

Uncover the Hidden Secret in the Wordle Results

Summary

Since Wordle has become a popular puzzle game, it has accumulated a large amount of data. In this paper, we define a series of metrics and build several models to explore the hidden information in Wordle results.

First, after **preprocessing** the given data and analyzing the time series diagram of the number of reported results, we found that the changes can be divided into 3 stages. To forecast the number of reported results, we developed a weighted optimization model based on ARIMA and BP neural network. The prediction interval is then given using the Bootstrap method. We packaged this process as **ARIMA-BP Interval Prediction Model Based On Bootstrap**. Thus, we finally predicted the interval prediction value obtained on March 1, 2023 at 95% confidence level to be about (19504.74, 20383.26)

Then, we defined 3 qualitative and 4 quantitative attributes of words and used them to build a **Multiple Linear Regression Model** with the percentage of hard mode's players. We found that the proportion will decrease by an average of 0.618 when the initial letter changes from a vowel to a consonant while it will increase by an average of 0.017 for each one-unit increase in word internal distance.

After that, we made the percentage distribution prediction of the reported results based on **LSTM Model**. To ensure the percentage is around 100%, we first processed the component data using a **spherical coordinate transformation**. Then we use them as output variables, the 7 word attributes and number of results as input to train our LSTM model. The prediction of EERIE based on this are [2%, 11%, 25%, 24%, 19%, 14%, 5%]. We changed the model's parameters and added noise to do **sensitivity analysis**. Meanwhile, we introduced COV to measure the uncertainty of the model prediction and found that it is around 0.4. For **error analysis**, we use MSE, RMSE and R to measure the prediction accuracy, and their values are shown in Table 7.

We extracted 6 indicators: RDC, TE, SK, NFC, NON, and HL to measure the difficulty of words. We built a **GMM Clustering Model** based on these indicators and thus classifying 5 difficulty levels. We classified the word EERIE as difficulty level III.

In addition, by counting the frequency of each letter in five positions, we found S as the initial letter has the most frequency and more specific statistical results are shown in Table 9. We also used the **Association Rule Model** based on **Apriori algorithm** to mine the word combination pattern in Wordle. Ideally, we found that the letters A, S and F, TL usually appear together in Wordle.

Finally, we evaluated and refined the model and reported the findings in a letter to the Puzzle Editor of the New York Times.

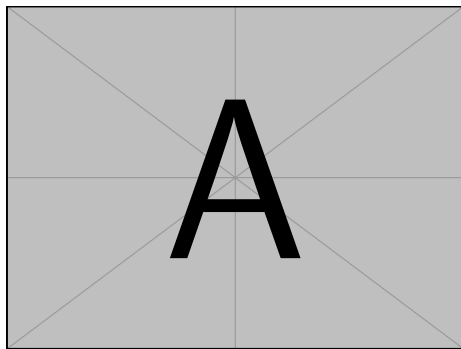
Keywords: ARIMA-BP, LSTM, GMM, Apriori Algorithm, Word Attributes

Contents

1 Introduction

1.1 Background

Wordle is an online word puzzle game invented by Josh Wardle during the epidemic. The New York Times newspaper, which is well known for the games it publishes, has bought Wordle on February 2021^[2]. Wordle only allows one game to be played per day, and every player in the world plays to guess the same five-letter word in six tries or less each day^[2]. Players can play in regular mode or hard mode. They can share their scores via Twitter, thus attracting more people to play and share.



Back in October 2021, less than 5,000 visits registered to its web page, while by January 2022 traffic had skyrocketed to over 45 million. Some of us also love this game. Figure 1 shows one result that we got. The green tile indicates that the secret solution word has the letter in the precise location. The yellow tile implies that the answer has the letter but not at the correct location. Grey tiles indicate the letters are not contained in the solution at all^[2].

Figure 1: Wordle Game

Now we have a file of daily results from January 7, 2022 to December 31, 2022. This file includes twelve key variables that are crucial in our later research. In the file, the percentages of the number of people for the seven tries may not sum to 100% due to rounding.

1.2 Problem Restatement

By analyzing the above background, we summarize the tasks that need to be addressed as follows:

- Develop a model to explain the daily variation in the total number of people reporting scores on Twitter and use it to give a prediction interval for the total number of people on March 1, 2023.
- Determine whether the attributes of words affect the percentage of players who choose the hard mode and explain the obtained results accordingly.
- If given a future date and the specific word, build a model to predict the percentage of 1-X tries in this day. After that, the word EERIE on March 1, 2023 should be used as a specific example of the model prediction, while analyzing the uncertainties of the model and the accuracy of the prediction.
- Develop a model to classify the difficulty of the words and identify the attributes of them under each category. Use this model to determine how difficult the word EERIE is? Finally, discuss the accuracy of the classification model.
- List and describe some other interesting features of the dataset.

1.3 Our work

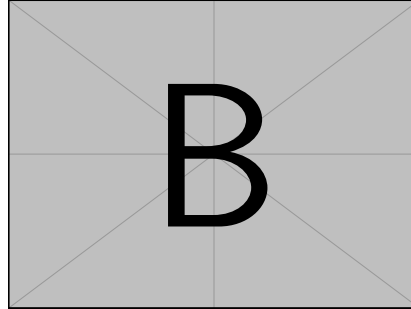


Figure 2: Our Work

2 Assumptions and Notations

Assumption 1: The number of daily online gamers in the data set is a time-dependent set of series and is independent of seasonal changes.

Explanation: Therefore, we can use ARIMA as well as LSTM models to predict a series of data for March 1, 2023.

Assumption 2: The scores reported by players on Twitter on a daily basis are normal and reliable.

Explanation: To ensure that the model we build based on the dataset can reliably predict a range of data for March 1, 2023.

Assumption 3: Assuming that Wordle's development process is consistent with the game's life cycle theory.

Explanation: This assumption reduces the impact of external uncertainties on Wordle game predictions, thus making the whole process of prediction and analysis more efficient.

In this work, we use the symbols in Table 1 in the model construction. Other non-frequent-used symbols will be introduced once they are used.

Table 1: Notations

Symbol	Definition
w_A	Longitude within the i -th Wildfire Grid
w_B	Latitude within the i -th Wildfire Grid
P_{Ai}	The area of the i -th grid
P_{Ai}	the distance d_{ki}
L_i	Score for evaluating the k -th wildfire grid
NF	Noise factor
C_i	the ratio of players at different guess counts

3 ARIMA-BP Interval Prediction Model Based On Bootstrap

3.1 Data Processing

By reviewing the given data, we found that there are no missing values, but there are outliers. One of them does not exist and two of them have less than 5 letters. We deleted these three rows of data due to the difficulty of obtaining the true values of these words. The two remaining outliers are caused by data entry errors. In order to avoid an excessive reduction in the amount of data, we changed them by combining the previous and later data as well as the semantics of them. In addition, the sum of the percentages in the original data are all in [98%, 102%], which is not much different from 100%, so they are reasonable and do not need to be processed. In summary, the preprocessing of the raw data is summarized in Table ??.

Table 2: Data Processing

Database Names	Database Websites
Google Scholar	https://scholar.google.com
Wikipedia	https://www.wikipedia.org
wolframalpha	https://www.wolframalpha.com

3.2 Point Prediction Based on a Combined ARIMA-BP Model

3.2.1 Variation Explaining

In the data preprocessing, although there are word entry errors, their corresponding numbers of reported results are not affected, so the complete reported data are analyzed in this problem. By observing and analyzing the characteristics of the changes in the number of reported results from Jan. 7, 2022 to Dec 31, 2022, we found that they can be divided into 3 phases, as shown in Figure 3.

Based on the game lifecycle and player lifecycle theory [4], and combined with Wordle's game features, we explain the reasons for the changes as follows.

- Phase 1: Rapid Growth Period (January 7 2022-February 2 2022)

Since the launch of Wordle's web version, it has been updated with only one puzzle a day. This artificial scarcity enhances players' desire for challenge and anticipation. The sharing function of Wordle uses emoji blocks to refer to the results of the game, which is very recognizable and easy to spread, while also avoiding spoilers, thus attracting more new players to play Wordle. In addition, the traditional popularity of crossword puzzles and the easy-to-play game mechanics have contributed to Wordle's further popularity. During this period, Wordle's overall growth in the number of results reported is 348.85%, with the number peaking at 361.908 on February 2.

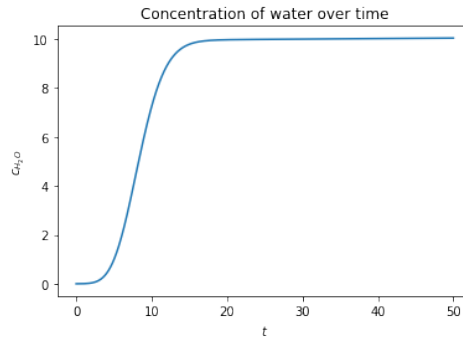


Figure 3: Changes in the number of reported results

- **Parse 2: Rapid Decline Period (February 3, 2022 - May 29, 2022)** During this period, Wordle experiences an overall 84.29% decline in the number of reported results, which is the general nature of internet fads. When the popularity of the game reaches its peak, it faces a significant loss of players due to gamers' declining interest and boredom with the game. The singularity of Wordle's game mechanic can contribute to this. Furthermore, the emergence of competing games in the market such as "Words with Friends" and Wordle's pirated games makes Wordle lose players further.
- **Parse 3: Stable Reduction Period (May 30, 2022 - December 31, 2022)** The overall decrease in the number of results reported for Wordle in this phase is 64.14%. Rapidly losing players in Phase 2 go to play Wordle generally because of its popularity among the general public and they will quickly leave when Wordle is no longer hot or the next trend emerges. The remaining players are often loyal Wordle players or players who have developed user stickiness due to the game's social circle. At this point, the number of reports will still drop but not by much.

3.2.2 Reasons for Model Selection

Predicting the number of reported results is a problem in the field of time series analysis. The traditional ARIMA model is capable of extracting deterministic information from historical data in order to predict the trend of variables over time, however, this model requires extensive testing before application, and the process of determining the order of the model is subjective to some extent [5]. In recent years, with the development of artificial intelligence, machine learning algorithms have been widely applied to the field of data classification and prediction. Among which, BP neural network model which is efficient and convenient can effectively break the limits of traditional time series prediction models. Considering these factors, we decide to combine the advantages of traditional time series models and machine learning models. It means that first of all, we use ARIMA model and BP neural network model to predict the number of reported outcomes separately, and then we use weighted average to combine the two models to obtain a more reliable ARIMA-BP combined prediction model.

3.2.3 ARIMA Forecasting

We first used the traditional ARIMA model for point prediction of the number of reported results, and this process is shown in Figure ??:



Figure 4: ARIMA model building process

First, we performed ADF smoothness test on the variable “Number of reported results in the preprocessed data set and found that it wasn’t smooth. Therefore, we differenced the series to the first order and tested the differenced series again. After that, we observed the ACF and PACF plot of the differenced series, and found that the ACF plot showed the characteristics of trailing tails and the PACF plot showed the characteristics of 1st order truncated tails. Therefore, the following ARIMA(1.1.0) model can be developed for the original series:

$$\begin{cases} (1 - \phi_1 B)(1 - B)x_t = \varepsilon_t \\ E(\varepsilon_t) = 0, \text{Var}(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t, \\ E(x_s \varepsilon_t) = 0, \forall s < t \end{cases}$$

where x_t is the t^{th} time series value, B is the delay operator, ϕ_1 is the coefficient of the moving self-averaging polynomial.

3.2.4 BP Neural Network Prediction

BP neural network is a multilayer feedforward neural network trained according to the error back propagation algorithm, which can iterate and repair its own weights continuously based on the relationship between input and output variables so as to finally estimate the exact functional relationship^[2]. Its algorithm is explained by the following pseudo-code. :

Algorithm 1: Prediction of the Number of Reported Results

Input: Training set D **Output:** Test set E Learning rate η ; Data normalization (normalization formula)

Create network; Train the network

repeat for D

Forward propagation; Backward propagation

until for reaches end condition

Using the network; Inverse normalization of data

BP neural network prediction with test set E **end:** The completed BP neural network trained

3.2.5 Analysis of the Results

The result obtained using the ARIMA(1,1,0) model prediction is shown in Figure ?? and using the BP neural network model prediction is shown in Figure ??:

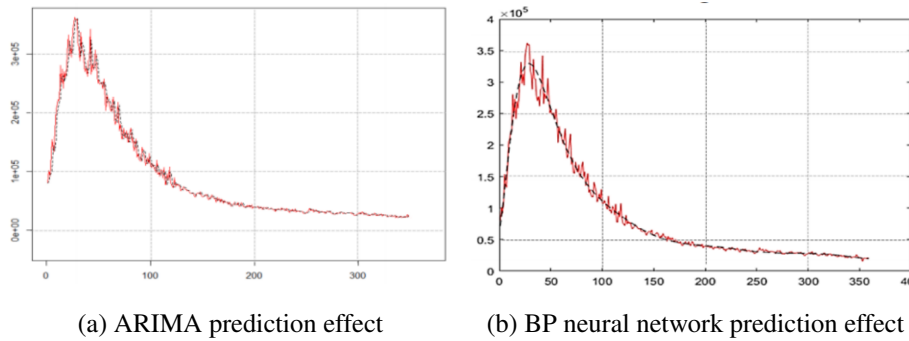


Figure 5: Three images

It can be seen that the ARIMA model has a better effect prediction, and the point prediction P_a obtained using this model for the number of results reported on March 1, 2023 is about 20598.84. BP neural network model has some bias in the early stage but has a relatively good prediction effect in the later stage. The point prediction value P_e obtained using this model on March 1, 2023 is about 19496.05.

3.2.6 Weighted Portfolio Model Forecast

In order to improve the accuracy and robustness of the model prediction, we derived a new combined ARIMA-BP prediction model based on ARIMA and BP neural network model by using the prediction errors to weight the two models. The design idea of this weighted combined forecasting model is shown in Figure ??: We first calculate the average prediction error for each of the two prediction models:

$$\overline{PEA^2} = \frac{1}{N} \sum_{i=1}^N (R_i - P_{Ai})^2, \overline{PEB^2} = \frac{1}{N} \sum_{i=1}^N (R_i - P_{Bi})^2.$$

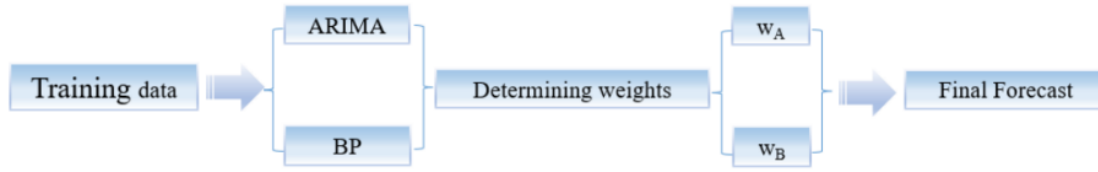


Figure 6: Combined model design

After that, we use the share of the average prediction error squared of either model in the total average prediction error squared of the two models to reflect the weight accounted for by the other model, i.e.

$$w_A = \frac{\overline{PEB}^2}{\overline{PEA}^2 + \overline{PEB}^2} \approx 0.406, w_B = \frac{\overline{PEA}^2}{\overline{PEA}^2 + \overline{PEB}^2} \approx 0.594$$

Thus, the point prediction value of the i^{th} series in the final ARIMA-BP combination model obtained is calculated as:

$$P_i = w_A P_{Ai} + w_B P_{Bi} \quad (1)$$

Through the analysis of the prediction errors above, it is obvious that the overall prediction error of ARIMA model is relatively high, while the BP neural network model may be overfitted. Therefore, the combined ARIMA-BP prediction model can correct the errors of these two models, thus making the prediction results more realistic. The final point fore-

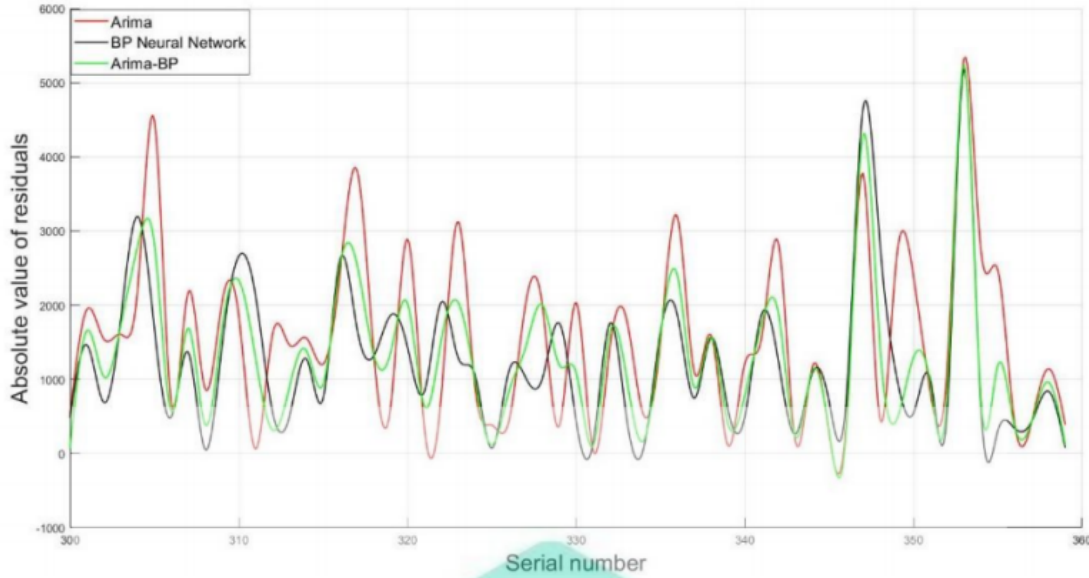


Figure 7: Prediction errors of the three models

cast for the number of results reported on March 1, 2023 using this combined ARIMA-BP model is approximately 19944.

3.3 Interval Prediction Based on Bootstrap Model

Bootstrap Method The Bootstrap method is an important method used in statistics for interval prediction, which first assumes that the data set obeys an unknown distribution, and later estimates the distribution interval of the sample by repeatedly sampling the given data set [7]. We took the data predicted under the weighted combination model for the next 60 periods as the sample set, after which we sampled it 1000 times with put-back to obtain the 1000 Bootstrap sample set, $y_B = (Y, Y_2, \dots, Y_{1000})$.

For each subsample, the distribution is in line with that of the sample, so we calculated the standard deviation (SD) of these 1000 subsample sets as the SD of the sample:

$$\sqrt{\text{Var}(y)} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2}$$

According to the central limit theorem, it is known that the set of samples obtained by conducting 1000 draws approximately obeys the normal distribution. Therefore, the upper and lower limits of the confidence interval of the sample at confidence level $1-\alpha$ be calculated with the help of z-statistic:

$$y = \hat{y} \pm Z_{\alpha/2} \times \sqrt{\text{Var}(y)}. \quad (2)$$

Taking the confidence levels of 95%, 90% and 80% respectively, the final prediction intervals for the number of results reported on March 1, 2023 are shown in Table ??

Table 3: Confidence Intervals

Confidence level	Lower limit	Upper limit
95%	19504.74	20383.26
90%	19575.33	20312.67
80%	19656.69	20231.31

4 Exploring the Impact of Word Attributes on Hard Mode Based on Multiple Linear Regression Model

Exploring the effects of certain attributes of words on the proportion of players choosing Hard Mode is a multiple-input variable, single-output variable modeling problem. We decided to develop a multiple linear regression model to derive whether word attributes affect the percentage of scores reported that were played in Hard Mode and the extent of their effects. The input variables X_1 - X_7 are defined as follows.

4.1 Defining Qualitative Attributes of Words

First we made an overview of all the words in the dataset, and drew a word cloud map for the 356 words after preprocessing in Figure 9. We found that each word is different and its frequency is 1, so its size in the word cloud map is about the same.

4.2 Defining Quantitative Attributes of Words

$$WID_n = \sum_{i=1}^4 |L_{i+1} - L_i|, n = 1, 2, 3, \dots, 357 \quad (3)$$

4.3 Building a Multiple Linear Regression Model

$$X_1 = \begin{cases} 1, noun \\ 2, verb \\ 3, adjective \\ 4, other \end{cases}, X_2 = \begin{cases} 1, common \\ 2, uncommon \end{cases}, X_3 = \begin{cases} 1, vowel \\ 2, consonant \end{cases}$$

The dependent variable (Y) is the percentage of scores reported that were played in Hard Mode, which can be calculated according to equation (??).

$$Y = \frac{\text{Number in hard mode}}{\text{Number of reported results}} \times 100\% \quad (4)$$

Descriptive statistics for all variables identified are shown in Figure 10 and Table 4. As seen in (a), there are more uncommon words in the dataset than common ones, and more words with initial consonant letters than vowel letters, as seen in (b), there are the most nouns in the dataset, accounting for 53%, 61% of words with only one vowel letter

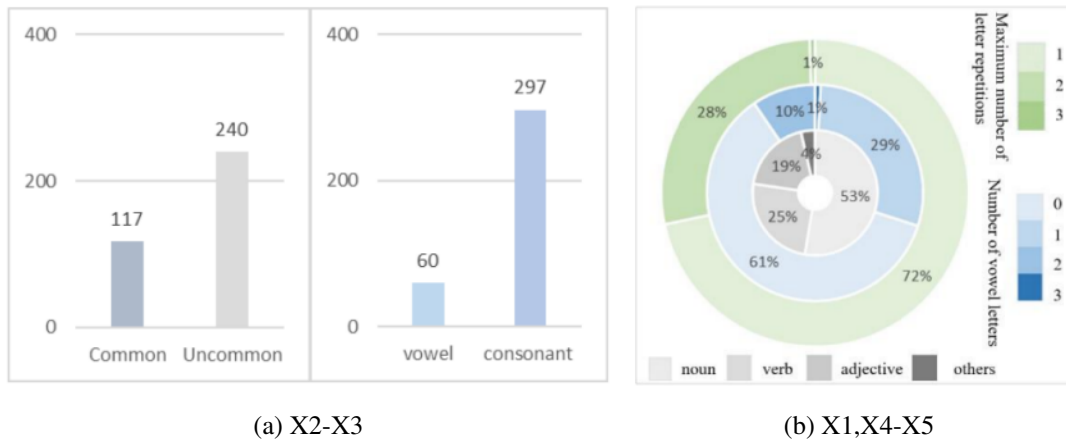


Figure 8: Three images

As seen in Table ??, the average value of emotional polarity value among the 356 words is 0.0355, which is positive, the average value of word internal distance is 33.3726; the average percentage of people reported to play hard mode is 7.52%

Table 4: Confidence Intervals

Confidence level	Lower limit	Upper limit	Average	Standard deviation
X6	-1	0.999	0.0355	0.6448
X7	7	71	33.3726	11.9655
Y	1.17%	13.33%	7.52%	0.0223

5 Conclusion

5.1 Summary of Results

5.2 Strengths

- The sensitivity analysis of the model demonstrates the effectiveness of the model under different parameter combinations and prove the robustness of the mod
- Second one ...

Table 5: Add caption

Letters that appear	Letters that may appear together	C_{conf}	Letters that appear	Letters that may appear together	C_{conf}
a/k/s	e	1	e/n/r	i	1
l/p/t	a	1	p/r/t	e	1
b/e/n	i	1	g/h/t	i	1
c/r/t	e	1	i/m/s	t	1
d/p/t	e	1	d/f	e/t	0.6
h/i/s	e	1	f/t	l/o	0.5

5.3 Weaknesses and Improvements

- The analysis of fish migration can be more accurate if we have more complete data;
- Some approximate analysis methods are applied to model the management of fishing companies, which may lead to a situation contrary to the actual one in extreme cases.

To:
From: Team
Date: Febru
Subject:

To:
From: Team #1887415157 of 2023 MCM
Date: February 31, 2023
Subject:

A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning.

A randomly generated piece of English
that has no real meaning. A randomly
generated piece of English that has no
real meaning. A randomly generated
piece of English that has no real mean-
ing. A randomly generated piece of En-
glish that has no real meaning. A ran-
domly generated piece of English that
has no real meaning. A randomly gen-
erated piece of English that has no real
meaning. A randomly generated piece
of English that has no real meaning.
A randomly generated piece of English

that has no real meaning. A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning.

This part must be a bit of an unordered list or an ordered list or something like that, otherwise it's really all text and will look super ugly.

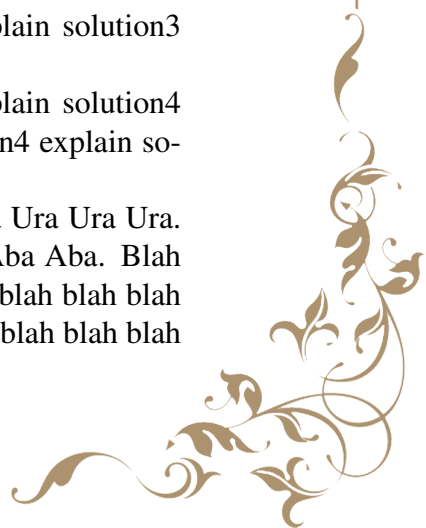
- Solution 1. **Build more shopping centers.** Explain solution1 explain solution1 explain solution1 explain solution1.
- Solution 2. **Build more shopping centers.** Explain solution2 explain solution2 explain solution2 explain solution2.
- Solution 3. **Build more shopping centers.** Explain solution3 explain solution3 explain solution3in solution1 explain solution1 explain solution1.
- Solution 4. **Build more shopping centers.** Explain solution4 explain solution4 explain solution4 explain solution4 explain solution4.

A part summarizes^[2, 2] the words of the table above. Ura Ura Ura Ura Ura Ura Ura. Aba Aba Aba Aba Aba Aba Aba Aba Aba Aba Aba Aba Aba Aba Aba Aba Aba. Blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah blah. Blah blah blah blah blah blah blah blah blah blah blah blah blah. Blah.

Note

1. Text Text Text Text
2. Text Text Text Text Text Text
Text Text Text
3. Text Text Text Text Text Text
Text Text Text Text Text Text
4. Text Text Text Text

1. Text Text Text Text
2. Text Text Text Text Text Text
Text Text Text
3. Text Text Text Text Text Text
Text Text Text Text Text Text
4. Text Text Text Text



Conclusion. This part of writing conclusive things, one or two sentences to summarize it, do not write too much.

References

- [1] Hao H, Wang Y, Xia Y, Zhao J, Shen F. Temporal convolutional attention-based network for sequence modeling. arXiv preprint arXiv:2002.12530. 2020Feb 28.
- [2] Harvey AC. 1990 Forecasting, structural time series models and the kalman filter. Cambridge, UK: Cambridge University Press.
- [3] Krizhevsky A, Sutskever I, Hinton GE. 2012 ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25 (NIPS) (eds F Pereira, CJC Burges, L Bottou, KQ Weinberger), pp. 1097–1105.