# GENERATION OF FAQ FROM Q & A DATASET

ARJUN AHUJA
CS14BTECH11004

SAFWAN MAHMOOD
ES14BTEC11017

AASISH PATOLE
CS14BTECH11026

# PROBLEM STATEMENT

- Given a set of questions from E-COMMERCE platform, find a list of generally asked questions by customers based on a category
- A category can be a product or a broader class like the one we used: cellphones and accessories.

http://jmcauley.ucsd.edu/data/amazon/

# UNDERSTANDING THE DATA

- We have tested our program on Cell Phones and accessories dataset, which contains around 85,865 questions.
- The data is in the following format:


```
{
  "asin": "B000050B6Z",
  "questionType": "yes/no",
  "answerType": "Y",
  "answerTime": "Aug 8, 2014",
  "unixTime": 1407481200,
  "question": "Can you use this unit with GEL shaving cans?",
  "answer": "Yes. If the can fits in the machine it will despense
hot gel lather. I've been using my machine for both , gel and
traditional lather for over 10 years."
}
```

# PREPROCESSING THE QUESTIONS DATASET

- Converted questions from a sentence to a vector model
- Removed stop-words(a, is, the, etc).
- Did Porter Stemming(Optional)
- Made an inverted index of words
- Made a general term document matrix for finding the TF-IDF.

# HANDLING SPARSE DATA

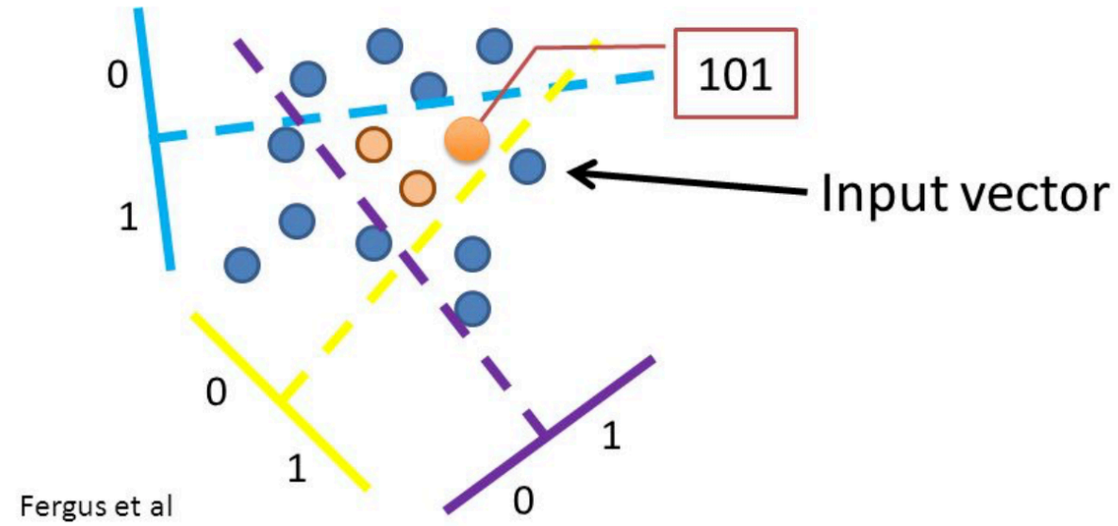
- We found 95% of zeros were there in the final TF-IDF matrix obtained during preprocessing of data.
- That's why we felt the need to do compressing of data, as normally to was taking too long to run LSH, with compression of data we observed that time consumed reduced by around 5-7 times.

# OUR APPROACH FOR THE PROBLEM

- For Finding the most frequently asked questions, we thought that why not use the same method as finding duplicates?
- We can use a similar approach but modify it slightly to have it transform into most frequently asked questions.
- So,  we thought why not see how much the bucket size in LSH is, as a bucket in LSH represents similarity between documents, if we can just take all the buckets having more than a particular number of documents our work is complete.
- We used a technique called random hyperplanes for constructing LSH.

# RANDOM HYPERPLANES AND LSH

- We used a technique called random hyperplanes for constructing LSH.
- For this given the number of hash-tables, for each hash table we create a number or hyperplanes, each one of which divides the space into two buckets(we classify them as 0 and 1)
- Each input point in a hash-table (current space) will thus be allocated a hash value, depending on which side of hyperplane it belongs to.



Fergus et al

# RANDOM HYPERPLANES AND LSH

- So overall in a hash-table, as each hyperplane(h) divides the plane in two buckets in total there will be around 2^h buckets for h hyperplanes.
- Now for each document we assign a hash value depending on term-document matrix.
- Finally to find the number of points in a bucket just check how many input points have the same hash value for this hash-table, and check if it is greater than frequency threshold.

# RESULTS AND OBSERVATIONS

```
------------best------------
Will this phone work in the UK, Germany and Turkey...??? 897
------------best------------


does this phone work with Verizon Wireless? 12
Does this work with the Panasonic KX-TS4200 phones? 298
can this phone work on US Cellular? 402
Will this phone work in Australia? 492
Does this phone work with T-Mobile in USA? 638
T Mobile: Does this phone work with Tmobile? 695
does this phone work in Spain and France? 803
Does phone work in Greece 804
will this phone works in Honduras? 806
will this phone work in India? 811
does this phone works with Straight Talk ? 813
Will this phone work in the Dominican Republic? 894
Will this phone work in the UK, Germany and Turkey...??? 897
------------ 10100011110011110100000010100001 ------------


------------best------------
Does this work with a Samsung NOTE 4? 64 1.0
------------best------------


Will this work in Israel? 14
does it work for Galaxy S 2 28
Does this work with Galaxy Tab 4? 37
Will this work in South Africa? 41
Will this work on the S4 ? 45
Does this work with a Samsung NOTE 4? 64
will it work on LG Optimus Dynamic 69
Does it work on the T-Mobile Samsung Galaxy S1 (USA Version)? 85
Will this work with a Plantroncs CT14? 146
Will this work in the Phillipines? 161
------------ 0000011111110101 ------------
```

# BAD RESULT RETURNED BY LSH:

```
------------best------------
I have a samsaung note 2 with a slim case on it , will it still fti in the p
ouch ?? 235 0.142857142857
------------best------------


and for the Galaxy S4?? 127
DOES NOT WORK WITH VERIZON SERVICE 137
I have a samsaung note 2 with a slim case on it , will it still fti in the p
ouch ?? 235
Music 295
DOES IT WORK WITHLG35G 304
Size 317
is this AA or AAA 367
------------ 00000000 ------------
```

# AS IT CAN BE CLEARLY SEEN THOUGH THIS BUCKET IS VERY CROWDED, NONE OF THE SENTENCES HAVE MUCH RELATION WITH EACH OTHER

# AVERAGE COSINE SIMILARITY(ACS)

- We used average cosine similarity metric, which is defined as $\Sigma$ cosine_similarity(a,b)/(n-1) , where n is the number of points in a bucket.
- We find the best representative from a set of questions, by finding the one with the highest average cosine similarity.
- And after this we check if this threshold is above a certain mark, we found that our results were best around ACS = 0.4.
- This removes the problem described in last slide.

# VARYING THE NUMBER OF HYPERPLANES (HASH-SIZE)

- Increasing the hash size gives us better result.

```
----------- 1000 -----------

------------best------------
will this phone work with straight talk? 72 0.159958684022
------------best------------


Will the work with lg optimus g ls970? 3
can I use it in Europe with a converter to 220V ? 11
Would they be activated to my personal Google account once I receive them? 2
1
Will this work for samsung galaxy s4 SGH-i337 (AT&amp;amp;T)? 36
Is this phone new, with 1 year manufacture warranty? 37
Does it come with all the current apps? 41
will this phone work with straight talk? 72
----------- 1000 -----------
```

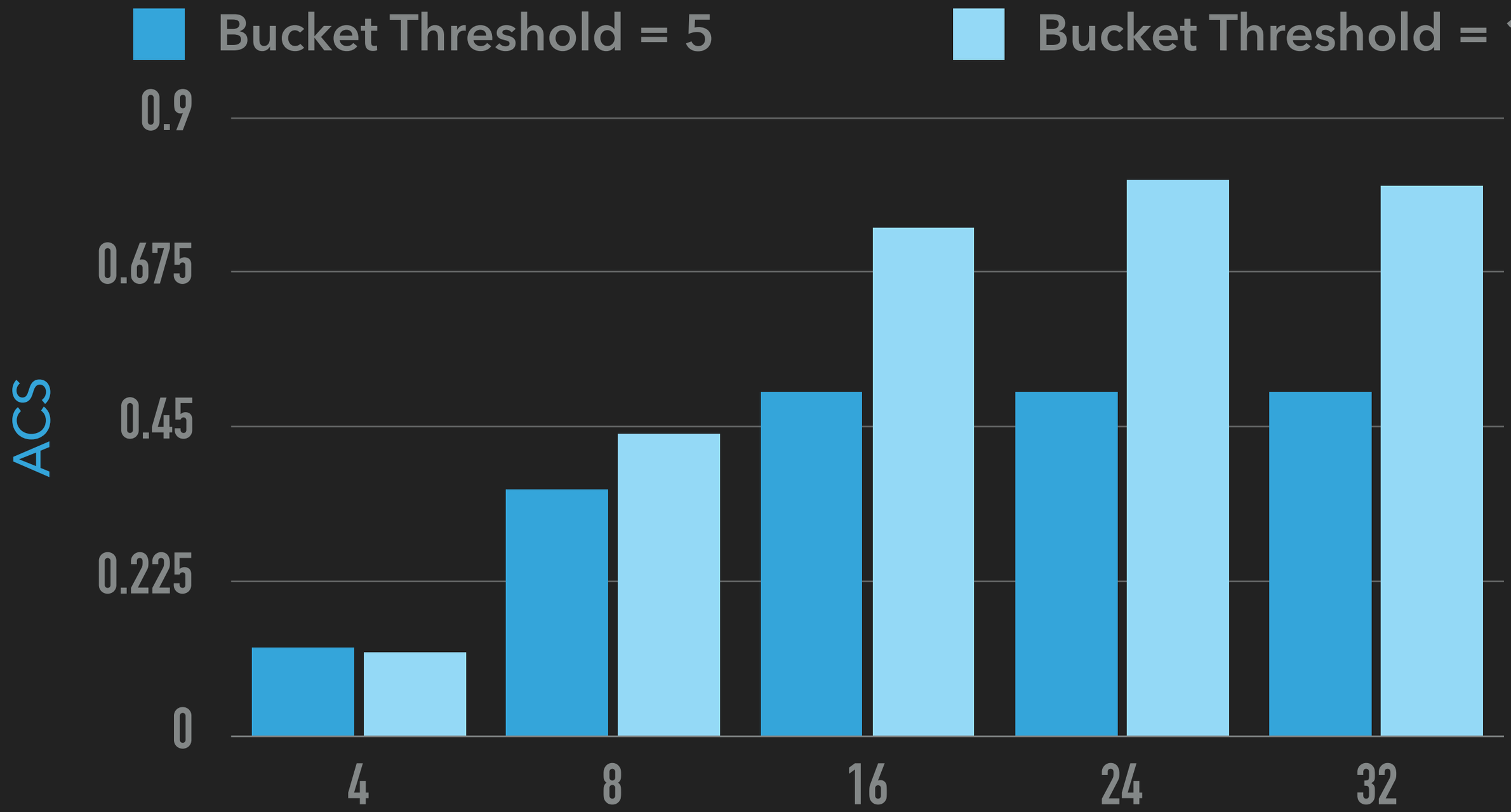
## CASE WITH HASH-SIZE=4

```
------------best------------
Will this work with a Plantroncs CT14? 193 0.944444444444
------------best------------


does it work for Galaxy S 2 13
Will this work in South Africa? 28
Will this work on the S4 ? 31
Will this work with Galaxy S3? 45
Does this work with the Moto G? 170
do this work with the CS50 181
Will this work with a Plantroncs CT14? 193
Will this work in the Phillipines? 222
Will this work in Israel? 244
Would they be activated to my personal Google account once I receive th
em? 262
Does this work with a Samsung NOTE 4? 272
will it work on LG Optimus Dynamic 277
Does it work with Motorazr2 V9m 309
it works with PS3? 353
Does this work with Galaxy Tab 4? 375
Does it work on the T-Mobile Samsung Galaxy S1 (USA Version)? 399
Will this work with Samsung Galaxy SIII? Thank you 400
Does this work with Kindle Fire? Does this work with Kindle Fire? 430
------------ 01110010 ------------
```

# CASE WITH HASH-SIZE=8

# AVERAGE COSINE SIMILARITY OF BEST RESULTS VS HASH-SIZE



# EXTENSION AND FUTURE WORK(JUST AN IDEA)

- While doing the project we observed that the output we obtained using ACS metric, did not actually cover all the cases. Example, in the given output: we see actually for the output, wouldn't it be better if we had something like this as output:
- Will this work in?
- Will this work on/with?
- Does it work with?

```
------------best------------
Will this work with a Plantroncs CT14? 193 1.0
------------best------------


does it work for Galaxy S 2 13
Will this work in South Africa? 28
Will this work on the S4 ? 31
Will this work with Galaxy S3? 45
Does this work with the Moto G? 170
do this work with the CS50 181
Will this work with a Plantroncs CT14? 193
Will this work in the Phillipines? 222
Will this work in Israel? 244
Does this work with a Samsung NOTE 4? 272
will it work on LG Optimus Dynamic 277
Does it work with Motorazr2 V9m 309
it works with PS3? 353
Does this work with Galaxy Tab 4? 375
Does it work on the T-Mobile Samsung Galaxy S1 (US
A Version)? 399
Will this work with Samsung Galaxy SIII? Thank you
 400
Does this work with Kindle Fire? Does this work wi
th Kindle Fire? 430
------------ 1001100010101010 ------------
```

# OBSERVATION

- We observed that selecting one sentence as a representative for the whole bucket, does not generally cover all the cases, as explained in previous slide.
- To fix this problem, we feel that we should take words that mostly influence the cosine similarity metric and ignore the ones that have minute, or no effect. Example in the previous slide words like Does, with, will, work, on, in, have a lot of influence in finding ACS.
- While words like kindle Fire have no, or minimal effect.
- What if we can tweak the cosine similarity metric slightly to accommodate this is place, with this we can get a better understanding of generally asked question type.

# THANK YOU