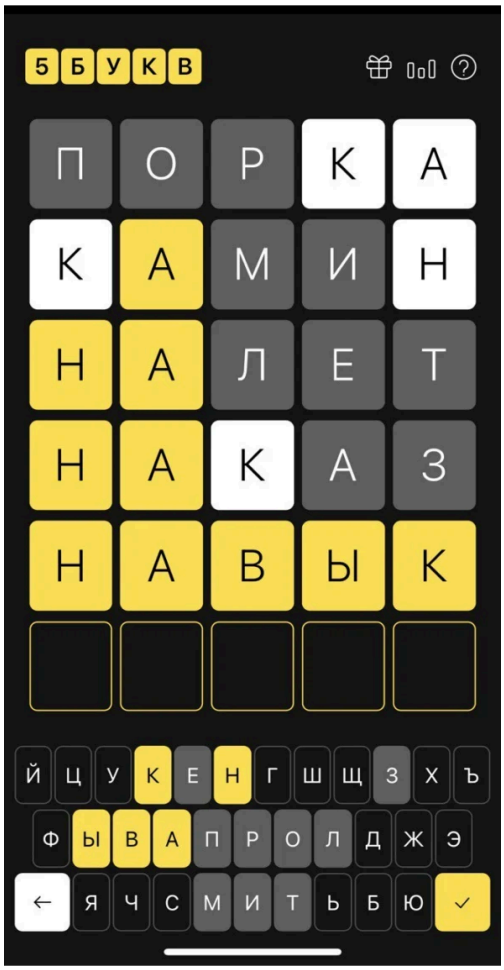
Игорь Белик
Офтоп 02.11.2022

# Как выигрывать в игру "5 букв" в приложении Тинькофф и получать повышенный cashback и при чем здесь порка

В Тинькофф сейчас идет 2й сезон игры "5 букв", победители игры получают бонусы (например, повышенный cashback). Это статья про то, как используя аналитику, я смог выработать оптимальную стратегию победы и получить бонусы.

## Суть игры “5 букв”.

Игра "5 букв" - аналог игры Wordle, где нужно угадать слово из 5 букв за 6 попыток. Каждая попытка должна быть существительным в единственном числе. Если буквы в этом слове нет, то она окрашивается серым, если есть, но на другом месте, то белым, если есть и на правильном месте, то желтым.




Один раунд в 5 букв

## Очевидная стратегия, которую подсказывает игра.

Нужно получать максимум информации из каждой попытки. Информация - это угаданная буква, угаданное место буквы, отсутствие той или иной буквы тоже

информация. Так постепенно можно узнать о слове все больше и зная 3-4 буквы загаданного слова можно угадать слово целиком.

В русском языке частотность букв (доля появления конкретной буквы к суммарному количеству букв в тексте) сильно различается, поэтому важно в первую очередь использовать буквы, которые с большей вероятностью будут выпадать в игре. Вики, как всегда, знает все и про частотность букв тоже. Статистика собрана на основе данных национального корпуса русского языка https://ru.wikipedia.org/wiki/Частотность#Частотность_букв_русского_языка Самые распространенные 5 букв русского языка — это О, Е, А, И, Н. К сожалению, из них не составить слово, но если добавить чуть менее частотные буквы, то обнаруживается, что самое лучшее слово для первого хода в этой логике – ОКЕАН.

Однако проблема такого начального слова в том, что гласных в русском языке всего 9 (в игре е и ё одна буква), а согласных 23, поэтому перебрав 3 гласных в первом слове, сложно будет сочинить второе без повтора выбывших букв, если таковых не оказалось. Поэтому лучше сразу придумать пару слов на первые 2 хода с максимально частотными и неповторяющимися буквами, например СОНЕТ-МИРАЖ, ТЕНОР-ВАЛИК, суммарная частотность 10 букв в этих словах превышает 60%, поэтому по матожиданию эти пары слов будут содержать 3 из 5 букв загаданного слова.

Дальнейшая часть стратегии тезисно такова:

• Перебирать слова идя от слов с более частотными буквами к менее частотным

• Если какой-то буквы в слове нет, постараться не использовать ее в дальнейших попытках, нужно пробовать другие буквы

• Если буква есть в слове, но не на своем месте, то в следующей попытке переставить ее в другое место. За 4 попытки можно перебрать все варианты

• Если буква есть и ее место угадано, но нет хорошей догадки о слове целиком, то лучше в следующей попытке ее исключить. Угаданная буква уже не принесет дополнительной информации, а вот на ее место можно поставить еще не использованную букву и получить информацию о ней.

• Есть проблема с повторяющимися буквами в слове. Каждая из них будет желтой на своем месте, поэтому после угадывания одной из них, вторая не будет искаться. Практика показывает, что проще найти остальные 2-3 буквы и по ним отгадать слово, чем тратить попытки на поиски дублей найденных букв.

> В принципе хорошая стратегия, позволять угадать за 4-5 попыток, но я решил пойти дальше.

## Более глубокая аналитика.

Мы отгадываем 5-буквенные существительные и частотность букв в них тоже может иметь свои особенности. Для того, чтобы узнать какие, я загуглил датасеты русских слов, нашел на гитхабе вот такой https://github.com/dkulagin/kartaslov/tree/master/dataset/kartaslovsent

Скачал, обработал в питоне, выделил 5-буквенные существительные, выгрузил в эксель и посчитал частотность как букв в 5-буквенных словах в целом, так и их частотность на определенных местах.

Получились следующие наблюдения:

• В 5-буквенных словах самые распространенные буквы – А и О из-за существительных в женском и среднем роде, причем А встречается 5й буквой в каждом 4м слове. Зато Е и И менее популярны некоторых согласных

• В 2 из 3 слов на 2м месте стоит гласная. Если слово не заканчивается на А, то скорее всего 4я буква будет гласная

• Буква К встречается значительно чаще, а вот Н наоборот реже, чем в целом в языке

• Есть буквы, которые встречаются на определенном месте: С или П скорее всего будут на 1м месте, У или Ы на 2м, Ь или Я на последнем

В связи с этим хочется пересмотреть выбор слов для начала игры. ОКЕАН уже не кажется таким хорошим словом, скорее нужно выбрать СОТКА или ПОРКА. Но в дальнейшем стратегия остается прежней.

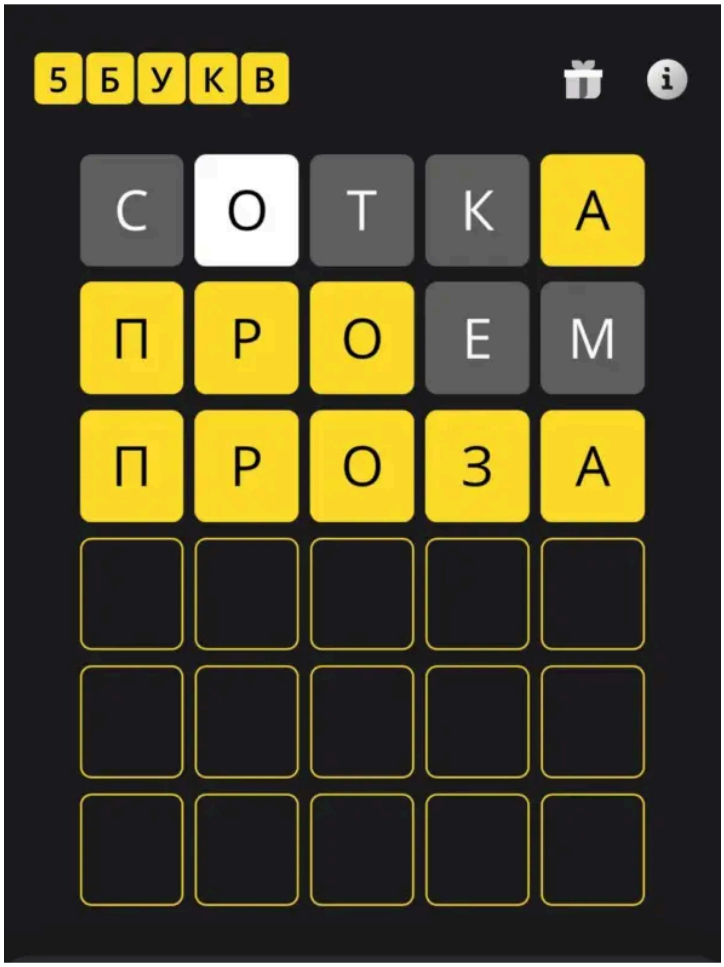
ИТОГО!

> Нужно начинать со слов СОТКА или ПОРКА, так как они содержат самые популярные сочетания букв, что позволит с первого хода собрать больше информации.
>
> В последующих попытках можно ориентироваться на таблицу, где представлены вероятности:

| Буква | 1я | 2я | 3я | 4я | 5я | итого в 5Б словах | в целом в языке |
|---|---|---|---|---|---|---|---|
| а | 3% | 19% | 5% | 9% | 24% | 12% | 8% |
| о | 6% | 15% | 4% | 14% | 5% | 9% | 11% |
| к | 8% | 3% | 3% | 10% | 10% | 7% | 3% |
| е | 0% | 11% | 6% | 11% | 3% | 6% | 8% |
| р | 5% | 7% | 9% | 4% | 6% | 6% | 5% |
| т | 5% | 5% | 6% | 4% | 7% | 5% | 6% |
| и | 2% | 8% | 4% | 9% | 3% | 5% | 7% |
| л | 4% | 4% | 8% | 4% | 3% | 5% | 4% |
| с | 10% | 2% | 5% | 3% | 4% | 5% | 5% |
| н | 4% | 1% | 5% | 6% | 6% | 4% | 7% |
| у | 2% | 9% | 2% | 3% | 1% | 3% | 3% |
| п | 9% | 2% | 4% | 1% | 1% | 3% | 3% |
| в | 7% | 1% | 4% | 2% | 1% | 3% | 5% |
| д | 4% | 1% | 4% | 2% | 3% | 3% | 3% |

Анализ частотности букв на различных позициях, в целом в 5-буквенных словах и в русском языке в целом

Попробовал эту стратегию сегодня и угадал слово за 3 попытки, чего и вам желаю. А вы пробовали в нее поиграть? Поделитесь своими результатами.




❤️ 10