the heavy usage of NLP in this project the cleanliness and accuracy of our data is of the upmost importance. It could be possible that we did not gather enough data or clean it as efficiently as possible. NLP models can be very sensitive to small changes in data, which could lead to very different results.

Another possible improvement could be to add more data, we only used 6,000 total samples out of the over 35,000 collected. We also only used the years 2020-2022. More data could create a more complete vocabulary and may produce better results.

Future plans:

Professor Waltenburg plans to use our metrics to see if more polar and complex decisions stem from more diverse courts as well as some other possible relationships. He mentioned using the subjectivity scores we came up with to look at if higher subjectivity relates to cases on more emotional decisions (such as same sex marriage and criminal defense cases as opposed to financial/tax cases). Professor Waltenburg also plans to see how our subjectivity scores compare to ideological scores his other research group came up with, which determines whether judges are more left/right leaning.