University *of Ljubljana*
Faculty *of Computer and Information Science*

IMapBook Collaborative Discussion Classification

Uroš Polanc

Abstract

Automatic text classification has been considered as a vital method to manage and process a vast amount of documents in digital form, which is widespread and continuously increasing. That brought a wave of new tools for student education, such as the IMapBook platform. It gives students a way to communicate and discuss the books they are reading. In our assignment, we will try to classify those discussions based on the book it's relevant to, and the type of message.

Keywords
IMapBook, Text Classification, TF-IDF, Word Cloud

Advisors : Slavko Žitnik

Introduction

With the rapid development of the internet and big data, digital information is increasing at a high rate. With this technologies and applications such as IMapBook [1], a web-based application that allows for discussions on reading materials as well as interactive games, also started emerging.

In our assignment, we will look at some IMapBook collaborative discussions on different books. For that, we first need to look at some vital information, that we extracted from the discussions and the book texts. We will create two different message classifications, based on :

- book ID (3-class),
- message type (16-class).

By checking the manual classifications in the dataset, we noticed some were misspelled and some were incorrectly written. Therefore, we applied some fixes to the manual classifications. The errors such as excess whitespace, lower and upper case, etc... were removed. Classifications where there were multiple classes, were simplified into one class, to simplify the classification process. The final fixed distribution is not intraclass uniform as can be seen in figures 1 and 2.

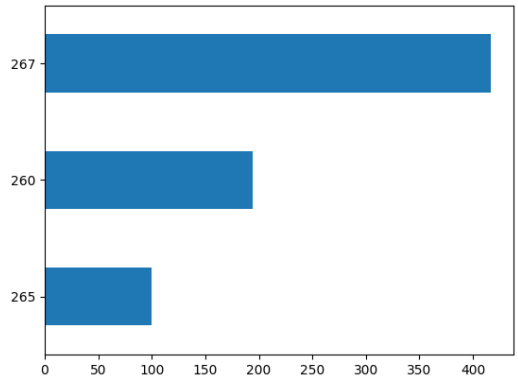

Figure 1. The distribution of book IDs in the dataset.

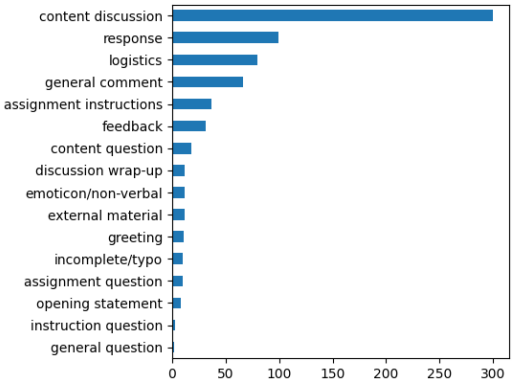

Figure 2. The distribution of message types in the dataset.

To better understand the contents we are trying to classify, we checked the top unigrams of the book texts and displayed the top 100 in a word cloud image (see figures 3, 4 and 5). We also calculated the TF-IDF of the books as well as the collaborative discussions (as a whole) and displayed the top 15 stemmed words for each (see figure 6).

Figure 3. The word cloud for book ID260 and ID261.

Figure 4. The word cloud for book ID264 and ID265.

Figure 5. The word cloud for book ID266 and ID267.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ID260 and ID261	king	door	arena	open	one	person	upon	fair	subject	youth	princess	accus	seen	would	lady
ID264 and ID265	maier	bag	bottega	veneta	said	elev	design	put	label	milan	tabl	interview	watch	fashion	one
ID266 and ID267	design	future	world	architect	one	crisi	earth	food	fundament	increas	climat	exhibit	power	contemporari	museum
XLSX Data	design	think	would	im	door	good	future	idea	respons	way	submitted	could	discuss	perspect	articl
XLSX Discussion	design	think	articl	need	would	chang	future	door	way	sustain	live	tiger	lover	prototyp	perspect

Figure 6. The top 15 words based on TF-IDF for each set.

From this we can say that the texts are contextually different, aside from a few words such as 'one', the top frequent words are more or less unique to their books.

Existing solutions

There are many approaches to automatic NLP text classification, which fall into three types of systems :

- rule-based systems,
- machine learning-based systems,
- hybrid systems.

Rule-based systems classify text into organized groups by using a set of handcrafted rules. But these systems require deep knowledge of the domain they are trying to classify. On the other hand, machine learning-based systems can learn the different associations between pieces of text by using pre-labeled examples as training data. Some of the more popular text classification algorithms [4]inlcude Naive Bayes, Support Vector Machines, and deep learning. While hybrid systems apply both handcrafted rules and machine learning approaches. The current state-of-the-art consists of approaches such as BERT [2] or ELMo [3].

Initial ideas

For our initial ideas, we plan to design some baseline classifiers, such as Naive Bayes, Random Forest, Logistic Regression, Support Vector Machine, and Majority Voting. From there we will proceed to include neural networks, and maybe try to include at least one of the state-of-the-art approaches, such as BERT or ELMo as well.

References

[1] Grandon Gill and Glenn Smith. Imapbook: Engaging young readers with games. *Journal of Information Technology Education: Discussion Cases*, 2:10, 01 2013.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019.

[3] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018.

[4] Tomas Pranckevicius and V. Marcinkevicius. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Balt. J. Mod. Comput.*, 5, 2017.