O.J. Simpson Criminal Trial:
a topic modeling analysis of the transcripts

Miriam Giuliani

April 7, 2022

1 Introduction

The objective of this project is to apply pre-processing and text-mining techniques to study the features of the OJ Simpson criminal trial for the murder of his ex-wife Nicole Brown and her partner Ron Goldman, occurred in 1995 in California. The choice of this corpus for my practical assignment is motivated by the fact that this court case was very controversial and under the public eye; the characteristics (duration, rhetoric, narrative, media exposure) of what has been defined as *the trial of the century* make the case particularly interesting. Both the defense and the prosecution exploited very visual and rhetoric strategies during the trial, as the very famous gloves fitting; I therefore think that it could be interesting to see whether these features are mirrored also in the official trial transcript and if they can be detected using the techniques studied during the course.
In Section 2 the creation of the data set and the pre-processing steps are described, while in Section 3 I explain the steps done to handle normalization, anaphora resolution and negation handling. The results of the Topic Modeling analysis are contained in Section 4, together with visualizations of the output procedure in Section 5. Section 6 concludes.

2 Building the data set and pre-processing

The data have been scraped using python libraries as BeautifulSoup from the following web site: http://simpson.walraven.org/ . I decided to focus on the transcripts for the period between January 1995 and September 1995 (when the verdict was reached), therefore excluding the Preliminary Hearings and juror interviews.

I then studied the structure of the transcript; here one sample is reported:

(THE FOLLOWING PROCEEDINGS WERE HELD IN OPEN COURT, OUT OF THE PRESENCE OF THE JURY:)
THE COURT: ALL RIGHT. GOOD MORNING, COUNSEL.
MR. COCHRAN: GOOD MORNING, YOUR HONOR.
MR. SHAPIRO: GOOD MORNING, YOUR HONOR.
THE COURT: ALL RIGHT. BACK ON THE RECORD IN THE SIMPSON MATTER. MR. SIMPSON IS AGAIN PRESENT BEFORE THE COURT WITH HIS COUNSEL, MR. SHAPIRO, MR. COCHRAN, MR. BAILEY, MR. BLASIER. THE PEOPLE ARE REPRESENTED BY MISS CLARK AND MR. DARDEN. COUNSEL, SO THE RECORD IS CLEAR, WE ARE IN THE MIDST OF OPENING STATEMENTS BY THE DEFENSE, THE PROSECUTION HAVING COMPLETED THEIR OPENING STATEMENTS.

The first step in pre-processing the text files, was to eliminate blank lines and lines between round brackets, used to describe events and details about the trial (e.g. (A CONFERENCE WAS HELD AT THE BENCH, NOT REPORTED.)). These operations have been done by using regular expressions.

I then made sure to keep only the text related to parts where people involved in the trial are talking; this was relevant because at the end of each transcript some information about that specific day are reported, as the list of people talking, the list of the witnesses, the main arguments discussed and some archive references. This was done by storing separately the part of the text file coming after the phrase " SUPERIOR COURT OF THE STATE OF CALIFORNIA FOR THE COUNTY OF LOS ANGELES ", which is the start of the ending part of the transcript where these information are present. I decided to store these part of the text in my final data set as the information present in this description could be useful in other parts of the project for validation purposes.

The same procedure was applied also to the first lines of each transcript, where the title of the record is reported, together with other information related to that specific day. I also decided to keep this part of the text file because it contains the date and the time of the document, which I might need when performing information extraction tasks. Both dates and times have been isolated using regular expressions.

Then, in order to isolate the different people talking and the correspondent content told by them, I made sure to eliminate `\n` characters which were present within paragraphs of the speech made by the same person; in this way I obtained a text document where each line starts with "`NAME OF THE PERSON TALKING:` ". To extract all the proper names of the different attorneys and people speaking during the trial, I exploited the html tags after scraping, given that the name of the entity talking was reported in bold; regex was used in this phase too.
Then, using again regular expressions I was able to store for each line the person speaking and the content of what they were saying at that point of the transcript. I stored this information as two columns (*person* and *speech*) in a data frame. I also added a column for the date and the time and a row (the last one, with `DESCRIPTION` as value for the *person* column) containing the information at the end of the transcript isolated in the previous step.

Figure 1: Head of the data set

	person	speech	date	time
0	#LOS ANGELES, CALIFORNIA; TUESDAY, JANUARY 31,...	(APPEARANCES AS HERETOFORE NOTED.)\n	JANUARY 31, 1995	9:36 A.M.
1	#THE COURT:	GOOD MORNING, COUNSEL. BACK ON THE RECORD IN T...	JANUARY 31, 1995	9:36 A.M.
2	#MR. COCHRAN:	YES, YOUR HONOR, IF I MIGHT APPROACH THE PODIU...	JANUARY 31, 1995	9:36 A.M.
3	#MR. DARDEN:	YOUR HONOR, MAY I OBJECT? THE LETTER SPEAKS FO...	JANUARY 31, 1995	9:36 A.M.
4	#THE COURT:	OVERRULED.\n	JANUARY 31, 1995	9:36 A.M.
...	...	...	...	...
2805	#MR. DARDEN:	NOT BEGIN. ADJOURN. 4:00 O'CLOCK OR 3:30?\n	JANUARY 31, 1995	9:36 A.M.
2806	#THE COURT:	4:00 O'CLOCK.\n	JANUARY 31, 1995	9:36 A.M.
2807	#MS. CLARK:	WHEN IS THE 3:30 DAY?\n	JANUARY 31, 1995	9:36 A.M.
2808	#THE COURT:	THAT WAS THE DAY THE JUROR HAD TO GO TO A DOCT...	JANUARY 31, 1995	9:36 A.M.
2809	DESCRIPTION	\nDEPARTMENT NO. 103 HON. LANCE A. ITO, JUDGE\...	JANUARY 31, 1995	9:36 A.M.

2810 rows × 4 columns

I performed these procedures separately for each data folder (one per month). This was necessary in order to monitor closely potential differences in the structure of the transcript across time and errors. Indeed, for instance, starting from the transcript of April 13th 1995, the content of the website was modified in another format, resulting in major differences with respect to the html structure and the way how non-dialog content was presented in the transcript (i.e. date and time format, positioning of the description of the transcript in the file) . Also, two transcripts did not contain the content description usually present at the end of the file and therefore a manual modification was necessary in order to obtain an homogeneous data structure.
I then merged all the monthly data sets, obtaining a data frame with 313.899 rows.

3 Normalization, anaphora resolution and negation handling

The next step was normalizing the data and solving potential anaphora resolution and negation handling problems.

The main issue encountered in this phase relates to witnesses' questioning. Indeed, when a witness is speaking, she/he is not denoted by her/his proper name, but by `THE WITNESS:` or, once the questioning has started, by `A:` (as for `A`(nswer)). In a similar fashion, the attorney asking the question is denoted by `Q:` (as for `Q`(uestion)).
This creates an identification problem because the same witness is very often questioned both by the

prosecutors and the defense team, multiple times during the same day. It also often occurs that multiple witnesses are questioned on the same day. Therefore, multiple people are encoded in the same way within the same document. Because of the number of witnesses questioned during the whole trial, a manual modification is not feasible.

To solve this, I decided to exploit some recurrent expressions that typically precede the name of the witness or of the attorney speaking.
For the name of the witnesses, this was possible exploiting the expressions `"-WITNESS NAME- , CALLED AS A WITNESS BY THE PEOPLE, WAS SWORN AND TESTIFIED AS FOLLOWS: "` or `"-WITNESS NAME- , THE WITNESS ON THE STAND AT THE TIME / HAVING BEEN PREVIOUSLY SWORN"` which are present before the starting of the questioning and contain the proper name of interest.
My first attempt was actually to use the expression `PLEASE BE SEATED AND STATE AND SPELL YOUR FIRST AND LAST NAME FOR THE RECORD`, stated by the clerk before each questioning and followed by the answer of the witness spelling her/his name. However, I then discovered that when the witness is questioned on different days, this expression is present only on the first day and not reported in later transcript; therefore this choice can lead to mistakes when identifying the person in other occurrences. Indeed, when manually setting the rules to assign the proper name of the witness I had to take into account all the possible instance combinations (e.g. one witness interrogated for the first time and the other witness interrogated for the second time), for which the positioning of the proper name of interest can change depending on how often the person has been already questioned.

Knowing this information and using regex, I iterated over all the rows of the data set and replaced `A:` with the corresponding name of the witness at specific points of the trial. I also stored the proper names of the witnessed questioned in each particular day, as this information could be useful for other tasks of the project. A manual check was randomly performed in order to assure that the substitutions were correct.

A similar procedure was applied to retrieve the name of the attorney questioning. In this case, I am exploiting expressions like `DIRECT EXAMINATION: BY -NAME OF THE ATTORNEY-` or `CROSS-EXAMINATION: BY -NAME OF THE ATTORNEY-` which precede the starting of the questioning and contain indeed the name of the attorney speaking. Also in this case all possible combinations have been taken into account when writing the regex rules (e.g. only direct examination, direct examination and cross-examination, resumed examinations after recess).

Note that because of the tasks performed before (i.e. removing `\n` characters) and because, for some reason, the description about the opening of the testimonies are not present in round brackets (like all the others details reported in the transcript that are not part of people's speeches, which have been removed in the beginning), the expressions used to find the proper names of witnesses and attorneys are now "attached" at the end of the speech given by the person talking right before, as they have not been eliminated in the first pre-processing steps. This is not a problem in this phase as I am looking for a match across all the text present in my data set, but it will be considered in the following pre-processing steps and such expressions will be removed as they contain information that are not actually part of what people are saying and therefore are not relevant.

I then proceeded by normalizing names used in the transcript referring to the same person. For instance, I normalized all the instances referring to O.J.Simpson, such as *Orenthal Simpson, O.J., the defendant, the suspect*; the same was done for the victims, family members, jurors, witnesses and experts.

Then, I made sure to handle anaphora resolution issues. I did that with a rule based approach, exploiting some features of the transcript. For instance, I substitute *I, me, we* with the name of the person talking, and you with the person talking next. Also, I identified the days in which family members of the defendant and of the victims had been interrogated and substituted words like *my/your brother, my/your son, my/your dad, your ex-wife* with the corresponding person.

Finally, I tried to take into consideration negations, by exploiting a *negation tagging* approach. This was done by appending the string NEG to words appearing between a negation instance (e.g. *not, no, none*) and a punctuation sign. In this way, the same word will be tokenized as two depending on the fact that it is close to a negation.

4 Topic Modeling

As said in the introduction, the trial was very visual and full of changes of narrative. It is reasonable to say that the official trial transcripts can potentially fail in mirroring closely what was actually perceived by the public and by the jury. The goal is therefore to use some text mining techniques to retrieve these features.

The book by Cotterill (2003) gives a more precise idea of the strategies used by the attorneys during the O.J. Simpson trial. While the prosecution decided to sustain their arguments presenting a single, linear and logical story, the defense exploited the opportunities to reinforce the jury's reasonable doubt. In particular, the prosecution's narrative was based on the relevance of evidences and witnesses available and on the past violent behaviour of the defendant against Nicole Brown. On the other hand, the so called *dream team* focused on proposing different alternatives to the opponent's theory, by introducing elements such as the real killer's story (theory based on which the murder was related to drug consumption) and sustaining that rush into judgment was in place because of the weakness of the evidence available (e.g. the murder weapon, never found) and non-professional and racist behaviour of the LAPD detective Mark Fuhrman, accused of planting evidence at the crime scene.

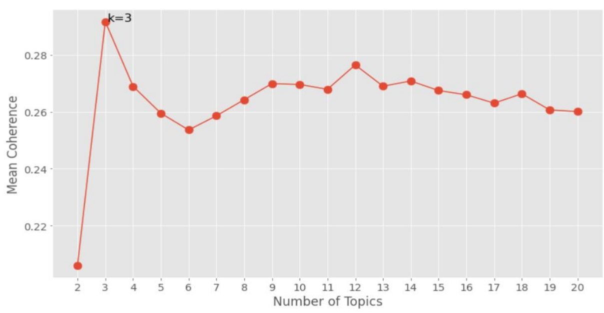
The goal of topic modeling is then to see whether we can trace these narratives, starting from the raw text of the transcripts.

I first tokenized and stemmed the text, removing punctuation and stop words. Then, I proceeded by computing a TF-IDF matrix over my data set, which resulted in a (234700×34708) object.

To perform topic modeling I decided to use Non-Negative Matrix Factorization (NMF), using the Python library sklearn.
Although I had a general idea of some of the main arguments discussed during the trial, I needed a more rigorous way to evaluate the topics and choose how many of them I should have look for. For this reason, I used a Skip-gram Word2Vec model, with the python library, in order to compute the coherence scores for models with different numbers of topics. This metrics can be interpreted as a measure of the semantic similarity between high scoring words assigned to a specific topic; by taking the average of this metric between topics within a model I can obtain a measure of quality, which can allow me to evaluate the model and choose the best number for k.

By using the whole data set, the plot of this metric for different numbers of topics is the following:

Figure 2: Quality of the model changing k



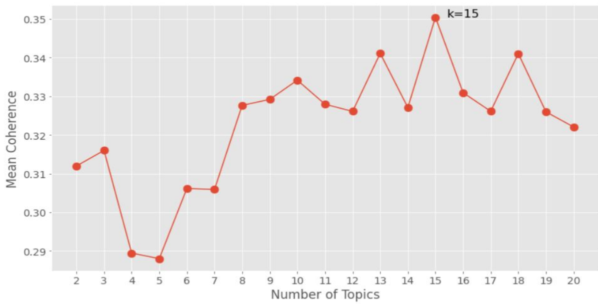
Based on what said before, the optimal value for k is 3.
In order to see which were the topics selected, I looked at the words with higher frequency for each group and the results showed that the model captured the proper names of the people involved in the trial. In particular, the first topic (see Appendix) had the Court as word with the highest frequency, followed by *Ms Clark* (main prosecutor) and other attorneys participating in the trial. Curiously, the second topic, had as first word *Mr. Crochran*, the team leader for the defense. Even if it is clear that more pre-processing needs to be done to find good topics, the fact that the name of Mr. Crochran does

not appear together with the ones of all the other attorneys (including his collegues in the dream team) can tell us something about his role in the trial, suggesting that his words may have a distinctive feature with respect to the other people involved. With respect to the last and third topic, the main words are *objection, speculation, hearsay*, suggesting that the words present here are the ones present during the witnesses' interrogation.

In order to continue the search for topics that better mirror the narrative of the trial, I eliminated the proper names of the court and the attorneys of the defense and prosecution team. Then, I computed the coherence metric again over the new text; this time, only two topics were selected. This result, together with the visualization of the words present in the groups suggests that more irrelevant words need to be eliminated. Indeed, tokens as "thats, see, may, let" are still the ones with the highest frequency within topics.

After eliminating irrelevant words and re-evaluating the model, the new best value for k is 14. The first 30 words per topic are reported in the Appendix.

Figure 3: Quality of the model changing k - Model 3



Looking then at the topics selected, finally we can start identifying some of the strategies cited before. In particular, out of the 14 topics selected, 6 (i.e. Topics 1, 2, 9, 11, 12 and 14) can be easily associated with the prosecutors' strategy, based on evidence and witnesses. For instance, in Topic 12 the most relevant words are " evidence, item, number, items, collected, fung, scene, trace, envelope, glove, respect, numbers, bundy, processing, stain "; similarly, Topic 9 contains the names of the main witnesses called on the stand by the prosecution, such as Kato Kaelin (actor and friend living at the defendant's house the night of the murder) and Rosa Lopez (Simpson's housekeeper), together with the names of the medical examiners called as experts.

On the other hand, Topics number 3, 7 and 10 show similarities with the narrative sustained by the defense. In particular, Topic 3 contains as main words "calls, speculation, argumentative, conclusion, facts, vague, forspeculation, strike, also, still, foundation ", which confirm the dream team's argument of the 'rush judgement'. Also, Topic 10 first word is "Fuhrman", that is the detective in charged of the investigation, who was at the center of the trial because of his past racist comments, played on tape for the jury. Finally, the Topic 7 contains as first tokens "lange, detective, vannatter, recall, ever, scene, recognize, arrived " which are words or proper names referring to the LAPD professional activity, one of the main strategy of the defense.

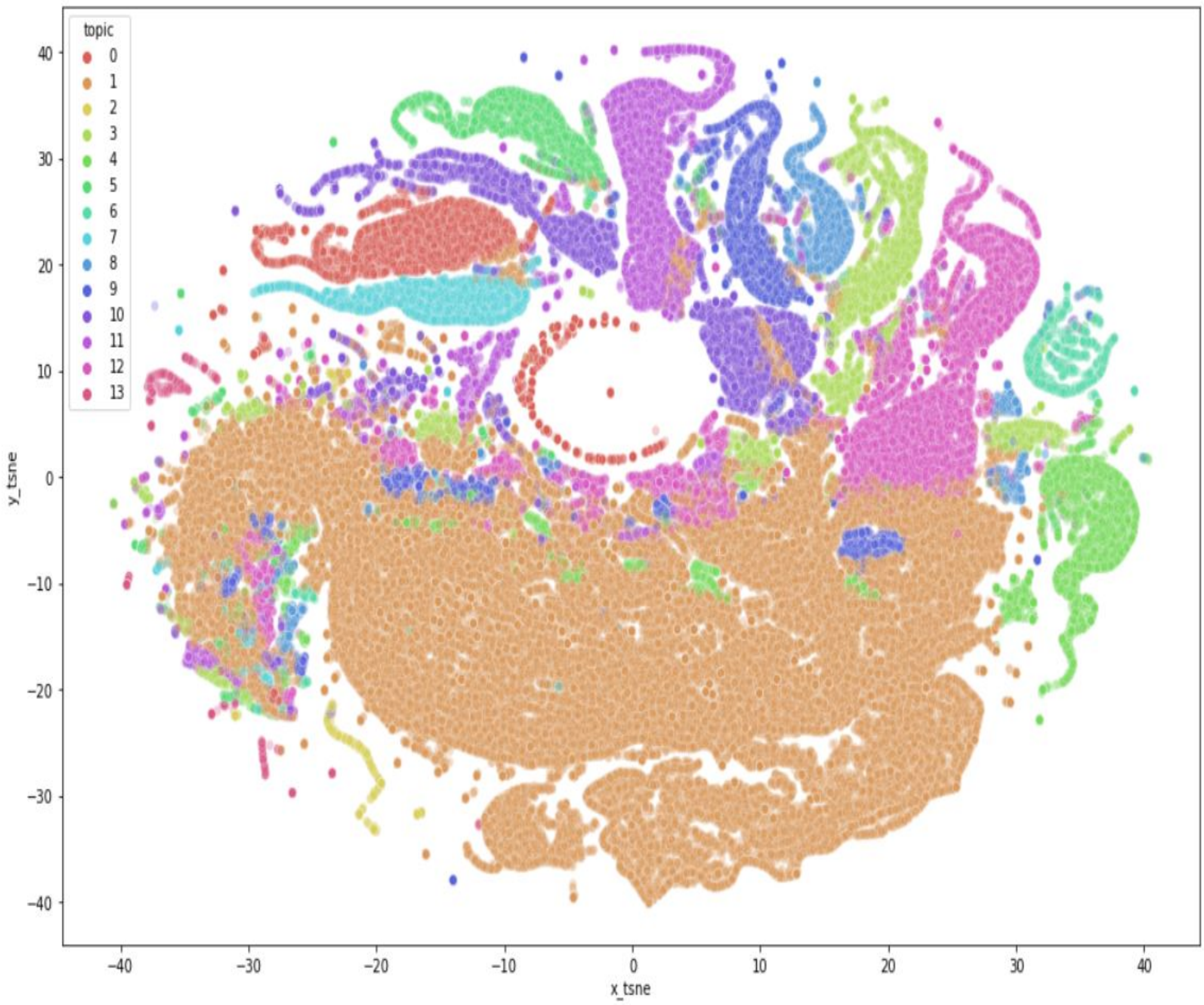
Between the remaining topics, we can say that Topic 6 and 13 are not easily associated with a particular narrative, as they contain an heterogeneous mix of tokens. Topic 5 contains words related to the concept of "time", but it is not clear whether it can be associated to a particular strategy. Finally, Topic 4 contains tokens typically occurring when the judge, the clerk or the attorney talk to the jury, such as jury, ladies, gentlemen, morning, record, open, reflect, stand.

To conclude this section, in terms of evaluation of the topic quality, this last model with 14 groups has the highest level of coherence when compared to the initial model, the selected model after removing the proper names of attorneys and any other model between the ones compared to select k.

5 Visualization

In order to visualize the topics selected with the final model, I first reduced the space to 2-dimensional using t-Distributed Stochastic Neighbor Embedding (or t-SNE), available with the sklearn library TSNE. Then, I plotted the topics as a scatter plot, as can be seen in the following image:

Figure 4: Topic 2 - First Model



This figure can help in visualizing the dimension and distribution of the topics. Topic 2 (in orange) is the biggest one and contains words related to the evidence and murder scene. Topic 5 (on the right in acquamarine), which is related to words about time lives alone in the space on the right of the figure. Finally, topic number 4 (lemon green, on the right), classified as the one of interaction with the jury is between topics 9 and 12, associated with the defense (Furhman topic) and prosecution respectively.

Then, to visualize how the topics evolved across time, topic rivers were performed:

An interesting feature we can see from this plot is the behavior of Topic 9, in dark blue, corresponding to detective Mark Fuhrman. Indeed, we see an increase in the intensity during the months of March and September. These two periods correspond respectively to the moment when the detective was called as a witness and the moment when the tapes in which he says racist affirmations are played to the jury. The same holds for Topic 7 (sage green), relative to the other detectives interrogated, for which we see an increase in February and March, when the questioning was performed. Finally, we see an increase at the very end for Topic 4, corresponding to moments when the attorneys talk to the jury. This makes

Figure 5: Topic River (baseline wiggle)

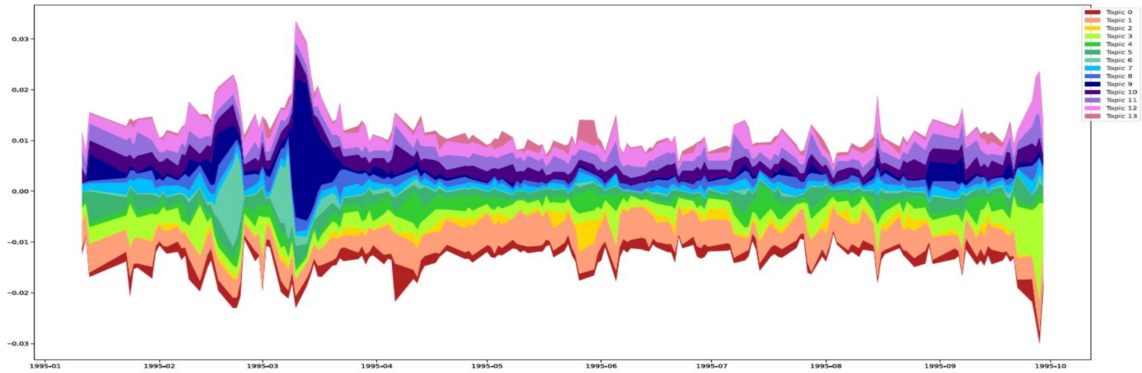
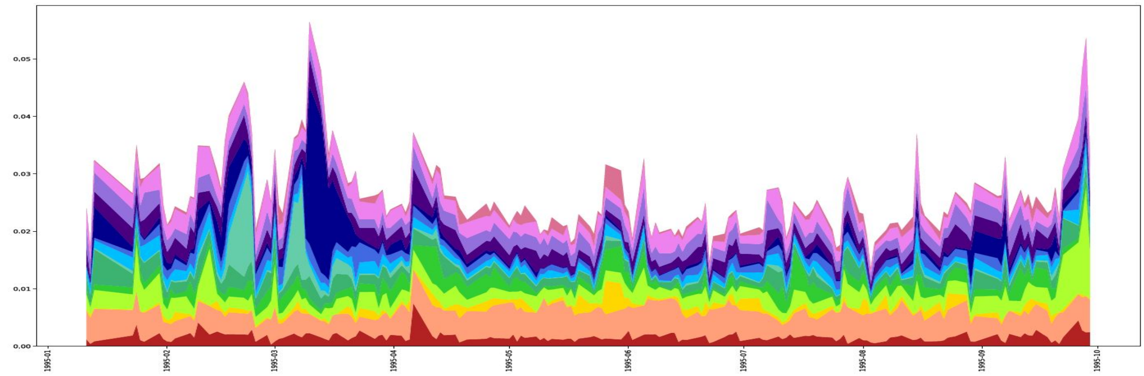


Figure 6: Topic River (baseline zero)



sense since this is the moment of the closing arguments and of the verdict.

In the appendix, the same plots for the first model can be found.

6 Conclusion

With respect to the different tasks described in this work, a relevant part of the time was dedicated to the pre-processing steps. This was due to different reasons.
In particular, the raw structure text was not homogeneous, which made necessary a series of modification of the rule-based approach used in the pre-processing phase. This was evident also for the normalization of the witnesses interrogation, which was characterized by a not small number of exceptions and errors. The same holds for the retrieval of other features such as date and time and name of the witnesses.
With respect to the operation of normalization and anaphora resolution , more sophisticated solutions, with an application of statistical and machine learning techniques might have brought to better results and less errors. Indeed, even after an intensive activity of rule creations, some irrelevant or not normalized words were still in the corpus. In particular with respect to negation handling, the solution adopted here did not work well for the scope of the project, as in Model 2 all the negative tagged words were clustered together, without contributing to the information extraction. Also to this respect a "less brute-force" solution could be preferable.
With respect to the results I think that the output obtained with Model 3 and 14 topics gave interesting insights as it was possible to extract topics mirroring the strategies of the defense and prosecution, in an unsupervised way.
As further steps, I would explore the topic modeling result even more, maybe proceeding in a hierarchical way in order to find more refined group. It could also be interesting to associate emotions to the topics, reflecting the perception of the jury towards O.J. Simpson.

References

[1] **J.Cotterill** "Language and Power in Court. A Linguistic Analysis of the O.J.Simpson trial. Palgrave Macmillan, 2003

[2] **Stevens, Keith, Kegelmeyer, Philip, Andrzejewski, David, Buttler, David.** Exploring Topic Coherence over Many Models and Many Topics - Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (Association for Computational Linguistics)

[3] **Sanjiv Das and Mike Chen. 2001.** Yahoo! for Amazon: extracting market sentiment from stock message boards. In Proceedings of the 8th Asia Pacific Finance Association Annual Conference.

[3] **Allan, James** Topic detection and tracking: event-based information organization.Kluwer Academic Publishers. (Editor), 2002

[3] https://towardsdatascience.com/using-nmf-to-classify-companies-a77e176f276f.

[3] https://towardsdatascience.com/visualizing-topic-models-with-scatterpies-and-t-sne-f21f228f7b02

7 Appendix

First 30 words per topic - Model 3

Topic 01: see, photograph, reporter, look, ever, sidebar, side, area, monitor, bar, kaelin, blood, picture, indicating, looking, tape, little, location, photographs, looked, riske, car, inside, bag, left, drlakshmanan, hand, top, videotape, white
Topic 02 two, case, also, blood, dr, doctor, area, want, saw, get, fact, mean, told, make, way, three, recall, left, nicolebrown, photographs, show, scene, drlakshmanan, people, sure, made, done, defense, many, put
Topic 03 calls, speculation, argumentative, conclusion, facts, vague, forspeculation, strike, also, still, foundation, phone, move, compound, nofoundation, motion, opinion, calling, conjecture, made, lack, legal, excuse, call, witness, yourhonor, overbroad, part, narrative, defense
Topic 04 jury, ladies, gentlemen, morning, record, open, reflect, stand, rejoined, recess, afternoon, members, panel, remember, beseated, witness, step, disregard, admonitions, seated, grand, resume, weve, present, jurors, need, proceed, oclock, matter, tomorrow
Topic 05 th, june, morning, july, date, fung, august, saw, night, bronco, taken, day, mryamauchi, september, january, mrfung, collected, february, hours, rd, recall, oclock, afternoon, april, evening, rockingham, early, pm, st, prior
Topic 06 point, foundation, leave, want, understand, conversation, make, phillips, made, get, still, riske, tape, come, mrheidstra, left, went, show, wanted, house, making, issue, need, vannatter, foundational, inside, dog, already, might, came
Topic 07: lange, detective, vannatter, recall, ever, scene, recognize, arrived, regard, rockingham, location, tom, shoes, saw, fung, morning, bundy, investigation, read, report, indicated, look, talk, phillips, body, arnelle, yesterday, inside, went, talked
Topic 08 know, far, want, mean, people, talking, need, name, wanted, long, came, phillips, person, kind, problem, havent, vannatter, taken, fung, hes, many, riske, kaelin, rossi, much, way, date, done, umm, lot
Topic 09 tell, kaelin, photograph, lopez, drlakshmanan, fung, truth, looking, happened, ever, recognize, riske, else, park, look, showing, us, mrfung, shown, drcotton, rossi, drgerdes, police, show, hard, exactly, house, dog, saw, found
Topic 10 fuhrman, detective, phillips, vannatter, mark, told, officer, conversation, scene, glove, rockingham, riske, saw, went, bundy, arrived, morning, detectives, walked, police, recall, call, bronco, location, okayand, house, door, ever, happened, long
Topic 11 next, witness, peoples, photograph, marked, order, exhibit, happened, stand, show, mark, call, record, board, defense, shown, showing, page, slide, whats, last, identification, day, recognize, seat, defendants, names, complies, resume, move
Topic 12 evidence, item, number, items, collected, fung, scene, trace, envelope, glove, respect, numbers, bundy, processing, stain, collection, room, dna, control, mrfung, rockingham, piece, photo, hair, lapd, particular, collecting, bag, biological, board
Topic 13 take, look, want, long, break, recess, need, minutes, place, notes, step, minute, care, photographs, remember, opportunity, matter, admonitions, board, second, conversation, get, reporter, vannatter, stand, else, dog, issue, later, monitor
Topic 14 back, record, come, went, present, came, put, get, house, rockingham, gate, parties, front, matter, bundy, jurors, deputy, kaelin, matterall, room, left, need, second, people, little, walked, inside, car, represented, call

Figure 7: Topic 1 - First Model

Figure 8: Topic 2 - First Model

Figure 9: Topic 3 - First Model

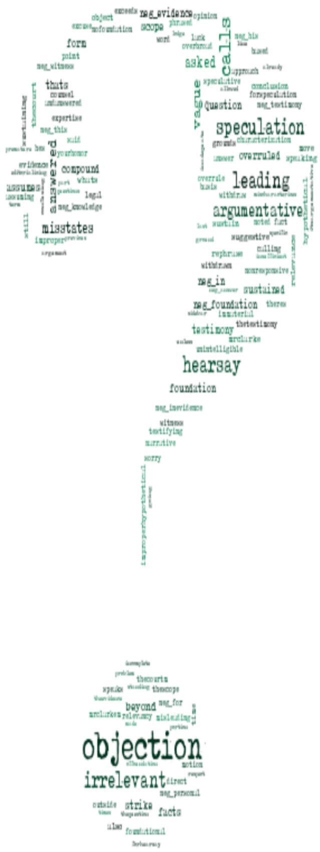


Figure 10: First Model - 2D Visualization of the topics

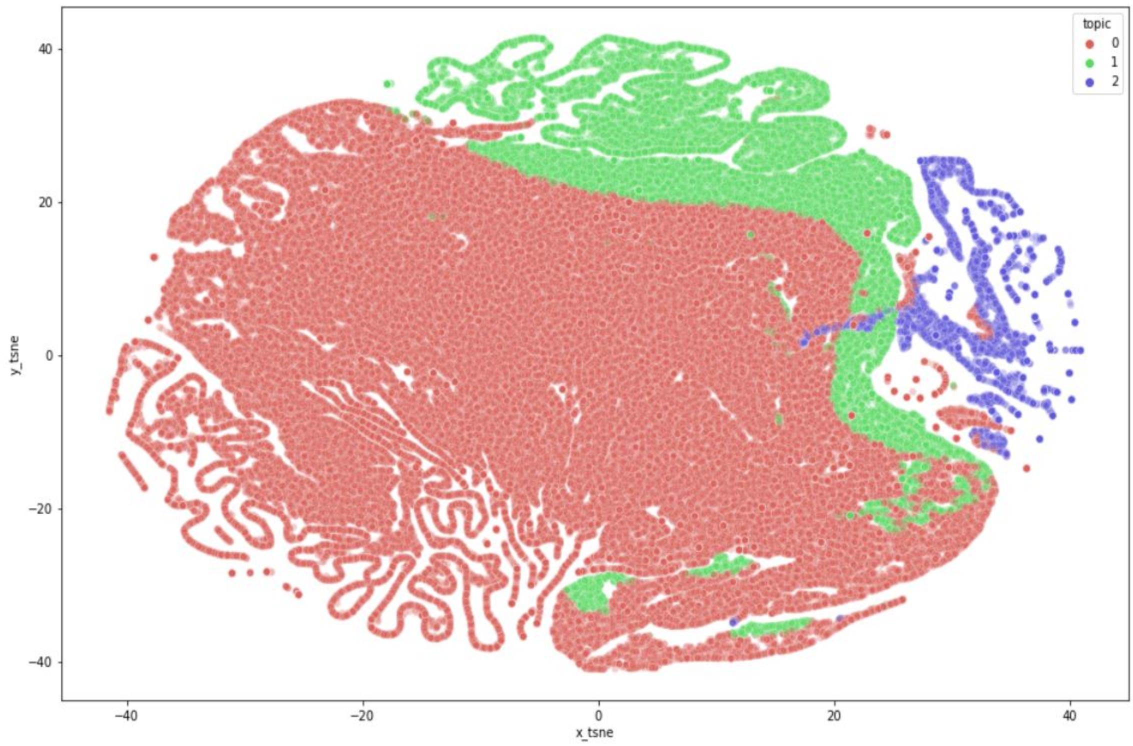


Figure 11: Topic River (baseline zero) - Model 1

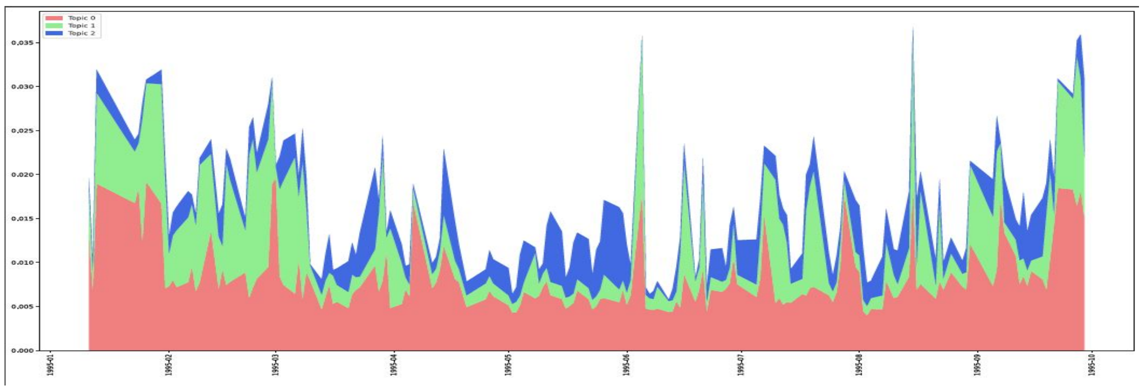


Figure 12: Topic River (baseline wiggle) - Model 1

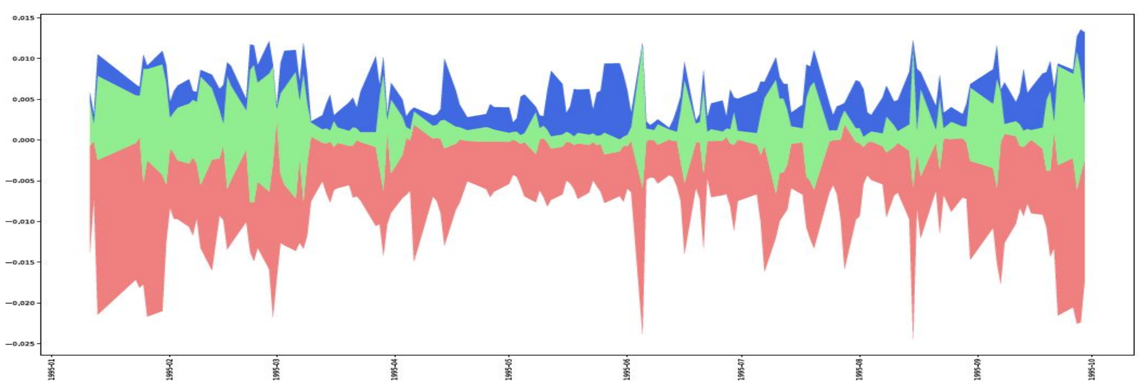


Figure 13: Topic 1 - Model 2

Figure 14: Topic 2 - Model 2