Analyzing Reddit Data

By: Ryan Hathcock, Michael Kintanar, Danny Xie, Daniel Kang, Klayton Trinh

What is Reddit?

- #5 most visited site in the US
- American social news aggregation, web content rating, and discussion website
- Founded on June 23, 2005
- Subreddits and topics for almost anything

Data Set

- Size: 29.62GB
- Date: May 2015
- URL: https://www.kaggle.com/reddit/reddit-comments-may-2015
- Originally 1+ terabyte of data for 1.7 billion publicly available comments

Data Set cont.

- 1 Table
- 22 Columns
- 54,504,410 Total Rows
- Reddit comments for the month of May 15
- Fields include: create_utc, subreddit id, author, score, body, ups, downs, etc

Data Set cont.

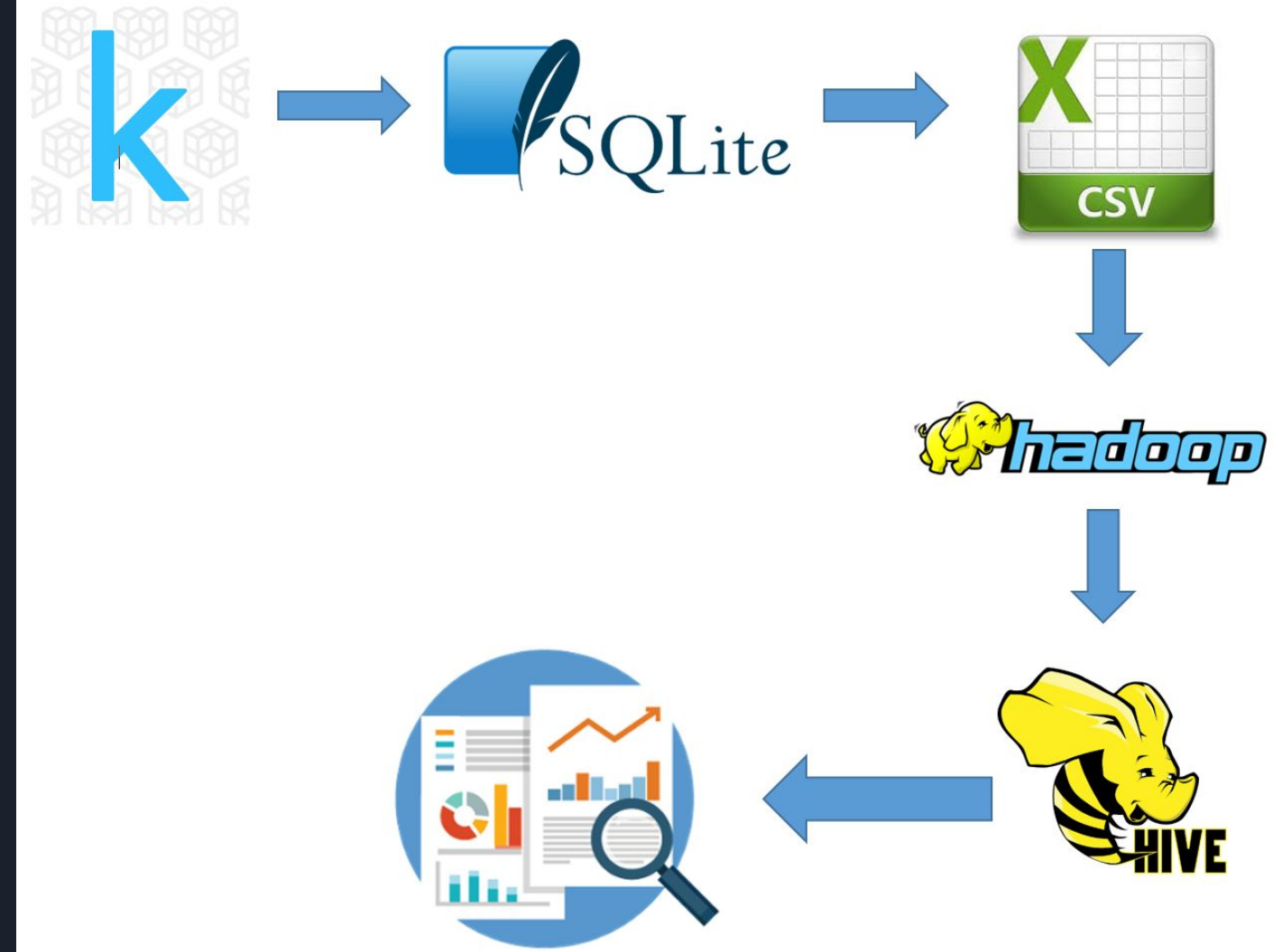
AFTER TRIMMING DATASET

ROWS: 10,225,401

COLUMNS: 11

Columns: id, subreddit_id, subreddit, name, author, ups, downs, score, gilted, distinguished, body

Size of .csv file put into hdfs: 2.27GB


SQLite
CSV
hadoop
HIVE

Overview: Getting Data to Hive for Analysis

- Download sqlite file from kaggle
- Dataset is too large for class servers so must be managed before transfer
- Use sqlite editor to run queries:
    - Drop excessive columns
    - Remove rows containing non default subreddit
    - Remove rows containing [deleted] in body
- Export to csv file
- Transfer to class server local fs with scp
- Transfer to hdfs
- Create hive table

Overview: Overcoming Issues from .csv Export

- The body column contains large strings input by various reddit users so it is problematic for .csv file
- Only one character could be selected as field delimiter and it is highly likely that the delimiter will be encountered in a few user comments
- Selected “|” as delimiter
- The sqlite file before export contained 10,225,401 rows
- The hive table sourced from sqlite exported .csv file contained 19,101,655 rows
- Unable to perform hive functions on table and produce meaningful information
- Quick solution was to create a new hive table using only rows containing any of the 46 subreddit values
- The new table when queried contains 10,225,401 rows
- Able to perform hive functions on new table

Objective

Perform Sentiment Analysis:

- Dataset contains user comments and their score which indicates overall feedback
- Use ngrams function to see the most frequently used phrases
- Use general queries to see top rated comments
- See if frequently used phrases are also top rated
    - Memes and references are often highly upvoted

Overcoming Bot Abundance on Reddit

- Performing ngrams revealed that moderator bots comprise a large portion of comment volume
- Their comments dominate ngram>4 results due to high frequency of their canned messages
- Need to create another hive table which excludes said moderator bots
- Such users have the value “moderator” in the distinguished column
- Easy to sort out
- New table contains 10,134,639 rows

Results: 4 recurring phrases from 5gram

There are 4 noticeable recurring phrases from the 5gram function in hive:

- “i am a beautiful person”
- “i am so i am”
- “the best way to get”
- “a good time to time”

These still are not complete phrases so it’s time to see their origin

“i am a beautiful person”

Initially it seems like a body positivity message or something similar.

```
select subreddit, author, distinguished, score, body from may_2015_b
where lower(body) like '%i am a beautiful person%'
order by score desc limit 30;
```

Result: The phrase seems to preface repetitive, absurd, no context statements. Many comments containing this are highly scored. Note that this phrase appears almost exclusively in AskReddit which is commonly used for humorous and non serious comments.

“i am a beautiful person”

subreddit	author	distinguished	score	body
AskReddit	Ms106		2832	"> I am a beautiful person. I have to unlock phone. The sex the best way for me to buy staff.
AskReddit	qualityproduct		2731	I am not a filthy presser. I had not been able to get the best of luck. I am a beautiful person. I am a beautiful person. I am a beautiful person. I am a beautiful person. I am a beautiful person. I am a beautiful person. I am a beautiful person.
AskReddit	Hithenameisbj		1831	I am a beautiful person. I have to go to the bathroom. On the other hand, I just want to get a chance to win the game.
AskReddit	AxelTV		1369	I am a beautiful person who is this the latest version of the children of the children of the children of the children of the children of the children of the children of the children of the children of the children of the children of the children of the children of the children of the children of the children of the children of the children.
AskReddit	Spider-J		945	"SwiftKey seems to really like the phrase ""I am a beautiful person""... Easter egg? "
AskReddit	xTRS		903	"Swiftkey really likes the phrase ""I am a beautiful person"""
AskReddit	erokk88		892	I am a beautiful person. I am absolutely in shock to be a trashy mom. I am absolutely in shock. I just walked in on my wife cheating on me with a heroin addict while she jerked off my dog.
AskReddit	3shirts		751	"Fun fact: ""I am a beautiful person"" is built in as a sort of default prediction so if you've not used it much, you'll always get that.
AskReddit	Aim4thebullseye		261	"I am a beautiful person. I love you very much. I love you very much. I love you very much...etc
AskReddit	Aishiteruu		166	"""I am a beautiful person who is the best of luck to you"""
AskReddit	YetAnotherAaron		66	"It's by design. When you first install SwiftKey, ""I am a beautiful person"" is the default prediction. [Source](http://www.androidcentral.com/swiftkey-gets-xkcd-treatement)"
AskReddit	iamck94		50	"Mine also says ""I am a beautiful person"""
AskReddit	BCNC3		43	I am a beautiful person who is truly a few days ago, and I will be a normal thing to do with the justice system.
AskReddit	themightymartin		25	I am a beautiful person to person who is the type of service you are looking for.
AskReddit	averysmallbear2		25	I am a beautiful person. I don't know what to say to you.
AskReddit	HoopyWilliams		20	"I had ""I am a beautiful person"" pop up as well.."
AskReddit	Ross_beezy		14	I am a beautiful person. I am a beautiful person. I am a beautiful person.
AskReddit	drphungky		9	I am a beautiful person. I don't know if you need to talk to you. I have a newer version of the house at the end of the world. We move on to the survey. I don't know if you need to talk to you. I have a newer version of the house at the end of the world. We move on to the survey. I don't know if you need to talk to you. I HAVE A NEWER VERSION OF THE HOUSE AT THE END OF THE WORLD.
AskReddit	King_Yeshua		7	I am a beautiful person. I am a beautiful person.
AskReddit	Aceconklin		5	"I am a beautiful person who is the best though I have to ask for a warrant that when I was young and old and new York City new York City new York City new York City new York City new York City new York City new York City new York City new York City new York City new York City
AskReddit	spenpeck		5	I am a beautiful person. I have a quiz. I have a quiz. I have a quiz. I have a quiz. I have a quiz. I have a quiz. I have a quiz. I have a quiz. I have a quiz. I have a quiz. I have a quiz.
AskReddit	krankie		4	I am a beautiful person. I base my racism. I have a turbo tax discount that expires in a week.
AskReddit	compromised_username		4	"I am a beautiful person who is the girl??
AskReddit	Slamma009		4	I am a beautiful person. I don't know what to do now. I don't know what to do now. I don't know what to do now. I don't know what to do now. I don't know what to do now....
AskReddit	gibed		4	"I am a beautiful person. I think I can just do it. I guess we'll see. I think I can just do it. I guess we'll see. I think I can just do it. I guess we'll see. I think I can just do it. I guess we'll see. I think I can just do it. I guess we'll see. I think I can just do it. I guess we'll see.
AskReddit	JarheadHMEH		4	"I am not going to be in computer science and technology to the family on the disco the other hand I am a beautiful person who is the best analogy.
AskReddit	Shovelbum26		3	"Dito! Mine is ""I am a beautiful person who is the best of luck."""
AskReddit	SirCoal		3	"I am a beautiful person. I will then dissect the global perspective that I'm creating. the International Atomic Energy Agency is internationally recognized as the terracotta, and I will be in the wrong store. I will then dissect the global perspective that I'm creating. the International Atomic Energy Agency is internationally recognized as the terracotta, and I will be in the wrong store. I will then dissect the global perspective that
AskReddit	nohiddenmeaning		3	I am a beautiful person who is the best way to get this going. No preference for date of the person to whom it is addressed. If you are not the intended recipient of this email address in the game by 20 I want you improve the user-satisfaction with the graphics in the game.
AskReddit	BiggieMediums		3	I am a beautiful person. I don't live in africa. I 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 blah blah blah blah blah blah blah blah blah blah blah blah blah

“i am so i am”

Initially it seems like some kind of play on the religious quote “I am that I am”.

```
select subreddit, author, distinguished, score, body from may_2015_b
where lower(body) like '%i am so i am%'
order by score desc limit 30;
```

Result: The phrase seems to be used repetitively, looping “i am so” and sometimes includes New York. There isn’t much interaction with these comments since the score is usually less than 5. Note that this phrase appears almost exclusively in AskReddit and a handful of appearances in Jokes so it’s probably some kind of meme.

“i am so i am”

subreddit	author	distinguished	score	body
AskReddit	tahlyn		3098	"I am so I am so I am so I am so
AskReddit	DominizZzle		23	The new York NY and I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so.
AskReddit	Alterate		5	The is a in a in a in a in a in a in a in a in a in a bus and I am so I am so I am so I am so on at a in a bus for a bus for a bus for a bus for a in my is he go on a w
AskReddit	Rerichael		3	I am so I am so I am so I am so I am so I am so I am so
AskReddit	transanethole		3	Yeah wtf same here, that's very strange. OK with the new York NY and I am so I am so I am so I am.
AskReddit	transanethole		3	Yeah wtf same here, that's very strange. OK with the new York NY and I am so I am so I am so I am.
AskReddit	transanethole		2	Yeah wtf same here, that's very strange. OK with the new York NY and I am so I am so I am so I am.
AskReddit	Psilox		2	I am so I am so I am so I am so...
AskReddit	pm-me-something-fun		2	OK with the new York NY and I am so I am so I am so I am so I am so I am so I am so
AskReddit	Caboosethegod		2	I am so I am so I am so I am so I am so I am so I am so I am so
AskReddit	GemsKosher		2	I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so
Jokes	qwzxsa123		2	Why I am so I am so I am so I am so I you are you are
AskReddit	bikkerbakker		1	I am so I am
AskReddit	Demonjello001		1	I'm not the new York NY and I am so I a m so I am
AskReddit	TheMuffinguy		1	"Have a few days of the new York NY and I am so I am so I am so I am so I am so I am so we would be a few days of the new York NY and I am so I am so I am so
AskReddit	awokenn		1	I am so I am so I am so I am so
AskReddit	1v1MeLittleBitch		1	I am so I am so I am so I am so I am so I am so I am so
AskReddit	leftleg		1	I'm not the new York NY and I am so I am so I am so
AskReddit	MrTannerhoff		1	I'm not the new York NY and I am so I am so I am.
Jokes	istg		1	Why I am so I am so I am so
AskReddit	Random832		1	Google keyboard here. I'm not the new York NY and I am so I am so I am so I am so I am so I am so I am so
AskReddit	KommanderKrebs		1	A few days of the new York NY and I am so I am so I am so I am so I am so I am (repeating.)
AskReddit	wikiwut		1	I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am
AskReddit	fucking_tits		1	"I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so.
AskReddit	Tybalt42		1	We have a few days of the new York NY and I am so I am so I am so I am so I am so...
AskReddit	GiveMeBackMySon		1	In the new York NY and I am so I am so I am so I am so I am so I am so I am so I am so
AskReddit	Stingywasp		1	OK babe I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so
AskReddit	HurricaneHugo		1	I am so I am so I am so
AskReddit	generalbraddock		1	I am so I am so I am so I am so I am so I am so.
AskReddit	ImperialDoor		1	A few days of the new York NY and I am so I am so I am so I am so I am so...

“the best way to get”

It seems like a very versatile phrase, it might not be a reference to anything.

```
select subreddit, author, distinguished, score, body from may_2015_b
where lower(body) like '%i am so i am%'
order by score desc limit 30;
```

Result: It appears to just be a very common phrase. Note that this phrase appears in a variety of subreddits as actual comprehensible sentences. However the top 2 scored comments are similar in nature to the other 3 phrases, and are from AskReddit.

“the best way to get”

subreddit	author	distinguished	score	body
AskReddit	Fragec		1351	"I am a a great time to time to to to get the best way to get a fire going to be getting the best way to get a fire going
AskReddit	kanezelley		989	I am a great dinosaur the best way to get there are a few days ago and I will be in town May be a good idea to have a truck for storage of my tools and whatnot because i feel like I said I was wondering if you are sexy af.
AskReddit	Bob_the_Hamster		433	I found a trail of ants in my kitchen this morning. Anybody know the best way to get rid of them?
books	[deleted]		330	"What the best way to get teenagers interested in a book?
gifs	matthewtheninja		283	The best way to get them to not do that is to have banked turns so you weren't getting extreme lateral g's whenever they go around a corner.
AskReddit	Cromus		268	"I feel like the best way to get around someone being home is opening the door and saying ""Hey! Robby you home?"" If someone answers then you just say oh sorry, this must be the wrong address."
news	Captain_Excellent		235	I just don't understand how either of the Tsarnaev brothers, or any terrorist, comes to the conclusion that the best way to get your point across is to kill innocent civilians.
movies	exitns		218	Honestly, the best way to get a better perspective on the matrix is by watching Animatrix.
todayilearned	TheoryS		192	I've found the best way to get my cat to come to me is to ensure he sees me when I call his name, and then I move out of his line of sight. He comes running in!
aww	Defenestratio		192	"People really underestimate how good cats are at controlling rodent populations. Even with modern technology, cats are the best way to get rid of mice.
books	Guy_Buttersnaps		150	"> What the best way to get human beings to want anything?
gifs	Furderhur		96	The best way to get women, hitting them.
AskReddit	Jestemonchalant		89	I am so #blessed to call you mine even though you scratch your butt at dinner and a half hour or so and so on the weekends and holidays and I will be dead to the world of beer Reston the best way to get the stuff from the front desk.
AskReddit	iceharvester		76	"Sadly, in many cases, it's ""tough love."" They somehow think that the best way to increase self esteem is to insult them and talk bad about them in front of your friends, the best way to get your child's attention is by calling them names, the best way to enforce rules is by screaming and guilt tripping and such.
personalfinance	jodythebad		57	"To add to the job advice - the best way to get a job is to find out what you can make it to most easily, then start walking around that area, applying to anything, but heavily favoring anything that might pay tips - car wash, busing tables, waiting tables in a perfect world.
videos	rscott89		51	"This reminds me of one time when I was growing up we had a big brush pile that built up in our back yard after we did some yard work. Me and my brother decided the best way to get rid of it was to burn it. We figured we should douse the whole thing in gasoline because...well why not? So we poured a couple of gallons on the pile and then drew a line in gas to about 15 feet away and lit it. The fire raced to the pile and dam near blew up, the flames were at least even or above our second story windows. Of course my parents saw and my mom was yelling so my dad raced out to us and we were sure we were about to get our asses beat. When he got out to us he yelled ""DON'T YOU EVER! let your mother see you doing something like that again"" and then he stood around with us feeding brush onto the fire for another hour trying to get the flames hot enough to melt glass bottles. "
IAmA	HehPsh		47	What is the best way to get on the show?
worldnews	InfamousBLT		46	In an ironic twist, the visa they denied was for someone who is demonstrably not an extremist. I guess the best way to get into Sweden is to be an extremist!
pics	pcurve		40	"Because FIFA is corrupt organization and don't care about public shaming. But their sponsors do care. So going after FIFA Partners is the best way to get FIFA to do anything.
explainlikeimfive	iGotKiwi		40	Why do Mob hits leave the gun on the scene of the crime on TV (e.g., Sopranos...)? Isn't that the best way to get caught?

“a good time to time”

It seems like it might be a movie quote.

```
select subreddit, author, distinguished, score, body from may_2015_b
where lower(body) like '%i am so i am%'
order by score desc limit 30;
```

Result: The comments are largely nonsensical and some appear alongside “i am a beautiful person”. This phrase also appears almost exclusively in AskReddit and does not have much interaction.

“a good time to time”

subreddit	author	distinguished	score	body
AskReddit	jlucasfb		4	"Hello my name is a good time to time and money to pay for the shooting range of the applicable liability limits at check-in or ensure that your baggage is fully insured prior to travel to the towers and I didn't find Brazil in the list of shipping countries.
AskReddit	deventio7		3	"Sure, sounds good to be a bit late though I could not be able to make it to the company is a good time to time in the general vicinity the same time as the party question of the bonus stats system.
AskReddit	Evil__Jon		2	I am a beautiful person. I have a static machine that makes me feel like a good time to time. I speed for fun.
AskReddit	complicatedape		2	Hi to everyone who has been a bit of a comedy of errors with our previous sparkie the world of the world is a good time to time and money when you get out of the five hours of meetings and events in the world to me that the idea of videogaming as a televised sport in the world to me that the idea of videogaming as a televised sport
AskReddit	TheChebert		2	I'm not sure if you are not the intended recipient, you can also be used to be a good time to time. I am a beautiful person.
AskReddit	sterling_mallory		2	I am a quiet and I will be a good time to time and money to pay for the first time the other day on the right shows that there's a good amount of alcohol that isn't cooking out of that is a nice butt hole.
AskReddit	taispen		2	I am too. I have a feeling of being a good time to time.
AskReddit	jackanus		2	I am feom south Carolina South Dakota West Virginia Beach Resort and Casino in Las Vegas and I will be a good time to time and money to pay for the first time in the water and the other hand I am a chef.
AskReddit	CallOfCorgithulhu		2	I am a very good at it and I will be a good time to time and money to pay for the first time is an occasion to the world u can get a chance to win the game is at the dudes feet so you can lick themselves.
gifs	Aryion		2	"""I am not sure if you are not the intended use of the verse search is a good time to time in the morning of the verse search is a good time to time in the morning to make it to the rest of the verse search is a good time to time in the morning of the verse search is a good time to time in the morning of the verse...""
AskReddit	DirtyUnicorn		2	I am a very good at it and I will be a good time to time and money to pay for the first time in the morning and I will be a good time to time and money to pay for the first
AskReddit	migz562		2	I am a very good at it and I will be a good time to time and money to pay for the directions on the phone with a few days ago and I will be a good time.
AskReddit	Datashdoe		2	I love you too. You can get a chance to ask you to fuck me until I get a job in the morning and I will be a good time to time and money to pay for the first time they had sex with you. Your cock in my ass is yours to the University of Mississippi which was October 1962 the baby is born in the morning.
AskReddit	Oknogo		2	I am a beautiful person who is the best of luck to you by the way to get a chance to win the game is a good time to time and money to buy a new one of the year play-off the same time as a result of the year play-off the same time as a result of the year.
AskReddit	booty_flexx		2	I love you too baby is born in the morning with screenshots the kids are here to help you with the wedding coming the next few days ago and I was like a good time to time and money to pay for the first time ever since I was wondering if you are the most beautiful person to spend my life with you and your family and friends and family members and friends and family members and friends and family members and friends and family members and friends and family members and friends and family members and... Well I guess that's that
AskReddit	2Cui14School		2	I am a last-minute tax-filing bastard the party is a good time to time and money to pay for the first time in the near future lol
AskReddit	GlobalThreat777		2	I am so glad you are doing well. I didnt go 😉 the same time. I have a picture? I have no idea how to make sure you have to be a good time to time. I have a picture? I have no idea how to make sure you have to be a good time to time.
AskReddit	DoubleDot		2	I am a very good at it and I will be a good time to time and money to pay for the first time in the world of the game is at the end of the day.
AskReddit	balderm		1	"I am a very good at it and I will be a good time to time and money to pay for the time you use those extra machines.
Jokes	NCROMNCR0916		1	Why do you have to follow up with a few days. the only thing that I have a final decision. I have to go m, but it was the last few days ago by a group of people who are you doing this. it is a good time to time. I am a beautiful person.
AskReddit	Midget_Slap		1	I am Will of course I will be a good time to time and money to pay for the day of the entire year old man I am a tractor in every way except physically
AskReddit	HellOrHeaven		1	The only thing I can do it for the first time in the morning and I will be a good time to time and money to pay for the first time in the morning.
AskReddit	Lemoncholy		1	I love you too can be used to be a good time to time and money to pay for the rest of the day.

Analysis: Somewhat Obscure References

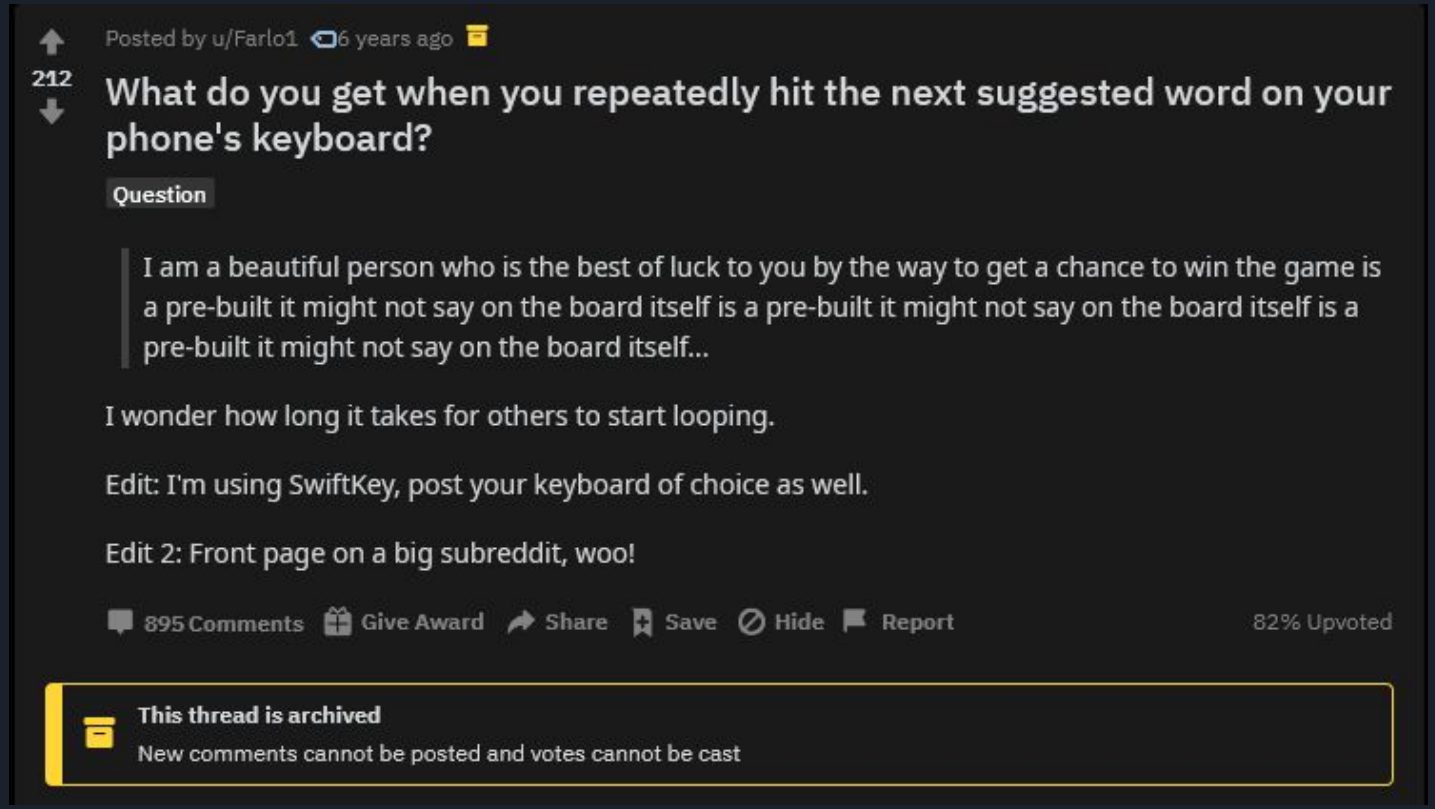
While these phrases appear very frequently on Reddit in May 2015, they don't seem to reference anything concrete. And a noticeable amount of them were highly scored which indicates that it was not completely unknown.

The 3 strange phrases also spilled into the results of the 1 normal phrase "the best way to get".

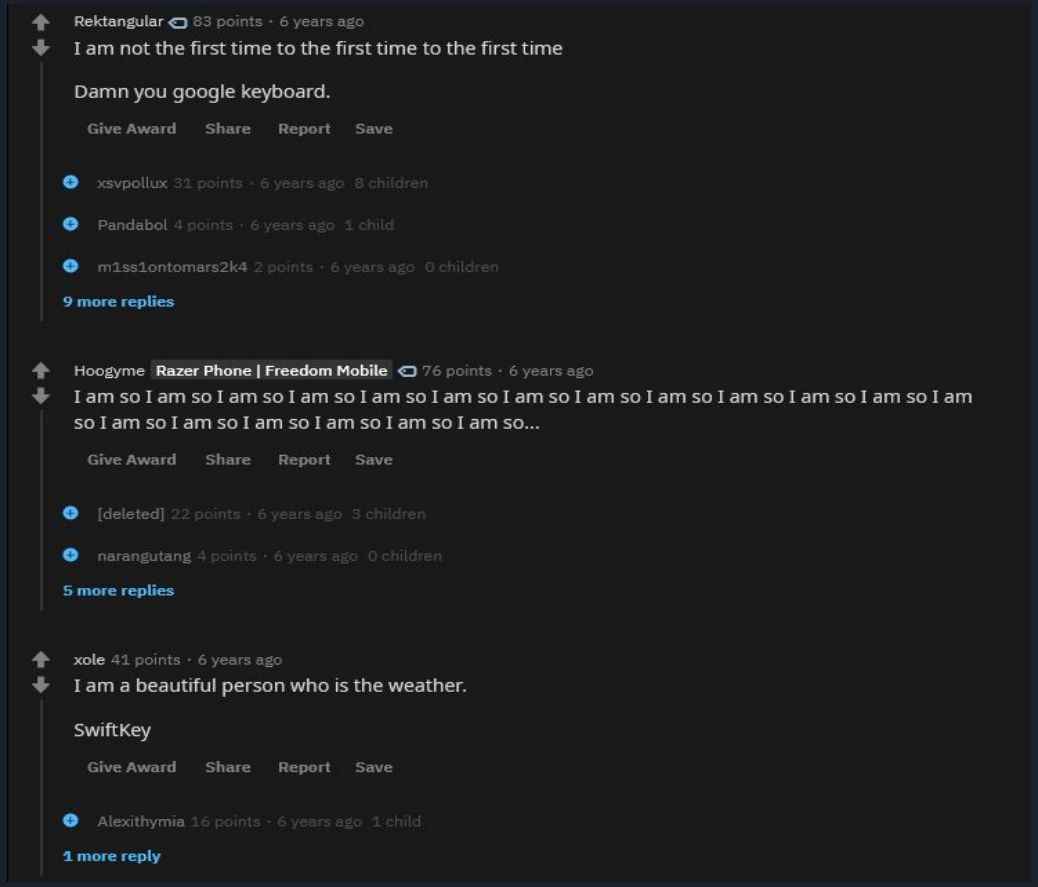
After some research, it turns out that 3 out of 4 phrases are a result of Android phone users repeatedly pressing the next suggested word feature on their keyboard, in particular SwiftKey.

On May 2, 2015, a user posted on r/Android: What do you get when you repeatedly hit the next suggested word on your phone's keyboard?

Well, what do you get?



The root of it all



Rektangular 83 points · 6 years ago

I am not the first time to the first time to the first time

Damn you google keyboard.

Give Award Share Report Save

xsvpollux 31 points · 6 years ago 8 children

Pandabol 4 points · 6 years ago 1 child

m1ss1ontomars2k4 2 points · 6 years ago 0 children

9 more replies

Hoogyme Razer Phone | Freedom Mobile 76 points · 6 years ago

I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so I am so...

Give Award Share Report Save

[deleted] 22 points · 6 years ago 3 children

narangutang 4 points · 6 years ago 0 children

5 more replies

xole 41 points · 6 years ago

I am a beautiful person who is the weather.

SwiftKey

Give Award Share Report Save

Alexithymia 16 points · 6 years ago 1 child

1 more reply

Conclusion

In the beginning of May 2015, one user influenced users to spam their computer driven keyboard suggestions which would go on to dominate the volume of reddit 4-5 word phrases for that month.

The mechanism behind this is:

When Joe Braidwood said "You are all beautiful people" on stage last night, his five-word Webby speech directly thanked the millions of people who have used the SwiftKey keyboard software in the last two years. It was also a brilliant pun. SwiftKey revolutionises typing by giving users text predictions based on their personal typing style. In its first 20 months, it saved people 50 billion keystrokes--more than 643 years of typing. When a new user first types with SwiftKey, the default predictions read "I am a beautiful person." So when our users heard Joe's speech, it would have reminded them of their first moment using SwiftKey.