ASPECT EXTRACTION FROM BOARDGAME COMMENTS

Date: 28/06/2024

Course: Natural Language Processing

Instructor: Alfio Ferrara

Author: Hasan Kuşpınar

TABLE OF CONTENTS PAGE

A. INTRODUCTION

Aspect Extraction is a part of sentiment analysis which focuses on identifying and extracting people's opinions from texts over different platforms such as blogs and forums (Rana et al. 2016). There are different types of classifications for sentiment analysis, which are document level, sentence level and aspect level (Rana et al. 2016). In this project the focus will be on aspect extraction. The main goal of aspect extraction is to find some aspects that are mentioned in a text either explicitly or implicitly. In explicit aspect extraction the aspects are used in the text which makes the task easier. However, in implicit the aspects are not mentioned explicitly so they need to be derived from the text.

B. METHODOLOGY

1. The goal of the project

In this project it is aimed to study some user comments from a boardgame website to extract aspect from them. The aspects are luck, bookkeeping, downtime, interaction, bash the leader and complex vs complicated.

To deal with this problem, firstly a chat-gpt based aspect extraction mechanism is developed. This mechanism uses chat-gpt in the same way that the ordinary users use it to give a prompt and receive a response from it. Chat-gpt is used for this purpose since according to Xiancai Xu, it performs pretty similar and, in some cases, better than the other models (Xu et al. 2023). Afterwards, the mechanism is evaluated with the given boardgame dataset.

2. Data

The data is gathered using Boardgamegeek website api. First of all, manually downloaded csv is tried however the file does not have a clear ranking; therefore, ranking based on 'rank', 'bayesaverage' and 'average' columns is checked. Bayes is found to be the best option and the dataframe is ranked based on it. After ranking the dataframe, it is compared to the ranking on the website and it can be seen that they differ (see Figure 1).

Board Game Rank ▲		Title	Your Rating	Geek Rating	Avg Rating	Num Voters
1		**Brass: Birmingham** (2018) Build networks, grow industries, and navigate the world of the Industrial Revolution.	N/A	8.416	8.60	45595
2		**Pandemic Legacy: Season 1** (2015) Mutating diseases are spreading around the world - can your team save humanity?	N/A	8.381	8.53	53399
3		**Gloomhaven** (2017) Vanquish monsters with strategic cardplay. Fulfill your quest to leave your legacy!	N/A	8.354	8.59	62047
4		**Ark Nova** (2021) Plan and build a modern, scientifically managed zoo to support conservation projects.	N/A	8.334	8.53	43124
5		**Twilight Imperium: Fourth Edition** (2017) Build an intergalactic empire through trade, research, conquest and grand politics.	N/A	8.242	8.60	23671
6		**Dune: Imperium** (2020) Influence, intrigue, and combat in the universe of Dune.	N/A	8.229	8.43	45119
7		**Terraforming Mars** (2016) Compete with rival CEOs to make Mars habitable and build your corporate empire.	N/A	8.214	8.36	98703
8		**War of the Ring: Second Edition** (2011) The Fellowship and the Free Peoples clash with Sauron over the fate of Middle-earth.	N/A	8.184	8.54	21255
9		**Star Wars: Rebellion** (2016) Strike from your hidden base as the Rebels—or find and destroy it as the Empire.	N/A	8.171	8.42	32357
10		**Gloomhaven: Jaws of the Lion** (2020) Vanquish monsters with strategic cardplay in a 25-scenario Gloomhaven campaign.	N/A	8.167	8.44	34120

	id	name	bayesaverage	usersrated
0	224517	Brass: Birmingham	8.41602	45595
103807	247030	Terraforming Mars: Prelude	8.40415	15151
1	161936	Pandemic Legacy: Season 1	8.38053	53399
2	174430	Gloomhaven	8.35400	62047
3	342942	Ark Nova	8.33394	43124
4	233078	Twilight Imperium: Fourth Edition	8.24157	23671
5	316554	Dune: Imperium	8.22939	45119
6	167791	Terraforming Mars	8.21379	98703
7	115746	War of the Ring: Second Edition	8.18431	21255
137583	363622	The Castles of Burgundy: Special Edition	8.17803	6121

Figure 1

Therefore, reading the ranking directly from the html seems to be the best way. After reading from the html the dataframe columns needs some cleaning and manipulation. Firstly, numerical info is extracted from the tuples. Then, the game id is extracted from the thumbnail image and the game name is extracted from the title. Also some advertisements are present and they are deleted as well.

Subsequently, having the top 100 games the comments for them are gathered using the api. Moreover, some comments seem to be not English; therefore, using py3langid library the comments are checked to see how many of them are English. And it is seen that only 0.06 are non-English comments (see Figure 2). Because majority is English the non-English comments are removed.

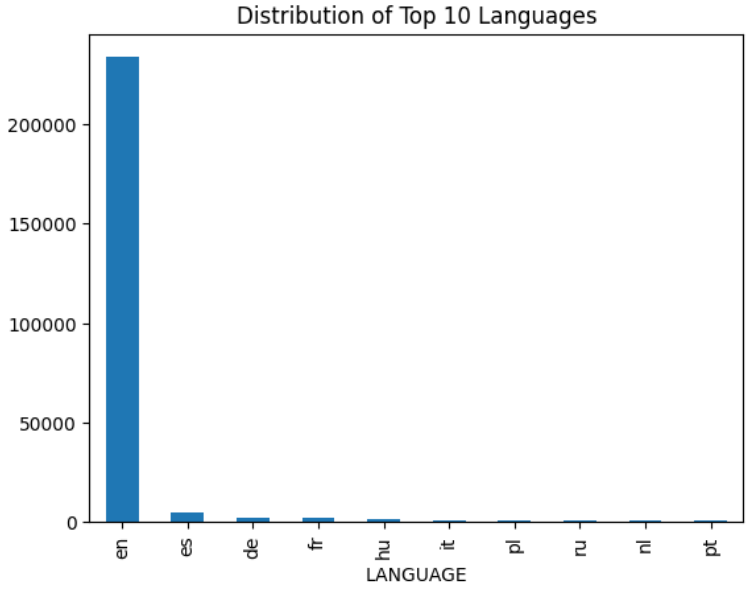

Figure 2

Afterwards, the length of the comments are checked and the histogram in figure 3 shows that as expected the lengths are skewed towards zero which means most of the comments are not that long.

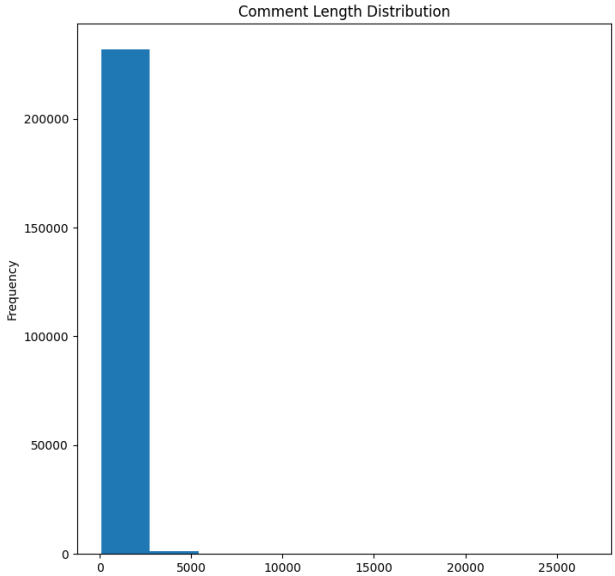

Figure 3

Finally, the number of tokens required to extract aspects from the comments is checked and the results show a pretty high number of tokens. Therefore, it would be better to use only 100000 of the comments.

3. Prompt

To use chat-gpt for aspect extraction a prompt needs to be given to it. This prompt needs to have a clear instruction of what it is going to do, an explanation of the aspects (context), some restrictions for it to follow and an output format (Xu etal. 2023). However, the explanation of the aspects part needs to be enhanced so that chat-gpt makes less errors and is more efficient. The explanation for the aspects are given to chat-gpt to be longer and more detailed. The resulting explanations are put in the prompt. The instructions given in the prompt explains that it needs to extract aspects from a given comment based on the aspect explanations given in the prompt and it needs to strictly follow the restrictions. The restrictions are emphasized and written in uppercase so that they are taken into consideration.

4. Aspect extraction

After the data is pre-processed and the prompt is prepared, aspect extraction needs to be performed on the data. The process is done like the following; by iterating over each game and their comments we get a comment for a game, that comment is sent to chat-gpt with the prompt that was prepared by using the openai class, the response got from it is assigned to the dataframe in a new column and finally the dataframe is turned into a csv and stored.

C. RESULTS

The stored csv's with the responses from chat-gpt has the labels in a list; therefore, they are extracted into a new column so that they can be better analyzed. Subsequently, the results are plotted to see how many times each aspect/category has appeared. After the looking at the plots in figure 4 it can be seen that the most frequent aspect is the 'bookkeeping' and the most frequent category is the 'complex'.

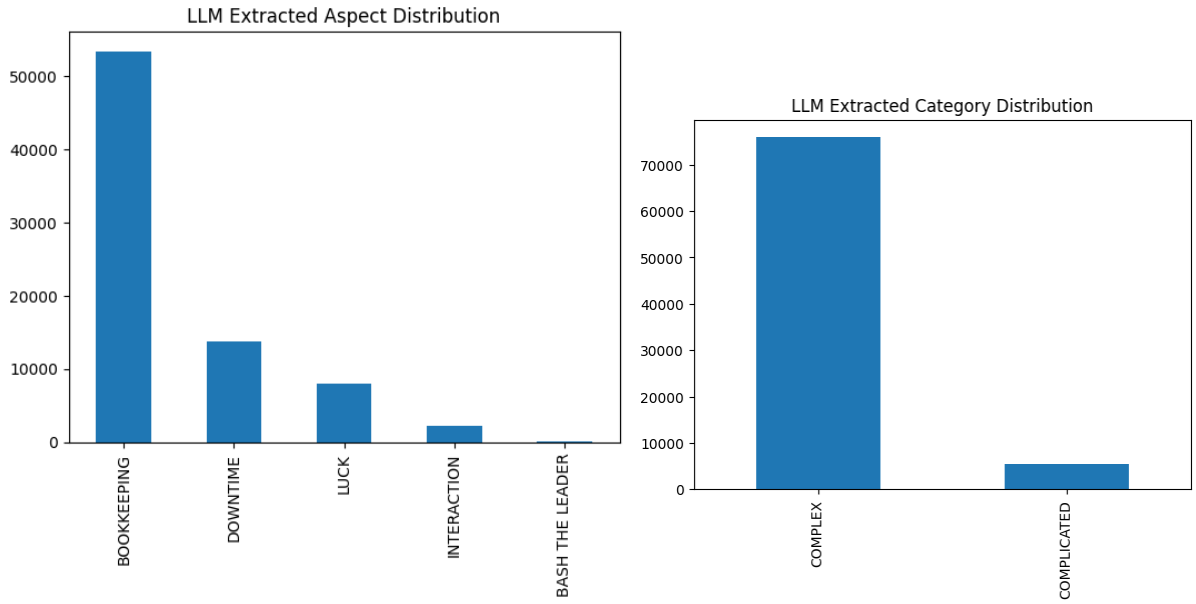

Figure 4

D. CONCLUSION

To sum up, it can be deduced that the current state of the chat-gpt works well for aspect extraction. However, there are downsides as well. First of all, there were some comments that chat-gpt was not able to find a aspect from. And also due to the requirements of high tokens not all of the comments were able to be used. Therefore, since chat-gpt is a developing model with the upcoming developments and updates it might perform a lot better in this task. And due to the interaction and mechanism of it being consistent, the process can be retried in the future after the new developments with minimal changes in the code.

E. REFERENCES

Rana, T. A., & Cheah, Y.-N. (2016). Aspect extraction in sentiment analysis: Comparative Analysis and Survey. *Artificial Intelligence Review*, *46*(4), 459–483. https://doi.org/10.1007/s10462-016-9472-z

Xu, X, & Zhang, J-D. & Xiao, R, & Xiong, L (October 2023). The Limits of ChatGPT in Extracting Aspect-Category-Opinion-Sentiment Quadruples: A Comparative Analysis. arXiv:2310.06502v1 [cs.CL] 10 Oct 2023