Problem Chosen	2024 MCM/ICM Summary Sheet	Team Control Number
C		2309397

Uncover the Hiddnd Secret in the Wordle Results

Summary

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

First, after **preprocessing** the given data and analyzing the time series diagram of thenumber of reported results, we found that the changes can be divided into 3 stages. Toforecast the number of reported results, we developed a weighted optimization modelbased on ARlMA and BP neural network. The prediction interval is then given usingthe Bootstrap method. We packaged this process as **ARIMA-BP Interval PredictionModel Based On Bootstrap**. Thus, we finally predicted the interval prediction valueobtained 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 tobuild a **Multiple Linear Regression Model** with the percentage ofhard mode's playersWe found that the proportion will decrease by an average of0.618 when the initial letterchanges from a vowel to a consonant while it will increase by an average of0.017 foreach one-unit increase in word internal distance.

After that, we made the percentage distribution prediction of the reported results basedon **LSTM Model**. To ensure the percentage is around 100%, we first processed thecomponent data using a **spherical coordinate transformation**. Then we use them asoutput variables, the 7 word attributes and number ofresults as input to train our LSTMmodel. The prediction of EERlE 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 predictionand found that it is around 0.4. For **error analysis**, we use MSE, RMSE and R tomeasure 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 dificultyof words. We built a **GMM Clustering Model** based on these indicators and thusclassifying 5 difficulty levels. We classified the word EERlE as dificulty level lll

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

Finally, we evaluated and refined the model and reported the findings in a letter to thethe 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 epidemicThe New York Times newspaper which is well known for the games it publishes hasbought Wordle on February 2021[?]. Wordle only allows one game to be played perday, and every player in the world plays to guess the same five-letter word in six triesor less each day[?]. Players can play in regular mode or hard mode. They can sharetheir scores via Twitter, thus attracting more people to play and share

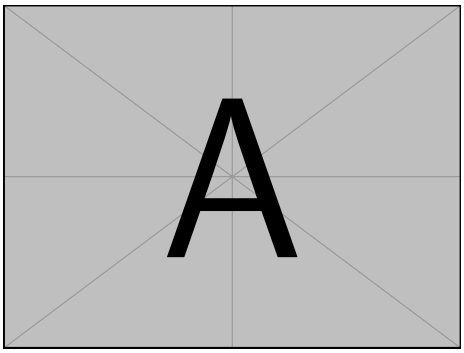


Figure 1: Wordle Game

Back in October 2021,less than 5.000 visits registered to its web page while by January 2022traffic had skyrocketed to over 45 million. Some ofus also love this game, Figure l shows one resulthat we got. The green tile indicates that the secretsolution word has the letter in the precise locationThe yellow tile implies that the answer has theletter but not at the correct location. Grey tilesindicate the letters are not contained in the solutionat all[?].

Now we have a file of daily results from January 7, 2022 to December 31, 2022. Thisfile 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 addressedas follows:

- Develop a model to explain the daily variation in the total number of peoplereporting scores on Twitter and use it to give a prediction interval for the totanumber of people on March 1.2023.
- Determine whether the attributes of words affect the percentage of playerswho choose the hard mode and explain the obtained results accordingly.
- If given a future date and the specific word, build a model to predict thepercentage of 1-X tries in this day. After that, the word EERlE on March 1. 2023should be used as a specific example of the model prediction, while analyzing theuncertainties of the model and the accuracy of the prediction.
- Develop a model to classify the diffculty of the words and identify theattributes of them under each category. Use this model to determine how difficultthe word EERlE is? Finally, discuss the accuracy ofthe classification model
- List and describe some other interesting features of the dataset

1.3 Our work

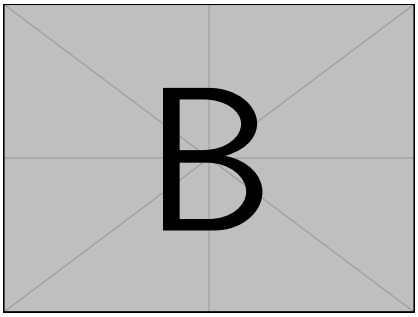


Figure 2: Our Work

2 Assumptions and Notations

Assumption 1: The number of daily online gamers in the data set is a time-dependentset of series and is independent of seasonal changes.
Explanation: Therefore, we can use ARlMA as well as LSTM models to predict a series ofdata for March 1.2023.

Assumption 2: The scores reported by players on Twitter on a daily basis are normaland reliable.
Explanation: To ensure that the model we build based on the dataset can reliably predict arange of data for March 1.2023.

Assumption 3: Assuming that Wordle's development process is consistent with thegame's life cycle theory.
Explanation: This assumption reduces the impact ofexternal uncertainties on Wordle gamepredictions, 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 none-frequent-used symbols will be introduced once they are used.

Table 1: Notations

Symbol	Defination
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 areive outliers, One of them does not exist and two of them has less than 5 letters. wedeleted these three rows of data due to the difficulty ofobtaining the true values ofthesewords. The two remaining outliers are caused by data entry errors. In order to avoid anexcessive reduction in the amount of data, we changed them by combining the previousand later data as well as the semantics ofthem, In addition, the sum of the percentagesin the original data are all in [98%, 102%], which is not much different from 100%, sothey are reasonable and do not need to be processed. In summary, the preprocessing ofthe 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 correspondingnumbers of reported results are not aflected, so the complete reported data are analyzedin this problem. By observing and analyzing the characteristics of the changes in thenumber of reported results from Jan.7, 2022 to Dec 31, 2022, we found that they canbe divided into 3 phases, as shown in Figure 3.
Based on the game lifecycle and player lifecycle theory[4], and combined with Wordle'sgame 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 aday. This artificial scarcity enhances players' desire for challenge and anticipation. Thesharing function of Wordle uses emoji blocks to refer to the results of the game, whichis very recognizable and easy to spread, while also avoiding spoilers, thus attractingmore new players to play Wordle. In addition, the traditional popularity of crosswordpuzzles and the easy-to-play game mechanics have contributed to Wordle's furtheipopularity. During this period, Wordle's overall growth in the number ofresults reportedis 348.85%, with the number peaking at 361.908 on February 2.

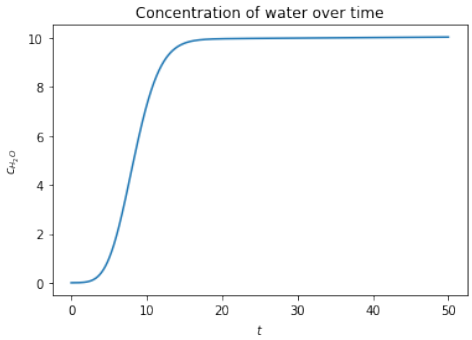


Figure 3: Changes in the number of reported results

- Parse 2: Rapid Decline Period (February 3, 2022 - May 29,2022) During this period, Wordle experiences an overall 84.29% decline in the number ofreported results, which is the general nature of internet fads. When the popularity ofthegame reaches its peak, it faces a significant loss of players due to gamers' declininginterest and boredom with the game. The singularity of Wordle's game mechanic cancontribute to this. Furthermore, the emergence of competing games in the market suchas "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 itspopularity among the general public and they will quickly leave when Wordle is nolonger hot or the next trend emerges. The remaining players are often loyal Wordleplayers or players who have developed user stickiness due to the game's social circleAt 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 oftime series analysisThe traditional ARlMA model is capable of extracting deterministic information fromhistorical data in order to predict the trend of variables over time, however, this moderequires extensive testing before application, and the process of determining the orderofthe model is subjective to some extent [5]. In recent years, with the development ofartificial intelligence, machine learning algorithms have been widely applied to the fieldof data classification and prediction. Among which, BP neural network model which iseficient and convenient can effectively break the limits of traditional time seriesprediction models. Considering these factors, we decide to combine the advantages oftraditional time series models and machine learning models. It means that first of all.we use ARlMA model and BP neural network model to predict the number of reportedoutcomes separately, and then we use weighted average to combine the two models toobtain a more reliable ARlMA-BP combined prediction model.

3.2.3 ARIMA Forecasting

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

Figure 4: ARIMA model building process

First, we performed ADF smoothness test on the variable "Number ofreported resultsin the preprocessed data set and found that it wasn't smooth. Therefore, we differencedthe series to the first order and tested the diferenced series again. After that, weobserved the ACF and PACF plot of the differenced series, and found that the ACF plotshowed the characteristics of trailing tails and the PACF plot showed the characteristicsof lst order truncated tails. Therefore, the following ARIMA(1.1.0) model can bedeveloped for the original series:

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

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

3.2.4 BP Neural Network Prediction

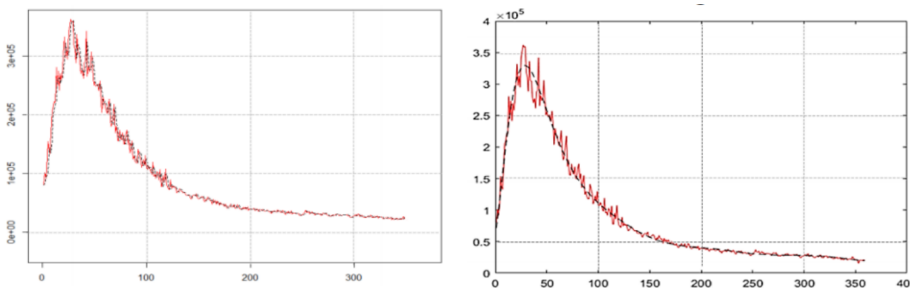
BP neural network is a multilayer feedforward neural network trained according to theerror back propagation algorithm, which can iterate and repair its own weightscontinuously based on the relationship between input and output variables so as tofinally estimate the exact functional relationship[?]. Its algorithm is explained by thefollowing 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 ARlMA(1,1,0) model prediction is shown in Figure **??**.and using the BP neural network model prediction is shown in Figure **??**:



(a) ARIMA prediction effect (b) BP neural network prediction effect

Figure 5: Three images

It can be seen that the ARlMA model has a better effect prediction, and the pointprediction Pa 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 buthas relatively good prediction effect in the later stage. The point prediction value Peobtained 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 anew combined ARlMA-BP prediction model based on ARlMA and BP neural networkmodel by using the prediction errors to weight the two models. The design idea of thisweighted combined forecasting model is shown in Figure **??**: We first calculate the average prediction error for each of the two prediction models:

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

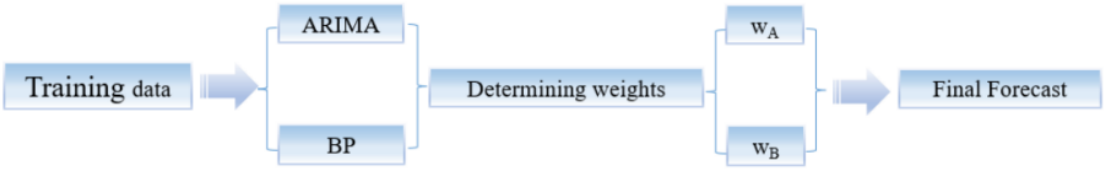


Figure 6: Combined model design

After that, we use the share of the average prediction error squared of either model inthe total average prediction error squared of the two models to reflect the weightaccounted 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^thseries in the final ARIMA-BP combination model obtained is calculated as:

$$P_i = w_A P_{Ai} + w_B P_{Bi} \tag{1}$$

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

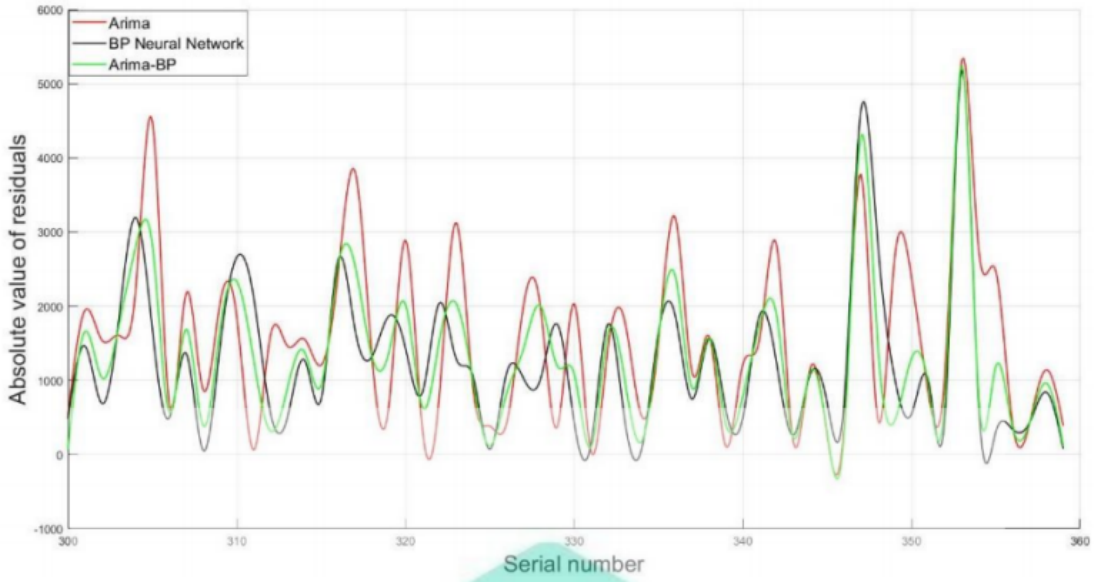


Figure 7: Prediction errors of the three models

cast for the number of results reported on March 1, 2023 using thiscombined 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 estimatesthe distribution interval of the sample by repeatedly sampling the given data set [7].Wetook the data predicted under the weighted combination model for the next 60 periodsas the sample set, after which we sampled it 1000 times with put-back to obtain the1000 Bootstrap sample set, y_B=(Y,Y_2,••••••, $Y_1$000).
For each subsample, the distribution is in line with that ofthe sample, so we calculatedthe standard deviation (SD) ofthese 1000 subsample sets as the SD ofthe sample:

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

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

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

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 playerschoosing Hard Mode is a multiple-input variable, single-output variable modelingproblem. We decided to develop a multiple linear regression model to derive whetherword attributes affect the percentage of scores reported that were played in Hard Modeand the extent oftheir effects. The input variables X1-X7 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 mapfor the 356 words after preprocessing in Figure 9. We found that each word is differentand its frequency is l, so its size in the word cloud map is about the same.

4.2 Defining Quatitative Attributes of Words

$$WID_n = \sum_{i=1}^{4} |L_{i+1} - L_i|, n = 1,2,3,\cdots,357 \tag{3}$$

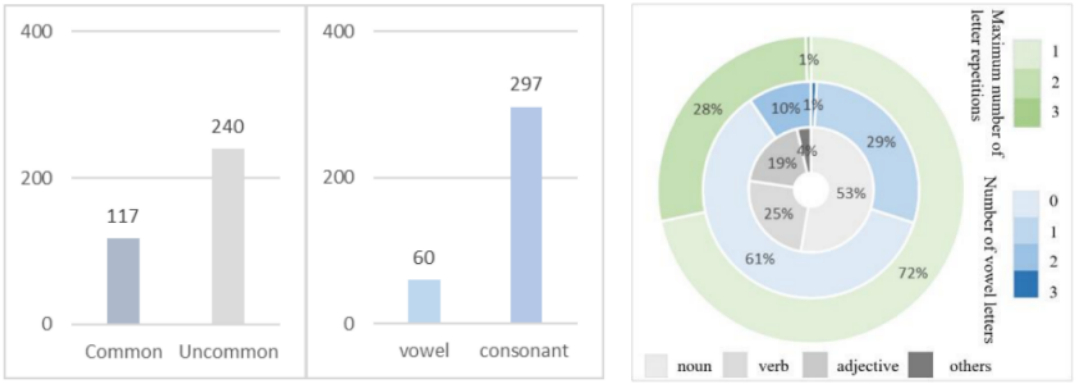
4.3 Building a Multiple Linear Regression Model

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

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

$$Y = \frac{Numberinhardmode}{Numberofreportedresults} \times 100\% \tag{4}$$

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



(a) X2-X3 (b) X1,X4-X5

Figure 8: Three images

As seen in Table **??**, the average value ofemotional polarity value among the 356 wordsis 0.0355, which is positive, the average value of word internal distance is 33.3726; theaverage 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 ...

Figure 9: Letters with the top 8 frequencies in 5 letter positions

First Letter	Frequency	Second Letter	Frequency	Third Letter	Frequency	Forth Letter	Frequency	Fifth Letter	Frequency
s	51	o	51	o	51	e	51	e	73
c	32	a	47	i	44	a	44	y	49
t	29	l	41	a	42	r	42	t	46
a	28	r	36	e	29	t	29	r	35
p	23	h	33	u	28	n	28	l	24
f	22	e	28	n	23	l	23	d	20
b	20	i	23	r	19	i	19	h	18
m	19	n	17	p	14	e	14	k	16

Table 5: Word combination pattern

Letters that appear	Letters that may appear together	Conf	Letters that appear	Letters that may appear together	Conf
a/k/s	e	1	e/n/r	i	1
l/p/t	appear together	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.

Memorandum

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

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

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 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 explain solution4 explain solution4.

A part summarizes[?, ?] the words of the table above. Ura 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 blah blah blah blah.

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.