Extracting quotations from long form documents

Rajasekhar Mekala
rmekala@uci.edu

Sai Vineeth Doddala
sdoddala@uci.edu

Agniraj Baikani
abaikani@uci.edu

1 Abstract

The accelerating decline in book reading among youth indicates an imminent issue with their marketing. If the current trend continues, literary reading will disappear in half a century. On the contrary, filmmaking has figured out various ways to gratify the audience instantly. We believe providing powerful quotations while previewing books can engage readers.

Formally, given a book, we extract quotes from it that people could find interesting. We use Transformer models to identify potential quotes in a book, similar to reading comprehension and summarization tasks, and further prune the predictions with classification techniques. Initial results have corroborated that the language models can capture such intriguing lines in long-form documents.

Key Words: Information extraction, Transformers, Reading Comprehension, Multi-head attention

2 Introduction

[1] According to (Jerrick, 2013), 96% of the college students believed movie trailers were effective means of movie marketing. The impact of trailers on audiences is intangible. Unfortunately, providing sample chapters is the only way of marketing when it comes to novels, comics, and storybooks. Over the last few years, book reading has been on the decline. Studies like (Bradshaw, 2004) attribute this to the lack of instant gratification in reading books and the ever-reducing human attention spans. It is hard to communicate the book's essence even after the user spends a lot of time and effort going through the sample chapters. Currently, there are no effective ways to capture an audience's attention in a short time. Conversely, people are often intrigued by inspiring quotations on social media, which mainly originate from books. Quotes often help understand the characters in the books. Although some of the latest books have these quotes on their covers, they may not cover all the quotes. Also, this tradition is recent and was not prevalent in books before this decade. With a similar motivation to movie trailers, good quotes from a book should provide better insights.

Although the problem at hand is to deal with quote extraction, the inherent task can be generalized to controlled information extraction. Practical applications in this direction include specific sentiment extraction, precise attributes-related data extraction, and risk analysis. Most of these applications require manual effort and technical expertise. Currently, large-scale datasets for controlled extraction are very limited except for Question answering and the SOTA approaches are not scalable to long-form documents.

We approach this problem as information extraction and summarization tasks and use two-stage funneling to use out-of-domain knowledge. Our best model achieves an F1 score of 63.12% for SQuAD v1 and ROUGE-L - Recall-Oriented Understudy for Gisting Evaluation (Lin, 2004) score for Extractive summarization of 26.53. Implementing a binary classifier on top of SQuAD v1 improved model performance. We observe that these metrics are not only a good estimate of the model performance and discuss the appropriate metrics that can be used in detail in Section 5. Results indicate that the SQuAD approach outperforms summarization. Although quantitative evaluation metrics were satisfactory, human evaluations are encouraging to extend these approaches to similar tasks as mentioned above.

[1] Code Link:https://github.com/rajasekharmekala/quote-extractor

3 Data

We used the following data for our experimentation.
Goodreads-quotes: A Kaggle dataset (Faellie, 2020) which contains 345k user-added quotes from Goodreads with the corresponding book title, author, number of likes, and associated tags.
Some additional quotations were scrapped from the Goodreads API for the relevant epub books used.
Books: 40 Ebooks for quotes in the Kaggle dataset were collected in epub and pdf formats. The complete list of books is shared in Appendix A.

Data preprocessing is one of the critical challenges to this problem since having the desired quotes is crucial for model learning. After experimenting with many pattern-matching algorithms, about 62.8% were recovered from the Kaggle dataset. The steps followed for this extraction task are mentioned below in detail.

3.1 Pre-processing:

Ebooks were converted from epub to HTML format, and content in the books was preprocessed. Since the Kaggle dataset contained UTF-8 characters, the content from various encodings in various books was converted to standard UTF-8 encoding. As part of data cleaning, the content before the preface and appendix sections was ignored. All the spaces and linebreaks in-between paragraphs in each chapter of the book were ignored.

Even though there were 345k quotes in the Kaggle dataset, since it was a dump of user-uploaded content, many of the quotes were observed to be overlapping. Multiple users have referred to the same books with slightly different book titles. Pattern match (particularly fuzzy match) algorithms were used to make the book titles consistent. As the intent of this work is to identify small independent quotes, the smaller quotations were retained for the overlapping quotes, and on average, about 20% of the quotes had been filtered from each book. Also, the dataset consisted of long paragraphs where the users were trying to provide the context of the quote. Figure 1 shows the quote length histogram in the Kaggle dataset. All quotations containing more than 200 characters were removed. The final processed data contained 191.374k quotations.

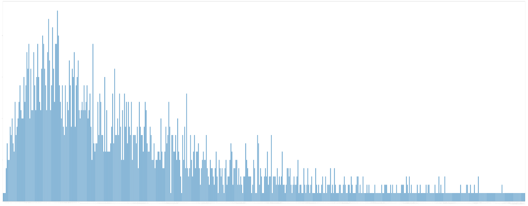

Figure 1: Histogram of quotation lengths in kaggle dataset

3.2 Pattern matching challenges

To create the dataset with contextual paragraphs from ebooks with their corresponding target quotations as labels, the exact positions of a quote were extracted from the paragraphs. Analysis of the Kaggle dataset suggested that human-written quotes did not match the exact text in the books due to typographical errors, multiple versions of books, regional versions(American-English, British-English), and users providing their context to the quotes.

Multiple variants of pattern matching algorithms were applied to find near similar matches to address these issues. At first, a regular-expression fuzzy search allowing three-character mismatches was performed, which resulted in roughly 50% of matches. In the remaining text, sentences in each chapter with the exact first and last words as the quote are considered a probable match. SimHash similarity(finds near similar sentences) with a 90% threshold has been used to identify the most probable matches. Finally, for the remaining quotes, keywords like ADJs, and VERBs (No NOUNS) were identified using POS tagging, and this ordered sequence of keywords was searched in each chapter. As the second and third pattern matching approaches were computationally expensive, a multi-threaded code version has been implemented where books can be parallelly preprocessed.

Below is a quotation from the book “A Game of Thrones” where our pattern matching algorithms did not work, which illustrates the difficulty of this task.

Sentence in the Book: “Ned knew the saying. “What the king dreams,” he said, “the Hand builds.”
Quote in the Kaggle dataset: what the king dreams, the hand builds.

4 Approach

It is subjective to define whether a sentence is an interesting quote. Nevertheless, some attributes can imply if a sentence is a potential candidate. Generally, in books, a paragraph comprises a topic sentence, the developing details, and a concluding sentence. Quotes tend to be the concluding sentence or setup sentence in the paragraphs and capture their essence. They usually make sense as independent sentences without much context and communicate strong human emotion. Often in a book, the author uses a witty line/quote and then explains it in more detail. In some sense, **the rest of the sentences attend** (Vaswani et al., 2017) **to the original quote**.

Although there has not been much prior work on this task, it is similar to the reading comprehension SQuAD - (Rajpurkar et al., 2016) and summarization tasks. Both can be viewed as information extraction tasks with different objectives. In closed-domain question answering, questions are asked from a given paragraph, and the system extracts the best answer phrase in the comprehension. Quote prediction does not require any inherent question. With enough training examples, we can expect models to figure out the attributes of the required phrase/sentence. Similarly, instead of a summary of comprehension, quotations can be targeted in the case of summarization.

With these intuitions, this problem was approached as an information extraction task. A two-stage funneling approach was used since the ebooks do not contain all the paragraphs for the quotes in the Kaggle dataset.

4.1 STAGE-1

The first stage uses transformers like BERT (Devlin et al., 2018) to extract the potential candidates. The following three variants of transformers were used.

- Given a paragraph, identifying the start and end positions of quotation similar to SQuAD (Rajpurkar et al., 2016).

- Given a paragraph, perform Abstractive and Extractive Summarization (Liu, 2019) with pre-trained language models.

- Given a paragraph, classify sentences based on Multi-head Cross Attention (Vaswani et al., 2017) using sentence embeddings.

4.1.1 SQuAD

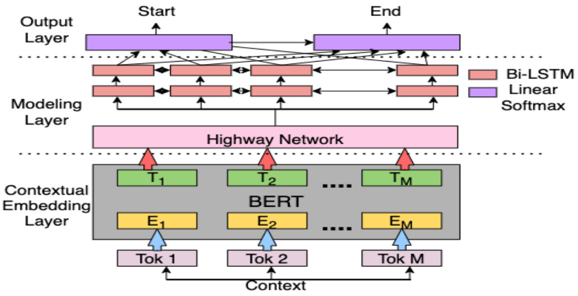

Figure 2: SQuAD Architecture

v1

As shown in Figure 2, The input is provided to BERT-like transformers, similar to question answering. The tokens of only the context are passed instead of tokens for both context and question. The target quote is provided in the original context with the start and end indices. The model tries to calculate the best probabilities for these indices by using the log probability as a loss function during training.

The model always predicts an extract as a potential quote in this approach. Since the number of good quotes in the book is very sparse compared to the total number of paragraphs, predicting a quote for each paragraph severely impacts precision and, in general, the actual motivation. Hence a binary classifier should be applied to the predicted sentences to filter redundant quotes.

v2

In SQuAD v2 training, the model predicts the position indices and the probability of this extract being a potential quote. Unlike v1, SQuAD v2 can handle paragraphs without any quotations. Hence, this can be used independently without another classifier by setting a threshold on the predicted probability.

4.1.2 Summarization

Abstractive

The primary intuition for implementing abstractive summarization is that the quotes capture the essence of a paragraph. The data was preprocessed to map each quote with its corresponding paragraph. The input to this T5-small transformer (Raffel et al., 2019) is a paragraph passed as a document. Furthermore, the target summary is the quote (can be more than 1) in the paragraph. The model tries to generate the summary by learning the most significant details from the paragraphs.

Extractive

Extractive summarization was approached as a binary classification task where each sentence in the context is classified if it is a part of the quotation. Figure 3 shows the architecture of the training procedure. The contiguous sentences are then assembled based on the predicted scores. The classifier used BERT representations for both sentence and context to predict the probability.

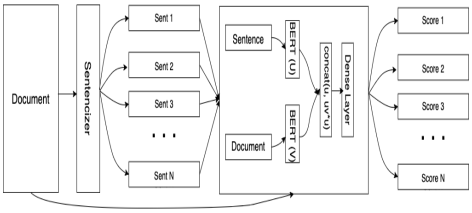

Figure 3: Extractive summarization (Victor, 2021)

4.1.3 Multihead attention on sentences

The third variant is based on the observation that all the sentences in the paragraph point to the quotation. These relations across sentences can be captured using cross attention on sentence representations. The context paragraph and the quote were split into sentences. The sentences were then encoded using a 192-dimensional LSTM. A self-attention block with six heads was used, followed by a dense layer for classification. All sentences that are part of the quotation were given a label of 1. Figure 4 shows the architecture of this approach. During inference, neighboring positively labeled sentences were concatenated to find the best prediction.

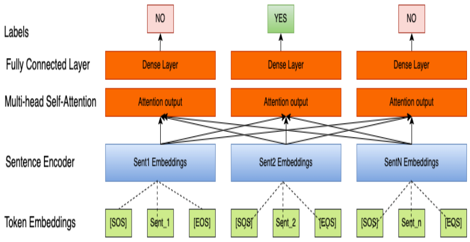

Figure 4: Multihead Sentence Attention

4.2 STAGE-2

The candidates obtained from the first stage are pruned using a binary classifier trained on the complete Kaggle dataset. The classifier was trained in two different settings. 1) Use only the content from the ebook data. 2) Use the Kaggle dataset along with the ebook data. The details of the classifier training are mentioned below.

4.2.1 Classifier

In this classifier, an LSTM-based (hidden-size 128) encoder was used as sentence embedding. In the first approach, quotations were extracted only from the ebooks data. The remaining text from the books was broken down into sentences and treated as negative samples in our training. As shown in Figure 5, there is a significant imbalance between positive and negative labels. To reduce this imbalance, the negative labels were sampled (50%) during data preparation.

As part of the second approach, we used quotes in the Kaggle dataset as positive samples. A similar approach as above was used to create the negative samples from ebooks data. No sampling is required as distribution is almost balanced.

5 Experiments

For all the experiments, the preprocessed text for each chapter is split into multiple blocks with 500 subword tokens, and the blocks are created with a stride of 100 subword tokens. The subword overlap is ensured to not lose any quotations due to this segmentation. The dataset is created by maintaining the id, book title, chapter name, and quotes for each contextual block. From all the 40 books, 9442 quotes were recovered, i.e. 62.8% of all the quotes available in the collected books.

5.1 Results

SQuAD v1: One of the fundamental goals is to answer if the task is hard enough to require larger transformers. To this end, experiments were conducted on different-sized transformer models. Table 1 shows the performance of various pre-trained models on the SQuAD task. Results indicate that the performance has improved as the size of the model increases.

SQuAD v2: Table 3 shows the comparative results of SQuAD v1 with v2. Exact Match, F1 score, and PPS (described in Section 5.2) metrics were used for evaluation. The results indicate that SQuAD v1 with a separate classifier significantly outperformed SQuAD v2.

Abstractive Summarization: ROUGE metrics are used for this task which is the most

Model	Exact Match	F1	PPS
distil-bert	23.01%	39.48%	51.04
bert-base	28.36%	45.96%	58.57
roberta	**30.27%**	**49.96%**	**63.12**

Table 1: Comparision of performance of transformers of different sizes for SQuAD v1

commonly used metric for summarization tasks. It is based on computing the precision and recall scores for the overlapping words. The results of abstractive summarization with the T5-small transformer were not encouraging, as most generated summaries cannot preserve exact quote sentences. Even understanding improvements with larger LM models (.i.e., T5-base, T5-large) was restricted due to memory constraints.

Extractive Summarization: Max Rouge score metrics were used between the given sentence and each sentence in quotes. As shown in Table 2, extractive summarization showed better results by extracting the exact quote sentences from a paragraph.

Multihead attention on sentences:
Since this approach involves breaking text into sentences, sentence-level Exact Match, and F1, PPS scores were used for evaluation. Table 4 shows the results in comparison with SQuAD v1. Although the results indicate that SQuAD v1 outperforms this approach, the results of this approach were satisfactory since this model does not require any pre-training and uses significantly fewer parameters than BERT-like models.

Classifier The purpose of training the classifier in two settings is to study if out-of-domain quotes (not in the ebooks) in the complete dataset help in enhancing the notion of quotation than just learning the in-domain quotations. Table 3 shows the results of this experiment. Results indicate a significant improvement in the classifier trained on out-of-domain quotations.

Model	Rouge1	Rouge2	Rougel
Abstractive (T5-small)	24.12	13.46	22.69
Extractive	29.61	15.27	26.53

Table 2: Performance comparison of summarizers

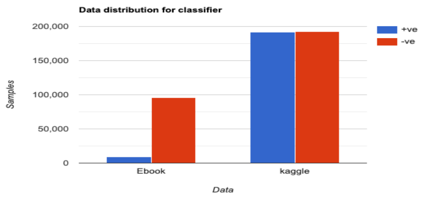

Figure 5: Total train, validation data sizes for classifier

5.2 Analysis

As shown in Appendix B, the phrases predicted by Roberta are not always complete sentences. We compute two metrics better than Exact Match and F1 scores to overcome this issue. 1) Partial Phrase Selection(PPS) - If the predicted phrase is a substring of the actual target (quotation), then the prediction is counted as correct. 2) Interpolated score(IS). The predicted phrase is interpolated to the closest sentence it is part of before computing the F1 and Exact match scores on sentences. These results can now be compared with the Multihead attention on sentences. Table 4 shows this comparison.

The results were separately computed on context blocks that have and do not have quotations. We observe that using classifier after transformer maintains almost all the correctly predicted quotes in stage-1 and filters some of the paragraphs that were predicted to have quotations. However, even after filtering using the classifier, many paragraphs are still predicted to contain good lines. It is interesting that the SQuAD v1, when used with a separate in-domain classifier, outperforms SQuaADv2, which is not trivial to reason.

As discussed in section 5.1, the results indicate that summarization models are not performing well in quote extraction. From our observations, abstractive generated summary mainly consists of influential words in a paragraph and cannot preserve exact quote phrases. Extractive improved it by extracting the exact quote sentences from a paragraph. From our observations, SQuAD v1 with binary classifier outperformed summarization. The primary issue for summarization is data imbalance due to a very sparse set of sentences being part of the target quotes—furthermore, memory constraints in using larger models for deriving representations for sentences and paragraphs for improved results. Exploring imbalance rectification approaches (like undersampling), larger LM models, and tuning hyperparameters

		Has-quote (665/5440)		No-quote (4875/5440)
Model	**Exact Match**	**F1**	**PPS**	**Precision**
Squadv1	30.27%	49.96%	63.12%	0.22 (empty prediction)
Squadv2	21.98%	28.15%	31.15%	10.96
Squadv1+classifier (ebook data)	28.50%	46.08%	60.38%	32.06
Squadv2+classifier (ebook data)	19.33%	25.98%	29.17%	31.02
Squadv1+classifier (kaggle data)	**29.67%**	**46.74%**	61.45%	**41.47**
Squadv2+classifier (kaggle data)	19.44%	26.13%	30.12%	33.42

Table 3: Comparision of SQuAD approaches(roberta-base) with and without classifiers on test data

Model	IS Exact Match	IS F1	PPS
roberta(kaggle data)	36.64	48.43	62.17
Multihead attention	34.23	43.12	57.92

Table 4: Comparing Multihead attention with Roberta

can improve results.

The predictions from the test dataset are shown in Appendix B. More Human evaluations (not included here) have surprised us that there is no quantifiable way to say that the sentences predicted by the model are not good lines. In fact, Some of them are great lines capturing motivations discussed in Section 4. In many cases, the user-provided quotations were not independent. However, the predictions by the model were better than the ground truth; the predictions were still treated as false positives. Although manual evaluation indicated high-quality predictions, instances like these were considerable in number, resulting in poor performance on evaluation metrics.

After these observations, it only makes sense to have an even bounded problem statement .i.e, (looking for particular emotions in the quotations, inspirational or motivational quotes, life quotes, e.t.c). The Kaggle dataset we used contains such tags for some of the quotes. A possible future work would be using these tags to extract more specific lines from books. Nevertheless, the tags data has its own set of challenges. Figure 6 shows the imbalance of the various tags, suggesting that it would require significant effort and can be considered a problem statement. Hence this approach is currently out of the scope.

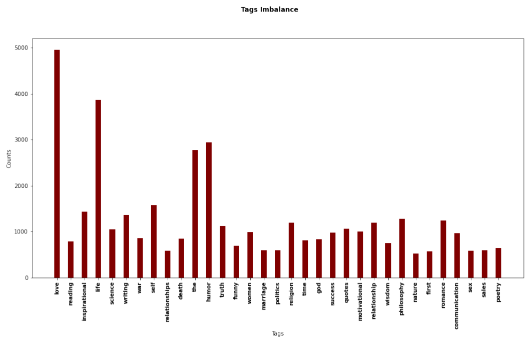

Figure 6: Class imbalance of tags in quotations

6 Conclusion

Overall, we observe that approaching quote extraction as a reading comprehension task is better than summarization. Our results show that having a global dataset capturing the required attributes is better than having an in-domain dataset. Though multi-head attention approach did not outperform BERT-like models, the results were good enough to identify local relations in long-form documents.

Given the variations in genres and authoring styles for this task, the observed results in Section 5 indicate that while the implemented approaches worked rather well, there is much scope for improvement. Since a quotation is a subjective entity, it may not be possible to completely solve this task for practical user data. This limitation fundamentally restricts our ability to extract all quotes in a book. We can only aim to predict quotations that are more probable. However, the results are encouraging enough to extend these approaches to other similar problems.

References

Tom Bradshaw. 2004. Reading at risk: A survey of literary reading in america.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lupe Faellie. 2020. Goodreads quotes.

David Jerrick. 2013. The effectiveness of film trailers: evidence from the college student market. *UW-L Journal of Undergraduate Research*, 16:1–13.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu. 2019. Fine-tune bert for extractive summarization.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Dibia Victor. 2021. How to implement extractive summarization with bert in pytorch.

A Book Titles

A Clash of Kings, A Game of Thrones, A Short History of Nearly Everything, Alexander Hamilton, All The Light We Cannot See, Americanah, Between the World and Me, Big Magic Creative Living Beyond Fear, Catch-22, Cloud Atlas, Comanche Moon, Creativity, Inc. Overcoming the Unseen Forces That Stand in the Way of True Inspiration, Dark Prince, Deep Work Rules for Focused Success in a Distracted World, Extremely Loud and Incredibly Close, Fahrenheit 451, Fire and Fury Inside the Trump White House, Gone Girl, Lean In Women, Work, and the Will to Lead, Life of Pi, Little Women, Madame Bovary, Quiet The Power of Introverts in a World That Cant Stop Talking, Ready Player One, Shantaram, Siddhartha, Steve Jobs, The Black Swan The Impact of the Highly Improbable, The Book Thief, The Fault in Our Stars, The Goldfinch, The Handmaids Tale, The Lonesome Dove Chronicles, The Perks of Being a Wallflower, The Sense of an Ending, The Unbearable Lightness of Being, Tools of Titans The Tactics, Routines, and Habits of Billionaires, Icons, and World-Class Performers, Tribe of Mentors Short Life Advice from the Best in the World, Tuesdays with Morri

B Human Evaluation

Table 5: Human evaluation examples

In the table below, examples 1 to 5 consists of exact matches. Examples 6 and 7 covers partial matches part of the quotes. Example 8 outputs different important sentence. Example 9 and 10 generates random or empty sentences

Example No.	**Predicted**	**Context** (only relevant part of paragraph is shown)
1	misery loves company, and madness calls it forth.	a sound without shape or colour sounds strange. to be blind is to hear otherwise. the words came again, " is someone there? " i concluded that i had gone mad. sad but true. **misery loves company, and madness calls it forth.** " is someone there? " came the voice again, insistent. the clarity of my insanity was astonishing. the voice had its very own timbre, with a heavy, weary rasp. i decided to play along.
2	love can be obtained by begging, buying, receiving it as a gift, finding it in the street, but it cannot be stolen.	which knows how to give so many sweet things! you are learning easily, siddhartha, thus you should also learn this : **love can be obtained by begging, buying, receiving it as a gift, finding it in the street, but it cannot be stolen.** in this, you have come up with the wrong path. no, it would be a pity, if a pretty young man like you would want to tackle it in such a wrong manner. "
3	stay humble or get humbled.	that right there, that's a guy who's going to make it, who's going to get it right. the arrogant guys, who lacked humility, they couldn't take criticism from others - - and couldn't even do an honest self - assessment because they thought they already knew everything. **stay humble or get humbled.** " on the importance of detachment " i was probably 20 or 21 years old. i was in my first seal platoon.
4	the most important thing in life is to learn how to give out love, and to let it come in.	" we're tuesday people, " he said. tuesday people, i repeated. morrie smiled. " mitch, you asked about caring for people i don't even know. but can i tell you the thing i'm learning most with this disease? " what's that?" **the most important thing in life is to learn how to give out love, and to let it come in.** " his voice dropped to a whisper. "
5	go and make yourself useful, since you are too big to be ornamental	mercy on us, this will never do, " thought jo, adding aloud, " go and sing to me. i'm dying for some music, and always like yours. " " i'd rather stay here, thank you. " " well, you can't, there isn't room. **go and make yourself useful, since you are too big to be ornamental.** i thought you hated to be tied to a woman's apron string? " retorted jo, quoting certain rebellious words of his own. " ah, that depends on who wears the apron! " and laurie gave an audacious tweak at the tassel.

6	he must learn to face his fears. he will not be three forever. and winter is coming.	she could feel the eyes watching her, but she did her best to ignore them. " arya is already in love, and sansa is charmed and gracious, but rickon is not quite sure. " " is he afraid? " ned asked. " a little, " she admitted. " he is only three. " ned frowned. " he must learn to face his fears. he will not be three forever. and **winter is coming.** " " yes, " catelyn agreed. the words gave her a chill, as they always did. the stark words. every noble house had its words.
7	so many people enter and leave your life!	i said, " that’s hilarious, " because it must have been for him to crack up so much. " hilarious! " he said. " it is! i never heard from her again! oh, well! **so many people enter and leave your life! hundreds of thousands of people! you have to keep the door open so they can come in! but it also means you have to let them go!** " he put a teakettle on the stove. " you’re wise, " i told him.
8	one opportunity leads directly to another, just as risk leads to more risk, life to more life, and death to more death.	it would inspire hans hubermann to come up with a plan to help the jewish fist fighter. and it would show me, once again, that one opportunity leads directly to another, just as risk leads to more risk, life to more life, and death to more death. **in a way, it was destiny.** you see,
9	i like dogs better than knights	" i like dogs better than knights. my father’s father was kennel-master at the rock. one autumn year, lord tytos came between a lioness and her prey. my grandfather lost a leg, so lannister paid him for it with lands and a towerhouse, and took his son to squire. the three dogs on our banner are the three that died, in the yellow of autumn grass. **a hound will die for you, but never lie to you.** and he’ll look you straight in the face. "
10	[empty]	tell me, my dear : you’re not taking control of your son’s upbringing? you don’t force him? you don’t beat him? you don’t punish him? " " no, vasudeva, i don’t do anything of this. " " i knew it. you don’t force him, don’t beat him, don’t give him orders, because you know that’soft’is stronger than’hard ’, **water stronger than rocks, love stronger than force.** very good, i praise you.