UDC: 811.322:

COMPUTATIONAL METHODS USING CHARACTER STATISTICS FOR THE WORD GAME

Ulugbek Salaev,
ulugbek0302@gmail.com
Urgench State University, Ph.D. student

Abstract. Wordle is a word game that intends six attempts to guess a five-letter word. This study uses character statistics of Uzbek five-letter words in Cyrillic script to determine the solving strategy of the word game. The proposed methodology covered computing the letter frequency (LF) and positional letter frequency (PLF) of Uzbek words in a Cyrillic script and calculated a word score based on the PLF to suggest the best probability words to obtain a solution in the game with minimal attempts.

Keywords: *Wordle, Character statistics, word-game, Uzbek words*

Introduction. The web-based word game Wordle [1] was implemented in November of 2021. This game is adopted into many languages of the world since it is open-source software. Among them, there is also the Uzbek language in both Latin and Cyrillic scripts. It gives players six attempts to try and guess a five-letter secret answer word. At each attempt, hints are offered to the player so that they may make subsequently “better” guesses by marking each letter with one of three possible colors, each has a specific meaning:
1) Green: The entered letter is in the expected solution and is in the expected position.
2) Yellow: The entered letter is in the expected solution but it is not in the expected position.
3) Gray: The entered letter is not in the expected solution.
Figure 3 shows examples of Wordle solutions with the tile colors providing hints to the player to progress. Also, there is no indication as to how many times the character would appear.

The work by Sidhu [3] attempted to find the best starting word from a linguistic perspective while the work by Anderson and Meyer [2] have used machine learning to find the optimal human strategy for solving the Wordle. The objective of this work is to derive the set of 3 optimum starting words for the English version of the game covering 15 different characters and ordered in the descending order of significance. Fully automatic solving software of Wordle has also been created for

the English language.

The proposed methodology in this article is targeted at the Cyrillic script version of the Uzbek language for the Wordle game. This version can be accessed via the webpage http://wordlar.uz/.

Methodology. The word database contains five-letter 5923 total words (W) in Cyrillic script that was manually prepared from the Uzbek dictionary [4] that has over 85k Uzbek root words. The majority portion of the dataset included root morphemes, because of the agglutinative character of the Uzbek language as many inflectional words can be formed by adding several suffixes to most root morphemes and it causes mismeasurements while calculating LF and PLF. From W, each of 3954 words has unique letters, which made up 67% of the total number of words. Table I shows the character occurrence count and its frequency, and the count of words that the character exists.

Table I: Calculated character frequencies

Letter	cnt.	freq.	*cnt.	Letter	cnt.	freq.	*cnt.	Letter	cnt.	freq.	*cnt.
а	4188	14.14	3439	м	1213	4.10	1167	ш	714	2.41	676
б	1050	3.55	997	н	1429	4.83	1365	ъ	31	0.10	31
в	416	1.40	411	о	1826	6.17	1736	ь	7	0.02	7
г	626	2.11	615	п	300	1.01	291	э	114	0.38	114
д	1070	3.61	1010	р	1400	4.73	1358	ю	178	0.60	176
е	541	1.83	535	с	1142	3.86	1089	я	230	0.78	229
ж	237	0.80	232	т	1226	4.14	1166	ё	308	1.04	308
з	680	2.30	665	у	1022	3.45	926	ў	650	2.19	649
и	3639	12.29	3133	ф	213	0.72	210	ғ	318	1.07	316
й	647	2.18	632	х	289	0.98	287	қ	929	3.14	853
к	829	2.80	763	ц	7	0.02	7	ҳ	352	1.19	350
л	1290	4.36	1229	ч	504	1.70	488				

**cnt – Word counts of the character exist*

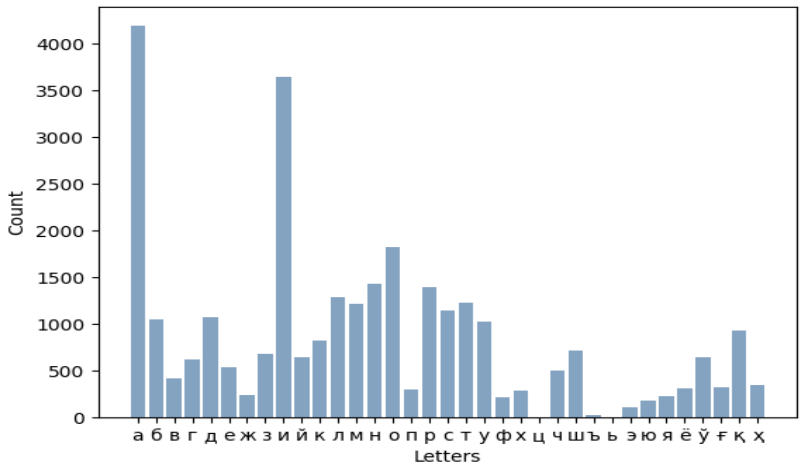

Figure I. Character frequencies

It can be seen from Table I that some letters are frequently used while some of them are very rare. Letters “а”, “и”, “о”, “н” and “р”, “л” occur 58%, 53%, 29%, 23%, 21% in words of *W* respectively, in which the 3 letters with the highest frequency correspond to the vowel letters. In the next step, we calculated the character distribution on each position (Table II) and kept it as a char frequency map.

Table II. Character count by positional

	#1	#2	#3	#4	#5		#1	#2	#3	#4	#5		#1	#2	#3	#4	#5
а	592	1068	237	1081	1210	м	164	81	373	226	369	ш	79	127	190	71	247
б	553	36	120	32	309	н	93	164	282	431	459	ъ		13	13	4	1
в	49	87	184	36	60	о	313	699	122	617	75	ь		1		2	4
г	74	18	198	267	69	п	115	62	101	12	10	э	111		2		1
д	197	44	330	427	72	р	86	247	505	150	412	ю	138	21	7	5	7
е	70	349	25	93	4	с	368	148	310	194	122	я	92	34	52	11	41
ж	108	25	69	16	19	т	440	178	329	119	160	ё	210	23	47	21	7
з	103	141	284	51	101	у	171	568	43	214	26	ў	235	408		6	1
и	256	594	351	1278	1160	ф	50	40	75	24	24	ғ	63	50	142	50	13
й	65	140	210	13	219	х	131	37	94	6	21	қ	359	107	204	53	206
к	284	74	181	81	209	ц			5		2	ҳ	133	42	131	9	37
л	90	184	583	238	195	ч	131	113	124	85	51						

We calculated the score for words of *W* based on Table II, this score indicates possibilities of words that can be offered as a solution. In the first position, the letter ‘a’ has the highest probability in the letter frequency map, indicating that ‘a’ is the most common starting letter for 5-letter words. Similarly, the 5th position shows that ‘a’ is the most common ending letter. This distribution was calculated only on words found in the 5923 words database. Figure II illustrates the most common five letter frequencies for the position.

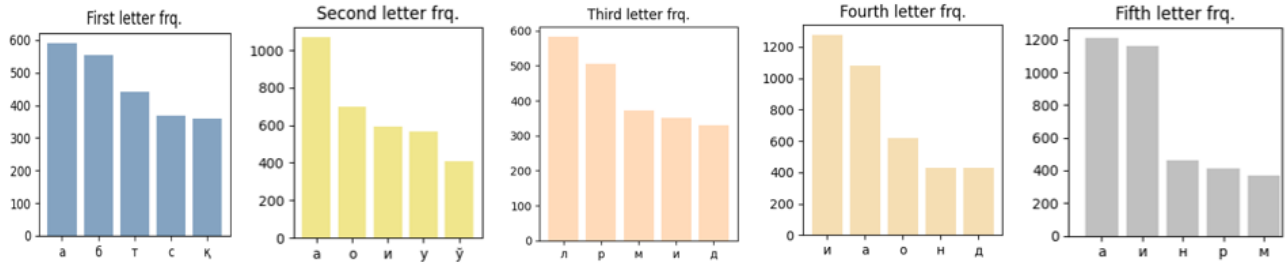

Figure II. Most common five letter frequencies for the position

To make a conclusion based on the frequency of letters in the Uzbek language, the highest frequency letters correspond to vowels. The best first guess has five different letters and it should include some of these three vowels. For this reason, it is suggested to start with vowels in the initial attempt when guessing a word. Using words with high frequency and not overlapping letters lead to a good process. We chose to focus only on the possible winning answers for this analysis because our goal is to find the best guesses for human players. Since machines can make perfect use of information from grays and yellows in choosing their next move, this method aims to optimize a Wordle strategy for a human player. Whereas with a green letter, a human can better construct their next move with the certainty that a green provides.

A detailed algorithmic description of the proposed methodology:

Data collection:

Collected Uzbek (Cyrillic script) 5-letter words from the dictionary and manually added other words
Load data:
$\boldsymbol{W} \leftarrow \{\boldsymbol{word_1}, \dots, \boldsymbol{word_n}\}$, $|\boldsymbol{W}| = \mathbf{5923}$, # List of five-letter words

$\boldsymbol{C} \leftarrow$ {"а", "б", "в", "г", … , "ҳ"}, $|\boldsymbol{A}| = \mathbf{35}$ # Alphabet, List of character

Calculate character frequencies:

$\forall \boldsymbol{w}\left[\forall \boldsymbol{c}\left[\boldsymbol{F}(\boldsymbol{c}) = \boldsymbol{F}(\boldsymbol{c}) + \sum_{\boldsymbol{l}\in\boldsymbol{w}}\begin{cases}\mathbf{1} & \textbf{if } \boldsymbol{c} = \boldsymbol{l} \\ \mathbf{0} & \boldsymbol{else}\end{cases}\right]\right]$, # l is a character in the word $\boldsymbol{w}$, F is character frequency map of the format <character, value>

$\underset{\boldsymbol{w}\in\boldsymbol{W}}{\forall}\left[\underset{\boldsymbol{p}\in\overline{\mathbf{1,5}}}{\forall}\left[\underset{\boldsymbol{l}\in\boldsymbol{w}}{\forall}[\boldsymbol{P}(\boldsymbol{p},\boldsymbol{l}) = \boldsymbol{P}(\boldsymbol{p},\boldsymbol{l}) + \mathbf{1}]\right]\right]$, $\boldsymbol{P}$ is a count of the letter in the position of the format<character, position, value>, $\boldsymbol{l}$ is a character in the word $\boldsymbol{w}$, $\boldsymbol{p}$ is a position of $\boldsymbol{l}$ in the word $\boldsymbol{w}$;

$\underset{\boldsymbol{w}\in\boldsymbol{W}}{\forall}\left[\underset{\boldsymbol{p}\in\overline{\mathbf{1,5}}}{\forall}\left[\underset{\boldsymbol{c}\in\boldsymbol{C}}{\forall}\left[\boldsymbol{W} = \sum_{\boldsymbol{l}\in\boldsymbol{w}}\begin{cases}\boldsymbol{P}(\boldsymbol{p},\boldsymbol{l}) & \textbf{if } \boldsymbol{l} = \boldsymbol{c} \\ \mathbf{0} & \boldsymbol{else}\end{cases}\right]\right]\right]$, update $\boldsymbol{W}$, word list map of the format <word, value> by adding the word $\boldsymbol{w}$ score, $\boldsymbol{l}$ is a character in the word $\boldsymbol{w}$;

Sorted $\boldsymbol{W}$ by a score in descending order;

$\acute{W} = \bigcup_{w \in W} \begin{cases} \{w\} & \textbf{if } len\left(\bigcup_{c \in w} \{c\}\right) = 5 \\ \emptyset & else \end{cases}$, the function makes $\acute{W}$ only contains words that have five unique characters by removing the words which have repeated characters;

Output data:
Output top k *highest scored words in a format <word, value>;*
Input parameters: non_exist_letters, exist_letters, positional_letters;
Filter $\acute{W}$ *by conditions given by input parameters;*
Output top k *highest scored words that matched the conditions;*
Go to input parameters and continue;
Exit when a user terminates;

Experiment and Result. When we proceeded with the algorithm, best probabilities words by computed positional letter occurrences were extracted for suggestions to use in an initial attempt. At this point, we came across a problem, in which some of the words that were suggested as options were invalid words by Wordle as they are not included in its dataset. To input, we will choose one of the highest-ranking suggestions until it is accepted in Wordle. However, it is avoided to output letter overlapping words to have to cover more characters. The highest-ran 2 words set accepted by wordle was {’қалин’, ’балиқ’}. Logically, these are the best words to use in the first attempt. Figure III shows an example of using one of the suggested words at the beginning of the game and the process of continuation based on the methodology. Additionally, any words from *W* may not be acceptable in the Wordle game.

Suggested words for initial attempt:
{'word': 'қалин', 'score': 3747}
{'word': 'балиқ', 'score': 3688}
{'word': 'салим', 'score': 3666}
{'word': 'бадир', 'score': 3641}
{'word': 'бақир', 'score': 3515}

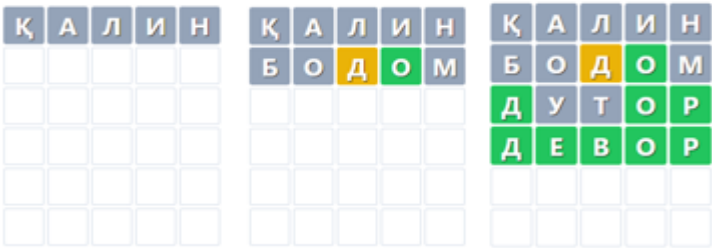

Figure III. Example of Wordle Solution

Left 220/5923 words
{'word': 'бодом', 'score': 2568}
{'word': 'бозор', 'score': 2565}
{'word': 'тумор', 'score': 2410}

Left 8/5923 words
{'word': 'дутор', 'score': 2123}
{'word': 'дучор', 'score': 1918}
{'word': 'девор', 'score': 1759}

{'word': 'бемор', 'score': 2304}
{'word': 'бедор', 'score': 2261}
{'word': 'дугоҳ', 'score': 1617}
{'word': 'ёдгор', 'score': 1481}

After providing the word on the third attempt, we got the following information: the solution does not have 9 characters [қ,а,л,и,н,б,м,у,т] and characters [д,о,р] are located on his exact position. And when we provide the above information to the algorithm, it presented only 1 possible option {'word': 'девор', 'score': 1759}. The solution was the same for this last suggestion word.

Conclusion. This article explored the letter frequency of Uzbek words in a Cyrillic script and calculate word scores based on the positional letter frequency among words. This methodology is aimed at offering the most matching words to win in the game with minimal attempts. According to the proposed algorithm the software was created. The complete list of words and the software is available at the address: https://github.com/UlugbekSalaev/wordle. It is a standalone program that is created in Python language and converted to an executable file.

References

[1] J. Wardle, “Wordle - a daily word game,” https://www.powerlanguage.co.uk/wordle/, accessed: 2022-02-04.
[2] B. J. Anderson and J. G. Meyer, “Finding the optimal human strategy for wordle using maximum correct letter probabilities and reinforcement learning,” arXiv preprint arXiv:2202.00557, 2022.
[3] D. Sidhu, “Wordle – the best word to start the game, according to a language researcher,” https://bit.ly/3qItHsI, 2022, [Online; accessed 07-February-2022].
[4] А.Мадвалиев, Э.Бегматов, “Ўзбек тилининг имло луғати”, Тошкент: Akademnashr, 2013. 520 б.
[5] M. Butterfield, “Science has determined the worst Wordle starting word,” https://bit.ly/3J9FOWj, 2022, [Online; accessed 07-February-2022].
[6] The Best Wordle Starting Word Has Been Figured Out With (Computer) Science. In: GameSpot [Internet]. [cited 28 Jan 2022]. https://bit.ly/35k7EB9
[7] Groux C. The 20 best Wordle starting words, according to science. In: Inverse [Internet]. [cited 28 Jan 2022]. https://bit.ly/3ISDcvn