SUMMARIZATION OF COMMERCIAL CONTRACTS

Keshav Balachandar[1], Anam Saatvik Reddy[1],
A. Shahina[1], Nayeemulla Khan[2]

[1]Department of Information Technology,
SSN College of Engineering, Chennai, India
[2]School of Computer Science and Engineering, VIT University, Chennai, India

ABSTRACT

In this paper, we propose a novel system for providing summaries for commercial contracts such as Non- Disclosure Agreements (NDAs), employment agreements, etc. to enable those reviewing the contract to spend less time on such reviews and improve understanding as well. Since it is observed that a majority of such commercial documents are paragraphed and contain headings/topics followed by their respective content along with their context, we extract those topics and summarize them as per the user's need. In this paper, we propose that summarizing such paragraphs/topics as per requirements is a more viable approach than summarizing the whole document. We use extractive summarization approaches for this task and compare their performance with human-written summaries. We conclude that the results of extractive techniques are satisfactory and could be improved with a large corpus of data and supervised abstractive summarization methods.

KEYWORDS

Text summarization, automatic summarization, commercial contracts.

1. INTRODUCTION

In today's day and age, contracts are drafted for every agreement between two parties, documents that companies, firms, and individuals deal with are increasing rapidly. It has become very difficult for corporate staff and chief officers to review contracts which could either be 2 pages or go beyond 100s of pages. To alleviate this difficulty, a large number of companies engage tools for summarizing contracts, extracting key pieces of information, and aiding in other such tasks. Summarization of the entire document is not fruitful as the summaries might be too vague and each line carries a different level of importance. This has been the main motivation behind our project. Thus we propose a solution to initially obtain the preferred topics/headings that are of importance to be included in the summary. We use existing systems and methods to generate summaries, with the novelty focusing on a domain-specific approach for commercial documents. The topics/headings from a given contract are made available to the user to choose from. This would make the generated summary accurate and caters to the unique needs of individual users. We have explored only the extractive ways to summarize a document. We have abstained from using abstractive summarization techniques as a large number of input documents are required to train a supervised model. This problem can be addressed by aggregating more input data with human-written summaries and using a supervised methodology to get better results. We look to expand on existing technologies and validate a tool for automatic summarization of legal documents that would most certainly be useful to lawyers, corporates, professionals to review

David C. Wyld et al. (Eds): MLIOB, SIPO, NET, DNLP, SOEA, AISCA - 2021
pp.19-26, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111202

various contracts. Even common men could potentially use it to obtain a general idea of the contracts they are about to sign or others concerning their interests. Having said that, it might not work for someone viewing a contract for the first time as they might fail to see the domain-specific importance that it carries.

2. RELATED WORK

Haghighi and Vanderwende [1] presented an exploration of generative probabilistic models for multi-document summarization. They started with a basic word frequency-based model and developed a sequence of models such as SumBasic, KL-Sum, TopicSum, and HierSum. HierSum was a hierarchical LSA-based summarizer, which gave the best ROUGE score.

Galgani et al. [2] compared traditional summarization methods with rule-based systems with a custom knowledge base and catchphrases acquired from legal documents acquired from the Federal Court of Australia. They show that the knowledge base created outperforms traditional summarization techniques.

Polsley et al. [3] proposed a tool called CaseSum for automatic text summarization of legal texts. They combined the word frequency method with additional domain-specific knowledge such as the involved parties, abbreviation of entity names. They used ROUGE as well as a custom domain expert to evaluate their approach.

Manor and Li [4] proposed a method for summarizing the Terms of Service. They tested out extractive summarization methods and compared them with human-written summaries. Their work and conclusions aligned most with our work and they are further discussed in the coming sections.

Erera et al. 2019 [5] proposed a novel method that generated summaries for research publication in the computer science domain. Each research paper was parsed from which tables, images, titles, and other metadata were extracted. Along with this they also extracted different types of entities and utilized a custom Unsupervised query focused multi-document summarization using the cross-entropy method. [6]

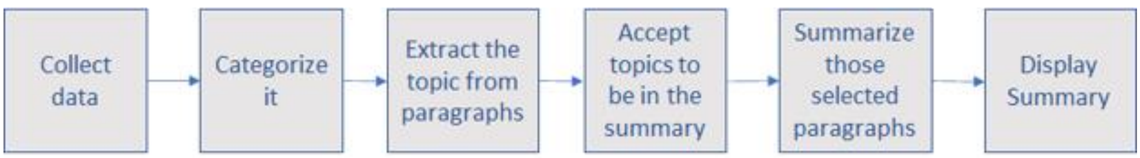


Figure 1. Workflow of the project.

3. PROPOSED METHODOLOGY

In this section, the various steps of the summarization process, as depicted in Figure 1, are discussed.

3.1. Collection of Data

The first step in building a model to summarize a text is to collect, categorize, and pre-process data. As mentioned earlier, we are considering the case of "Employment agreement". The total number of samples collected is 1000, taken from the open-source repository of LexPredict [7].

3.2. About the data and Categorisation

As mentioned above, we are focusing on the sub-domain of the “Employment Agreement”. There are two divisions for an employment agreement. One is a newly issued one, and the other is the amendments to the previous original agreement. It is observed that amendments usually contain less information. So we have our first 2 categories: “Amendments”, “Agreements”.

From the collected dataset, it is observed that some of the contracts are merely empty forms. So those are to be omitted. They are categorized as “Empty”. As mentioned before, a majority of documents contain headings/ topics succeeded by paragraphs. Further, the “Agreements” are categorized as those with “Headings”, and those with “Without Headings”. Since it is important to tokenize the documents as paragraphs and further into sentences, we must know how the paragraphs are segmented. Subsequently, the documents with “Headings” are further categorized as “Alphabets”, “AlphaNum”, “Number.Number”, “Number”, and “Roman”, meaning how they are indexed in the document. These are depicted in Figure 2.

The categorization is mainly done to find out how each topic/ heading is indexed so that it will be easier to extract them.

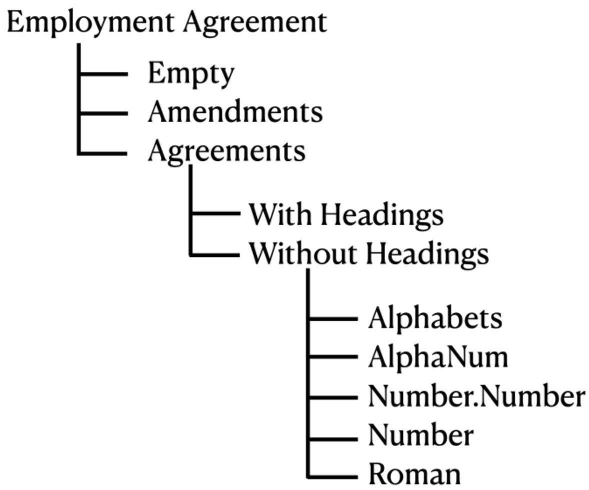


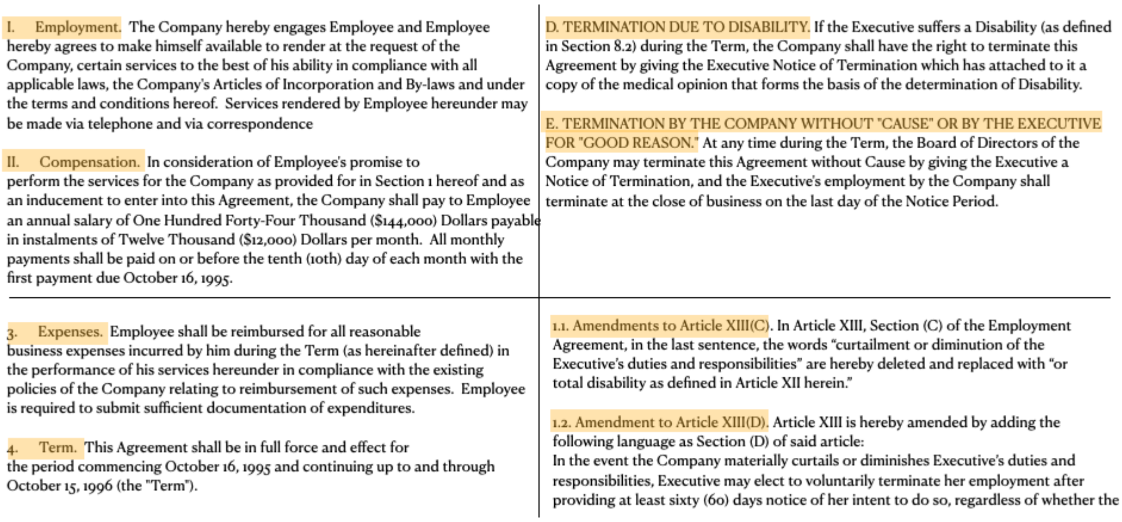
Figure 2. Categorisation of the dataset.

3.3. Data Pre-processing

Once the categorization is done, the data is subjected to cleaning and preprocessing for the task of summarization. Using basic python formatting techniques, the topic-paragraph pairs can be extracted and inserted into a dictionary. This is done for the entire document that is uploaded.

3.4. Topic Extraction

The topic extraction is based on the observation that the majority of the documents are indexed (Alphabets, AlphaNum, Number.Number, Number, Roman), and contain heading/ topic for the corresponding paragraphs, as seen in Figure 3. For the remaining documents, in the future, this project can be expanded where we can train a model to identify the topic of the paragraph and then map it with its corresponding content.



I. Employment. The Company hereby engages Employee and Employee hereby agrees to make himself available to render at the request of the Company, certain services to the best of his ability in compliance with all applicable laws, the Company's Articles of Incorporation and By-laws and under the terms and conditions hereof. Services rendered by Employee hereunder may be made via telephone and via correspondence

II. Compensation. In consideration of Employee's promise to perform the services for the Company as provided for in Section 1 hereof and as an inducement to enter into this Agreement, the Company shall pay to Employee an annual salary of One Hundred Forty-Four Thousand ($144,000) Dollars payable in instalments of Twelve Thousand ($12,000) Dollars per month. All monthly payments shall be paid on or before the tenth (10th) day of each month with the first payment due October 16, 1995.

D. TERMINATION DUE TO DISABILITY. If the Executive suffers a Disability (as defined in Section 8.2) during the Term, the Company shall have the right to terminate this Agreement by giving the Executive Notice of Termination which has attached to it a copy of the medical opinion that forms the basis of the determination of Disability.

E. TERMINATION BY THE COMPANY WITHOUT "CAUSE" OR BY THE EXECUTIVE FOR "GOOD REASON." At any time during the Term, the Board of Directors of the Company may terminate this Agreement without Cause by giving the Executive a Notice of Termination, and the Executive's employment by the Company shall terminate at the close of business on the last day of the Notice Period.

3. Expenses. Employee shall be reimbursed for all reasonable business expenses incurred by him during the Term (as hereinafter defined) in the performance of his services hereunder in compliance with the existing policies of the Company relating to reimbursement of such expenses. Employee is required to submit sufficient documentation of expenditures.

4. Term. This Agreement shall be in full force and effect for the period commencing October 16, 1995 and continuing up to and through October 15, 1996 (the "Term").

1.1. Amendments to Article XIII(C). In Article XIII, Section (C) of the Employment Agreement, in the last sentence, the words "curtailment or diminution of the Executive's duties and responsibilities" are hereby deleted and replaced with "or total disability as defined in Article XII herein."

1.2. Amendment to Article XIII(D). Article XIII is hereby amended by adding the following language as Section (D) of said article:
In the event the Company materially curtails or diminishes Executive's duties and responsibilities, Executive may elect to voluntarily terminate her employment after providing at least sixty (60) days notice of her intent to do so, regardless of whether the

Figure 3. Some of the topics extracted from documents. From left top corner, clock-wise: Roman, Alphabets, Numbers, Number.Number

3.5. Models Used

3.5.1. Tf-Idf Summarization

Term frequency-inverse document frequency is used as a weighting factor for term features. For each term in the document, the weight increases as the word frequency increases, but it is offset by the number of times the word appears in the entire data set. The logic behind this is that if a term or word appears frequently, it’s important. But if it appears frequently in other documents as well, it's probably not that important, and therefore alters its weight accordingly. This is the drawback that from using the bag-of-words model as it took into account all the frequent words without discrimination.

3.5.2. TextRank

The TextRank algorithm [8] was inspired by the famous PageRank algorithm, which models any document as a graph using sentences as nodes. It determines the relation of similarity between two sentences based on the content they both share. This overlap is calculated simply as the number of common lexical tokens between them, divided by the length of each to avoid promoting lengthy sentences.

3.5.3. LexRank

LexRank Algorithm [9] is similar to the TextRank algorithm as discussed before. It uses a modified version of the PageRank algorithm to rank the sentences in the document. It models the document as a graph using sentences as its nodes. But unlike TextRank, where all the weights are assumed as unit weights, LexRank utilizes the degrees of similarities between words and phrases. Then calculates the centrality of those sentences and assigns the weight to the node. Modified cosine similarity is then used to compare the similarity between two sentences.

3.5.4. Latent Semantic Analysis

Latent Semantic Analysis [10] is a technique that analyzes relationships between document sentences, first by constructing a document term matrix, which is a representation of each of the document sentences as vectors, where the rows correspond to the document sentences and the columns are unique words present in the vocabulary. Then Singular Value Decomposition is used to reduce the number of rows while still capturing the structure among the columns. Finally, cosine similarity is calculated between vectors formed by any two columns to determine the degree of closeness.

3.5.5. KL-Sum

Statistically speaking, KL-divergence [11] is a measurement used to find the difference between 2 distributions. KL-Sum is a greedy optimization approach that measures the divergence of the summary vocabulary words from the input document vocabulary words. It adds sentences to the summary so long as it decreases this divergence value. There are 2 main criteria for selecting a sentence to be in the final summary: The KL Divergence between the input vocabulary's set of unigrams and the output/ summary vocabulary's set of unigrams. And the number of words in the summary should be less than L. The algorithm, although is similar to PageRank and TextRank, at its core KL Sum uses the KL Divergence formula to measure how different each sentence is from one and other.

We made use of the package Sumy [12] for executing LSA, LexRank, TextRank, KL-Sum.

4. EVALUATION METRICS AND RESULTS

In this section, we discuss two ways to evaluate the generated summaries. Table 1 summarizes the evaluation results for the models used.

4.1. Rouge

Recall-Oriented Understudy for Gisting Evaluation is a set of metrics used for evaluating automatic summarization and machine translation. The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation. [13]

Recall in the context of ROUGE means how much of the reference summary is the system summary recovering or capturing.

$$Recall = \frac{number\ of\ overlapping\ words}{total\ words\ in\ reference\ summary}$$

Precision on the other hand measures how much of the system summary was relevant or needed.

$$Precision = \frac{number\ of\ overlapping\ words}{total\ words\ in\ system\ summary}$$

The F-measure considers both the precision and recall and is the harmonic mean of the two.

- ROUGE-N: Overlap of N-grams between the system and reference summaries.
- ROUGE-1: Refers to the overlap of unigrams (each word) between the system and reference summaries.

- ROUGE-2: Refers to the overlap of bigrams between the system and reference summaries.
- ROUGE-L: Longest Common Subsequence (LCS) based statistics. It takes into account sentence level structure similarity naturally and identifies longest co-occuring in sequence n-grams automatically.

Table 1. F-measure scores of the 5 models used.

Model/ Metric	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
LexSum	0.4916	0.1898	0.4421
TextRank	0.5098	0.2366	0.5096
KLSum	0.4799	0.1745	0.3957
LSA	**0.5382**	**0.2399**	**0.5099**
Tf-Idf	0.4902	0.1908	0.4286

5. USER INTERFACE

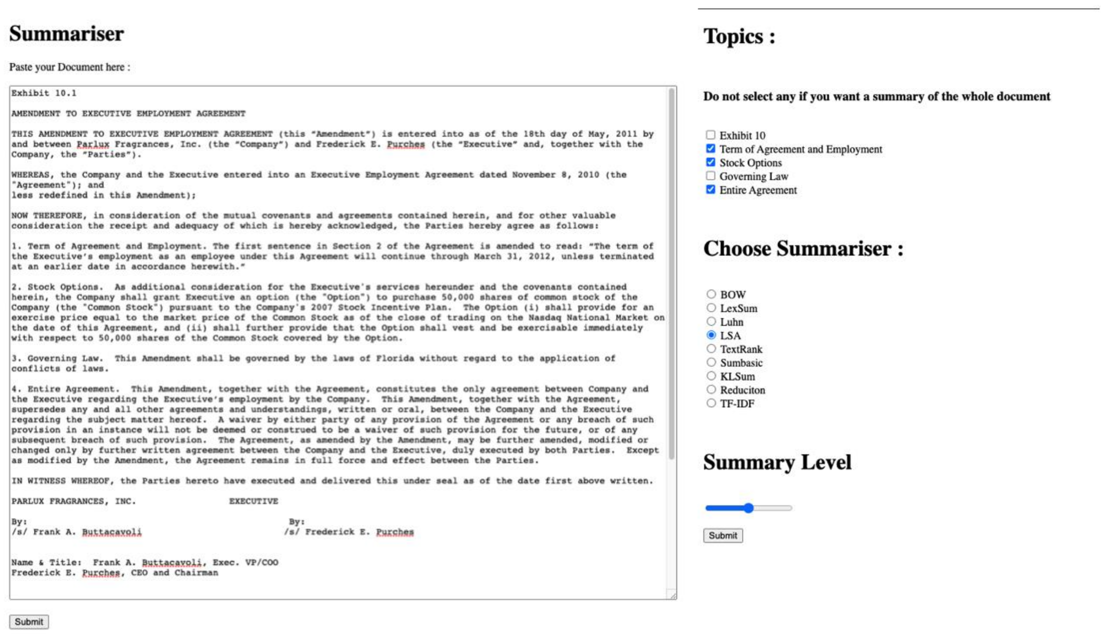


Figure 4. Uploading contracts and Topic extraction.

On the left-hand side of Figure 4, a sample employment agreement is uploaded. On the right-hand side of Figure 4, the topics are extracted and displayed to the user. The user selects the topics that are to be included in the summary and how detailed the summary has to be. (Choosing the Summarizer is for the paper's explanation point of view).

The summary of the uploaded contract is displayed in Figure 5. For the sake of simplicity, the best performing LSA is chosen to summarize the input document.

Original Sample Contract:

Exhibit 10.1

AMENDMENT TO EXECUTIVE EMPLOYMENT AGREEMENT

THIS AMENDMENT TO EXECUTIVE EMPLOYMENT AGREEMENT (this "Amendment") is entered into as of the 18th day of May, 2011 by and between Parlux Fragrances, Inc. (the "Company") and Frederick E. Purches (the "Executive" and, together with the Company, the "Parties").

WHEREAS, the Company and the Executive entered into an Executive Employment Agreement dated November 8, 2010 (the "Agreement"); and
less redefined in this Amendment);

NOW THEREFORE, in consideration of the mutual covenants and agreements contained herein, and for other valuable consideration the receipt and adequacy of which is hereby acknowledged, the Parties hereby agree as follows:

1. Term of Agreement and Employment. The first sentence in Section 2 of the Agreement is amended to read: "The term of the Executive's employment as an employee under this Agreement will continue through March 31, 2012, unless terminated at an earlier date in accordance herewith."

2. Stock Options. As additional consideration for the Executive's services hereunder and the covenants contained herein, the Company shall grant Executive an option (the "Option") to purchase 50,000 shares of common stock of the Company (the "Common Stock") pursuant to the Company's 2007 Stock Incentive Plan. The Option (i) shall provide for an exercise price equal to the market price of the Common Stock as of the close of trading on the Nasdaq National Market on the date of this Agreement, and (ii) shall further provide that the Option shall vest and be exercisable immediately with respect to 50,000 shares of the Common Stock covered by the Option.

3. Governing Law. This Amendment shall be governed by the laws of Florida without regard to the application of conflicts of laws.

4. Entire Agreement. This Amendment, together with the Agreement, constitutes the only agreement between Company and the Executive regarding the Executive's employment by the Company. This Amendment, together with the Agreement, supersedes any and all other agreements and understandings, written or oral, between the Company and the Executive regarding the subject matter hereof. A waiver by either party of any provision of the Agreement or any breach of such provision in an instance will not be deemed or construed to be a waiver of such provision for the future, or of any subsequent breach of such provision. The Agreement, as amended by the Amendment, may be further amended, modified or changed only by further written agreement between the Company and the Executive, duly executed by both Parties. Except as modified by the Amendment, the Agreement remains in full force and effect between the Parties.

IN WITNESS WHEREOF, the Parties hereto have executed and delivered this under seal as of the date first above written.

PARLUX FRAGRANCES, INC. EXECUTIVE

By: By:
/s/ Frank A. Buttacavoli /s/ Frederick E. Purches

Name & Title: Frank A. Buttacavoli, Exec. VP/COO
Frederick E. Purches, CEO and Chairman

Your Summary:

The first sentence in Section 2 of the Agreement is amended to read: "The term of the Executive's employment as an employee under this Agreement will continue through March 31, 2012, unless terminated at an earlier date in accordance herewith."As additional consideration for the Executive's services hereunder and the covenants contained herein, the Company shall grant Executive an option (the "Option") to purchase 50,000 shares of common stock of the Company (the "Common Stock") pursuant to the Company's 2007 Stock Incentive Plan.The Option (i) shall provide for an exercise price equal to the market price of the Common Stock as of the close of trading on the Nasdaq National Market on the date of this Agreement, and (ii) shall further provide that the Option shall vest and be exercisable immediately with respect to 50,000 shares of the Common Stock covered by the Option.

Figure 5. The original document (left), and the corresponding generated summary (right).

6. DISCUSSION AND CONCLUSION

From our results, we conclude that the summarization of legal/ commercial documents is a challenging task and could further be improved. From Table 1, we see that the F-measure scores for each of the extractive summarization models are satisfactory. LSA performs the best amongst others. LSA captures both the meaning of words as well as the similarity among the sentences. Also, Singular Value Decomposition (SVD) can reduce noise and model latent, the semantic relationship among words and sentences. This leads to an improvement in accuracy. The reason this is a challenging task is that firstly, the formatting and the representation adapted companies to draft legal documents to vary hugely, hence the task to text pre-processing is difficult. Second, the use of current SOTA supervised or unsupervised models for text summarization will fail to work because it is difficult for it to recognize legal jargon and taxonomy. As [4] rightly mentions there is no large dataset available for this domain. This task could further be attempted to solve by training a supervised abstractive summarization model, using Neural networks. This, of course, requires a large number of documents and their corresponding human-written summaries.

REFERENCES

[1] Aria Haghighi and Lucy Vanderwende. 2009. Exploring Content Models for Multi-Document Summarization. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Boulder, Colorado, 362–370. https://www.aclweb.org/anthology/N09-1041

[2] Filippo Galgani, Paul Compton, and Achim Hoffmann. 2012. Combining different summarization techniques for legal text. 115–123.

[3] Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. CaseSummarizer: A System for Automated Summarization of Legal Texts. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations. The COLING 2016 Organizing Committee, Osaka, Japan, 258–262. https://www.aclweb.org/anthology/C16-2054

[4] Laura Manor and Junyi Jessy Li. 2019. Plain English Summarization of Contracts. In Proceedings of the Natural Legal Language Processing Workshop 2019. Association for Computational Linguistics, Minneapolis, Minnesota, 1–11. https://doi.org/10.18653/v1/W19-2201

[5] haiErera, Michal Shmueli-Scheuer, Guy Feigenblat, Ora Nakash, OdelliaBoni, Haggai Roitman, Doron Cohen, Bar Weiner, Yosi Mass, Or Rivlin, Guy Lev, AchiyaJerbi, Jonathan Herzig, Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and David Konopnicki. 2019. A Summarization System for Scientific Documents.

[6] Guy Feigenblat, Haggai Roitman, OdelliaBoni, and David Konopnicki. 2017. Unsupervised queryfocused multi-document summarization using the cross entropy method. In Proceedings of the 40th International ACM SIGIR, pages 961–964.

[7] LexPredict, LLC, acquired by Elevate Services, Inc. in 2018. https://github.com/LexPredict/lexpredict-contraxsuite-samples

[8] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Barcelona, Spain, 404–411. https://www.aclweb.org/anthology/W04-3252

[9] Gunes Erkan and Dragomir Radev. 2011. LexRank: Graph-based Lexical Centrality As Salience in Text Summarization. Journal of Artificial Intelligence Research - JAIR 22 (09 2011). https://doi.org/10.1613/jair.1523

[10] Makbule Ozsoy, Ferda Alpaslan, and Ilyas Cicekli. 2011. Text summarization using Latent Semantic Analysis. J. Information Science 37 (08 2011), 405–417. https://doi.org/10.1177/0165551511408848

[11] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. The Annals of Mathematical Statistics (1951), 79–86.

[12] https://pypi.org/project/sumy/

[13] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Tex Summarization Branches Out. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://www.aclweb.org/anthology/W04-1013