Personalized content recommendation for StackOverflow data

Karthik Sheshadri, Nandini Rangaswamy

StackOverflow

- 32 million users every month
- 63% of users visit more than once daily
- A question is asked on StackOverflow every 8 seconds

Recommendation Systems

"Search is what you do when you're looking for something. Discovery is when something wonderful that you didn't know existed, finds you."

Recommendation Systems

- 67% of movies on Netflix are recommended
- 35% of Amazon sales are from recommendations
- Google news recommendations generate 38% clickthrough

StackOverflow Newsletter

Stack Overflow Newsletter

Tuesday, June 28, 2011

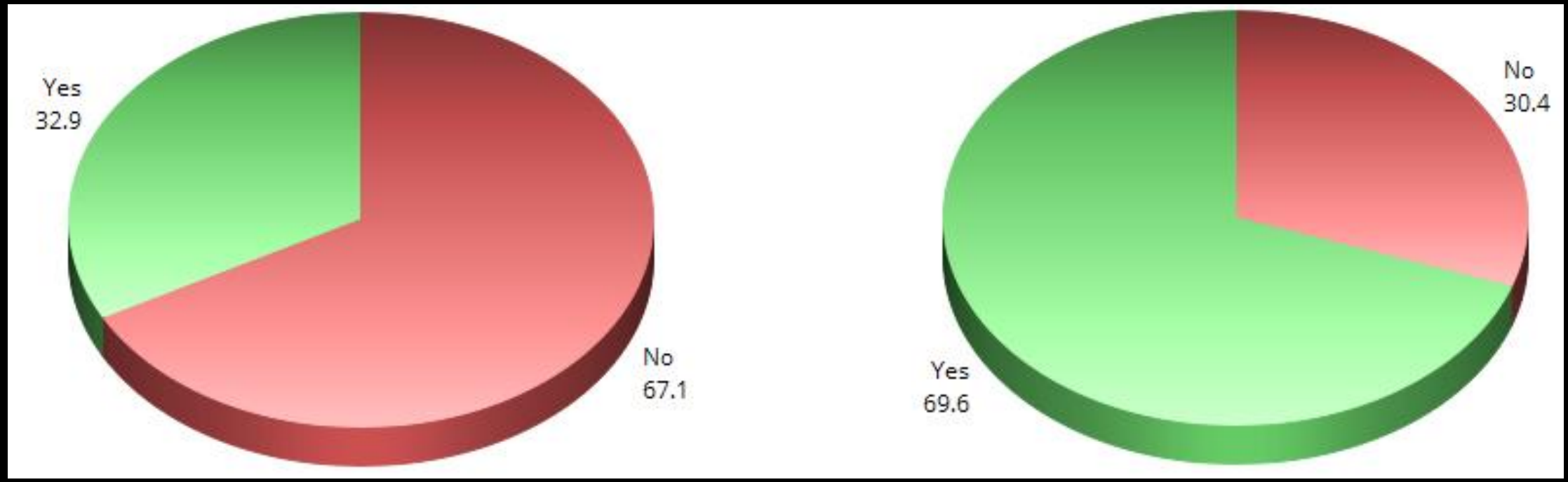
Hot questions this week:

- Eric Lippert answered "Can a local variable's memory be accessed outside its scope?!"
- Stephen Canon answered "Protecting executable from reverse engineering?"
- In C++, why should `new` be used as little as possible?
- Ignacio Vazquez-Abrams answered "parseInt(null, 24) === 23... wait, what?"
- Armen Tsirunyan asks, "What differences, if any, between C++03 and C++0x can be detected at run-time?"
- James McNellis asks, "How can I reliably get the address of an object?"
- Optimizations for pow() with const non-integer exponent?

Can you answer these?

- Random Access of Large Media Files on a Remote Web Server
- Producing a WAR file from a django project with SQLite
- Text align problem when using Arabic font.

User Interest Survey



Left:Are you satisfied with the SO newsletter?
Right:Would you like personalized recommendations from SO?

StackOverflow Data dump

- Anonymized data dump of 2014-2016
- 40 Gigabytes of data
- We work off of a 2% downsample

Content Based Filtering

Recommendation are generated by **matching** the **features stored** in the user profile **with those describing the items** to be recommended.

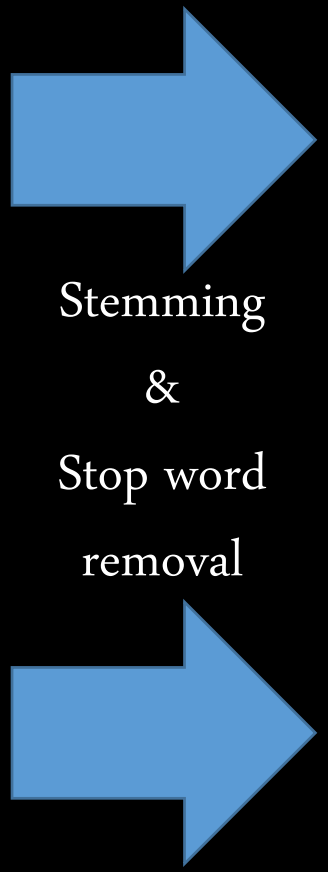
user profile

items

Preprocessing

Stackoverflow Post

In my android application I have a text view that displays text containing special characters. The text view somehow automatically breaks strings at the characters '/' and '-'

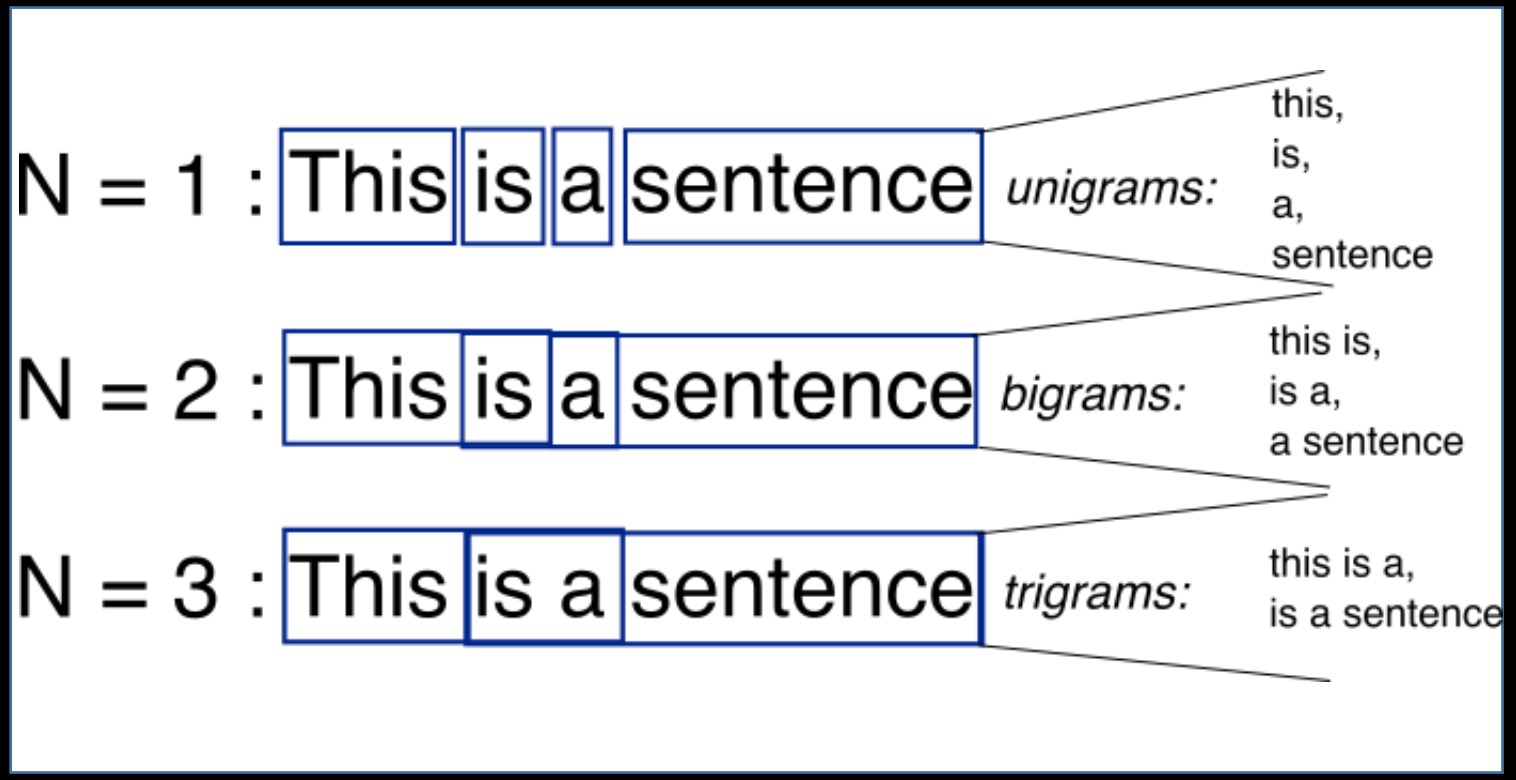


Android

Application

Text view

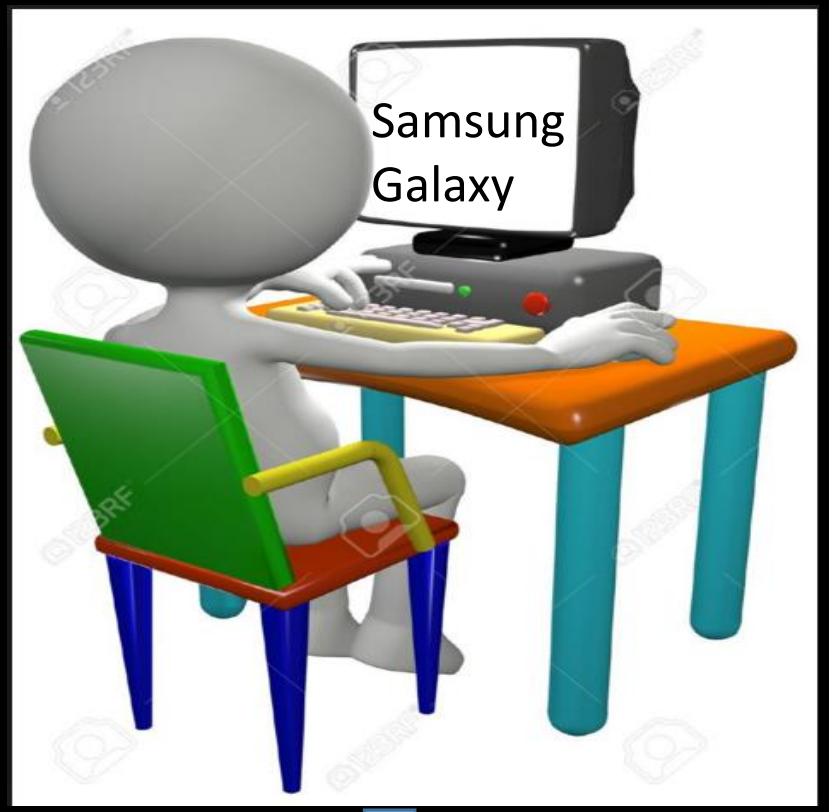
N-gram + edit distance



N-grams, an example

N-gram + edit distance

User Input



N-grams

Preprocessed text

Android

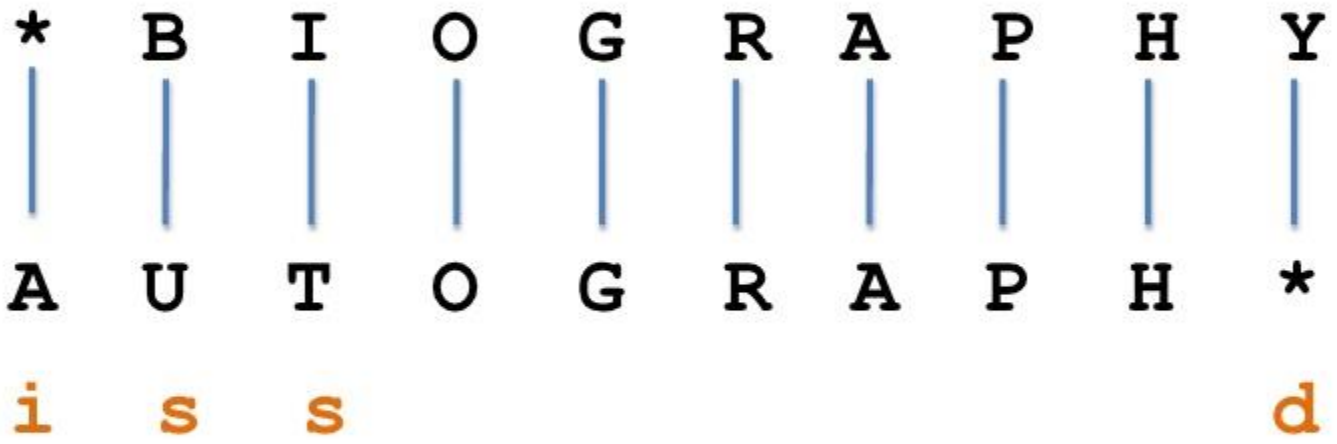
Application

Text view

N-grams

N-gram + edit distance

Minimum Edit Distance (Example)

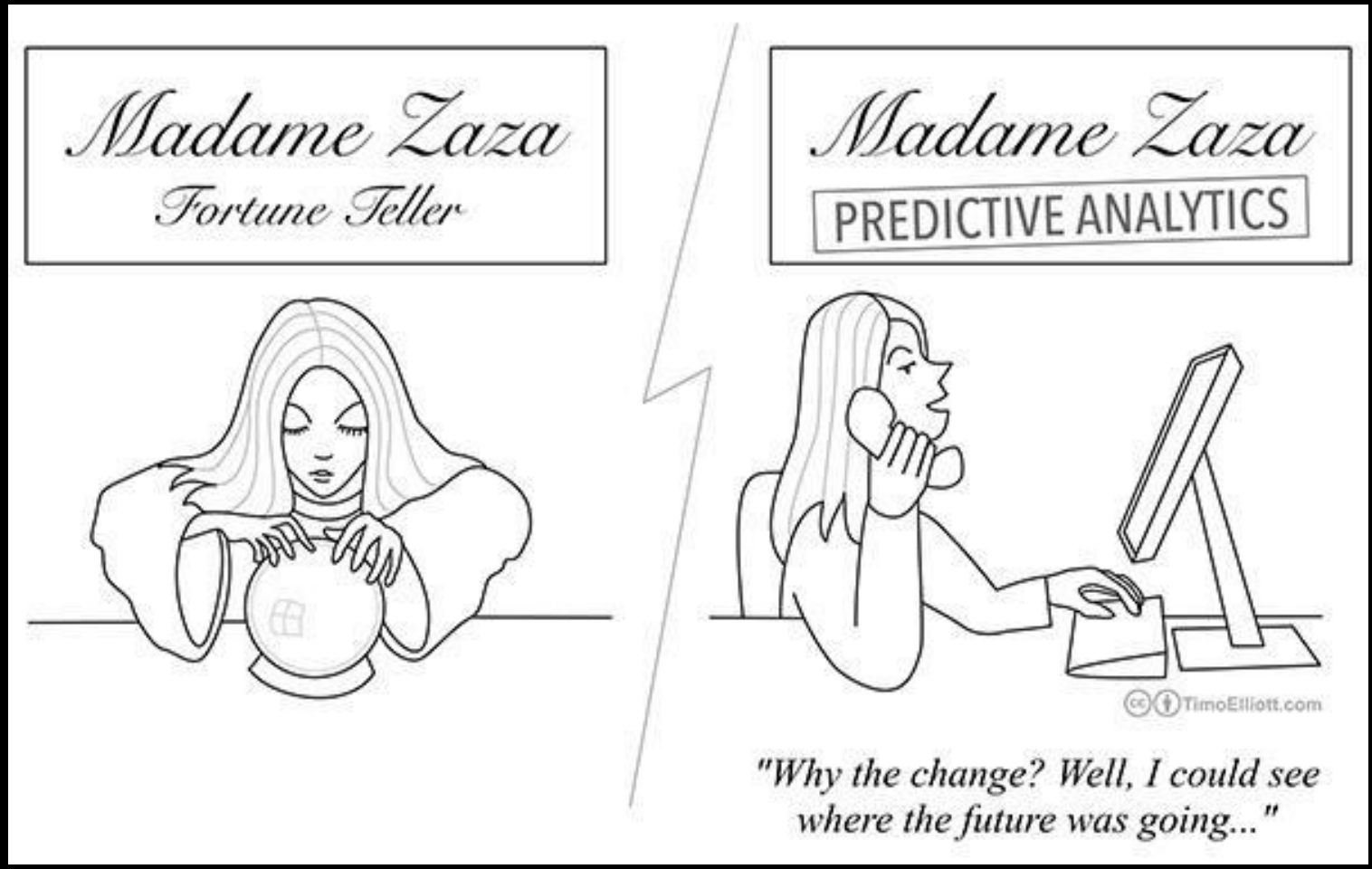


- Let cost of each operation be 1
  - Total edit distance between these words = 4

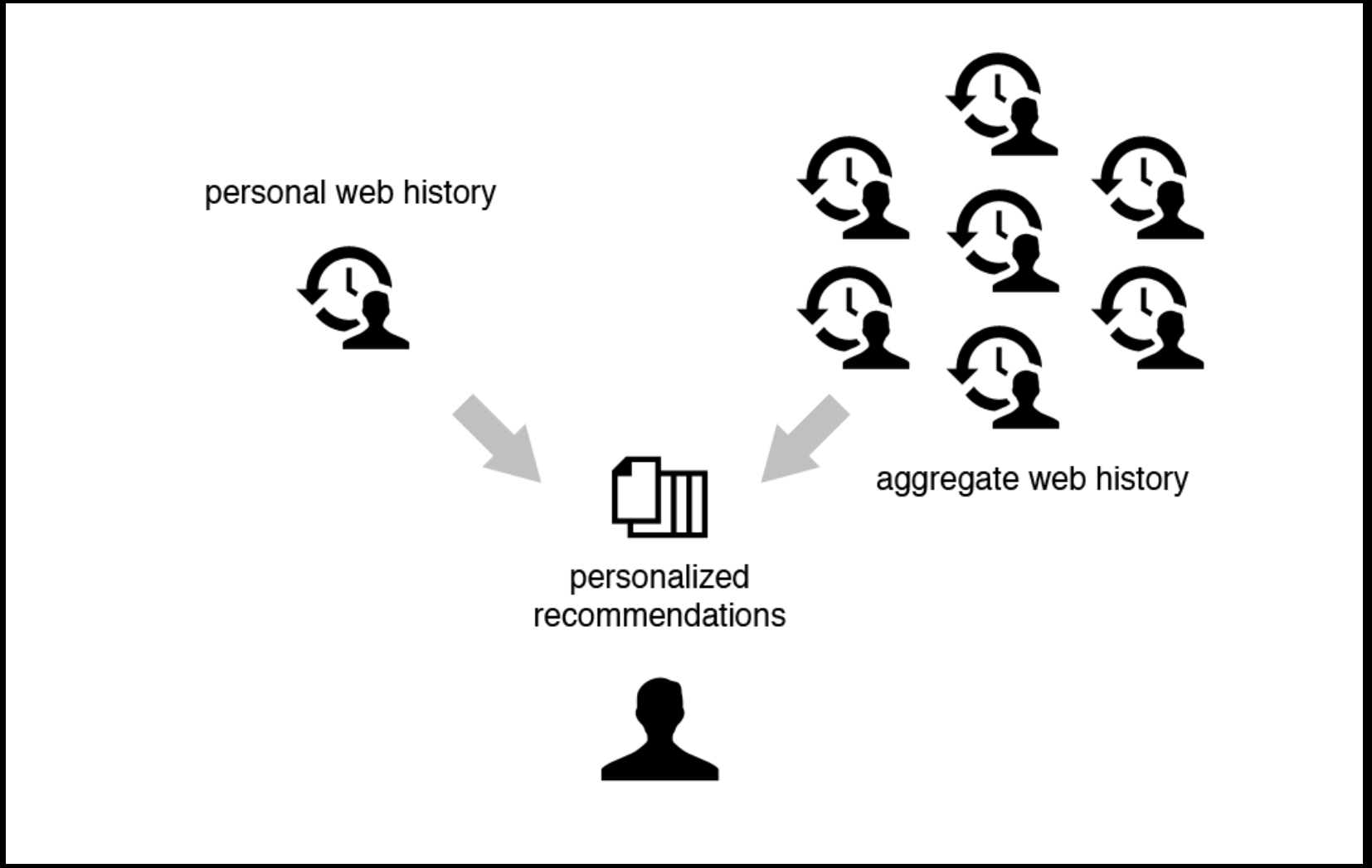
N-gram + edit distance

Return top k posts that minimize edit distance
to nearest n-gram

“Learning” user preferences

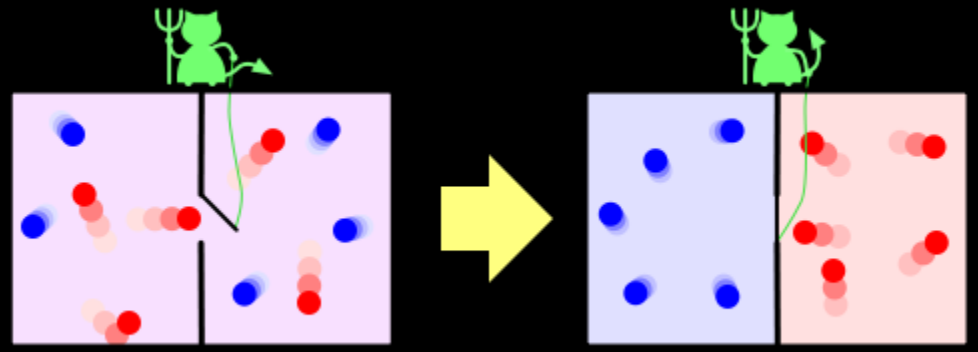


Collaborative Filtering

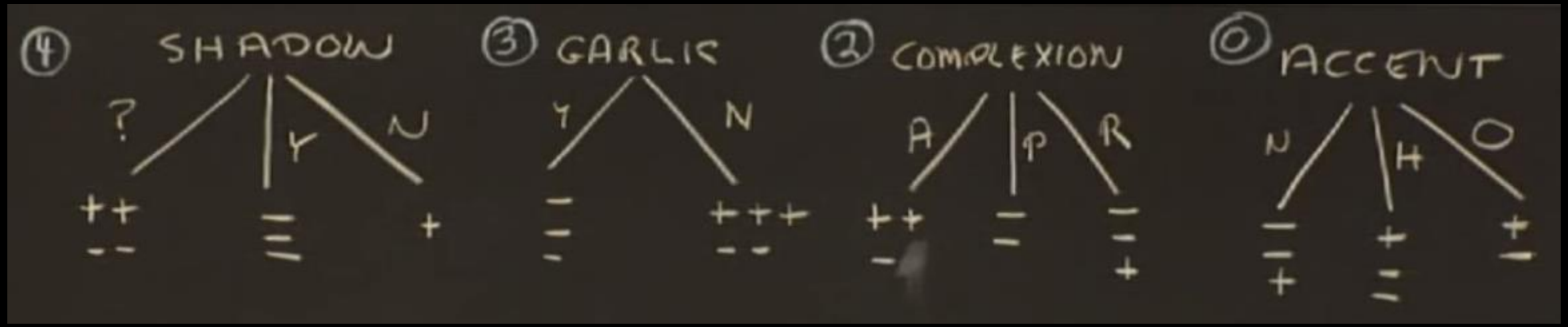


Random Forests

Vampire?	Shadow?	Garlic?	Complexion?	Accent?
No	?	Yes	Pale	None
No	Yes	Yes	Ruddy	None
Yes	?	No	Ruddy	None
Yes	No	No	Average	Heavy
Yes	?	No	Average	Odd
No	Yes	No	Pale	Heavy
No	Yes	No	Average	Heavy
No	?	Yes	Ruddy	Odd



Vampire?	Shadow?	Garlic?	Complexion?	Accent?
No	?	Yes	Pale	None
No	Yes	Yes	Ruddy	None
Yes	?	No	Ruddy	None
Yes	No	No	Average	Heavy
Yes	?	No	Average	Odd
No	Yes	No	Pale	Heavy
No	Yes	No	Average	Heavy
No	?	Yes	Ruddy	Odd



TF-IDF

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

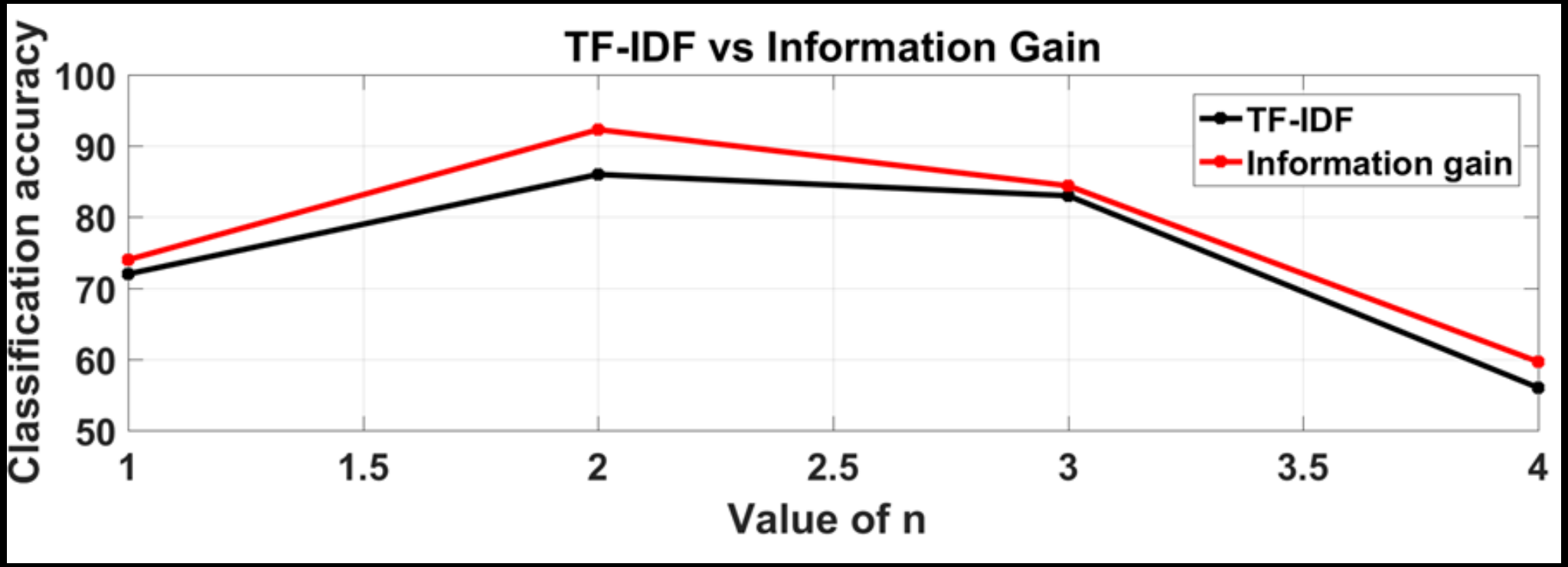
Term x within document y

$tf_{x,y}$ = frequency of x in y

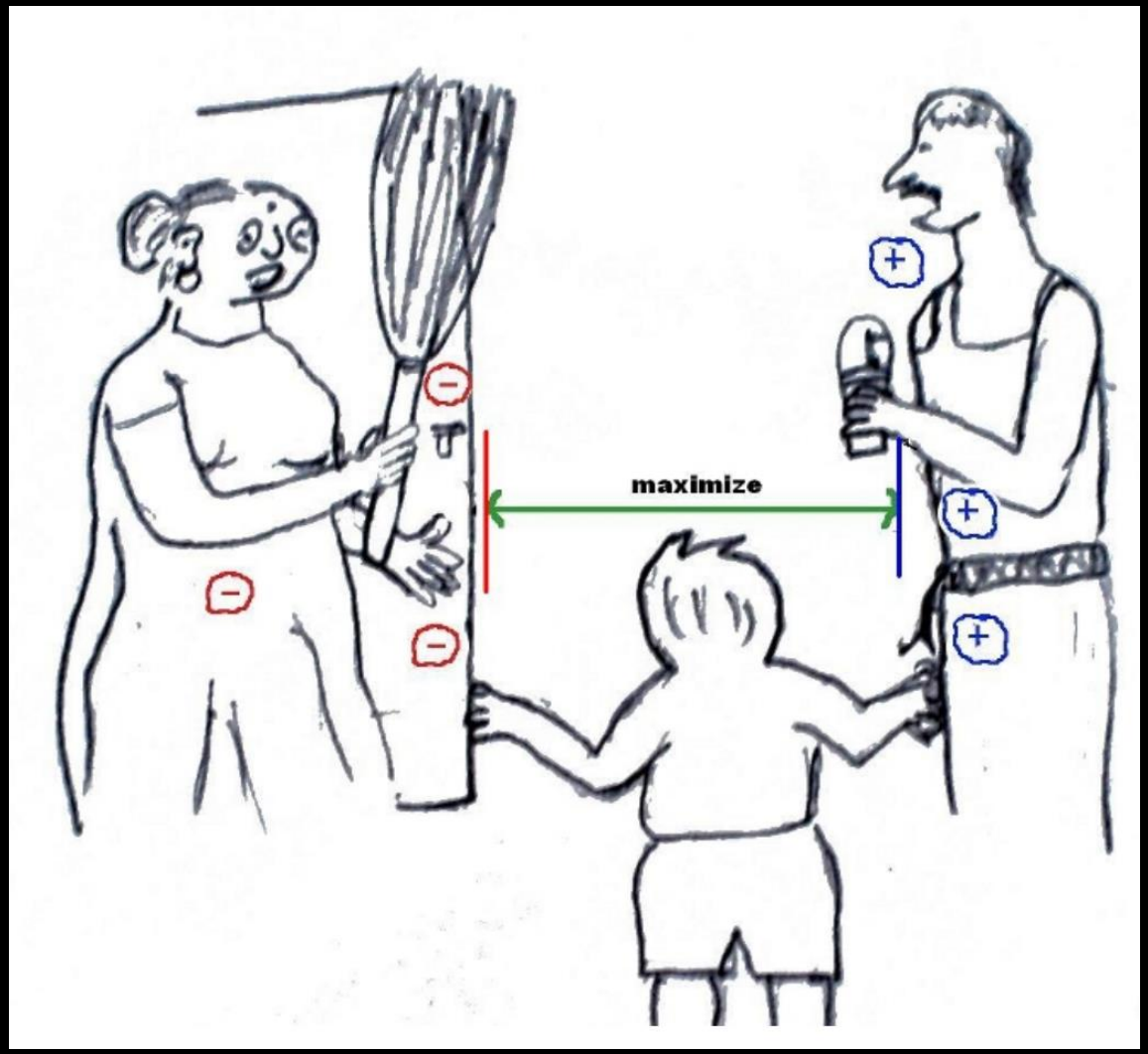
df_x = number of documents containing x

N = total number of documents

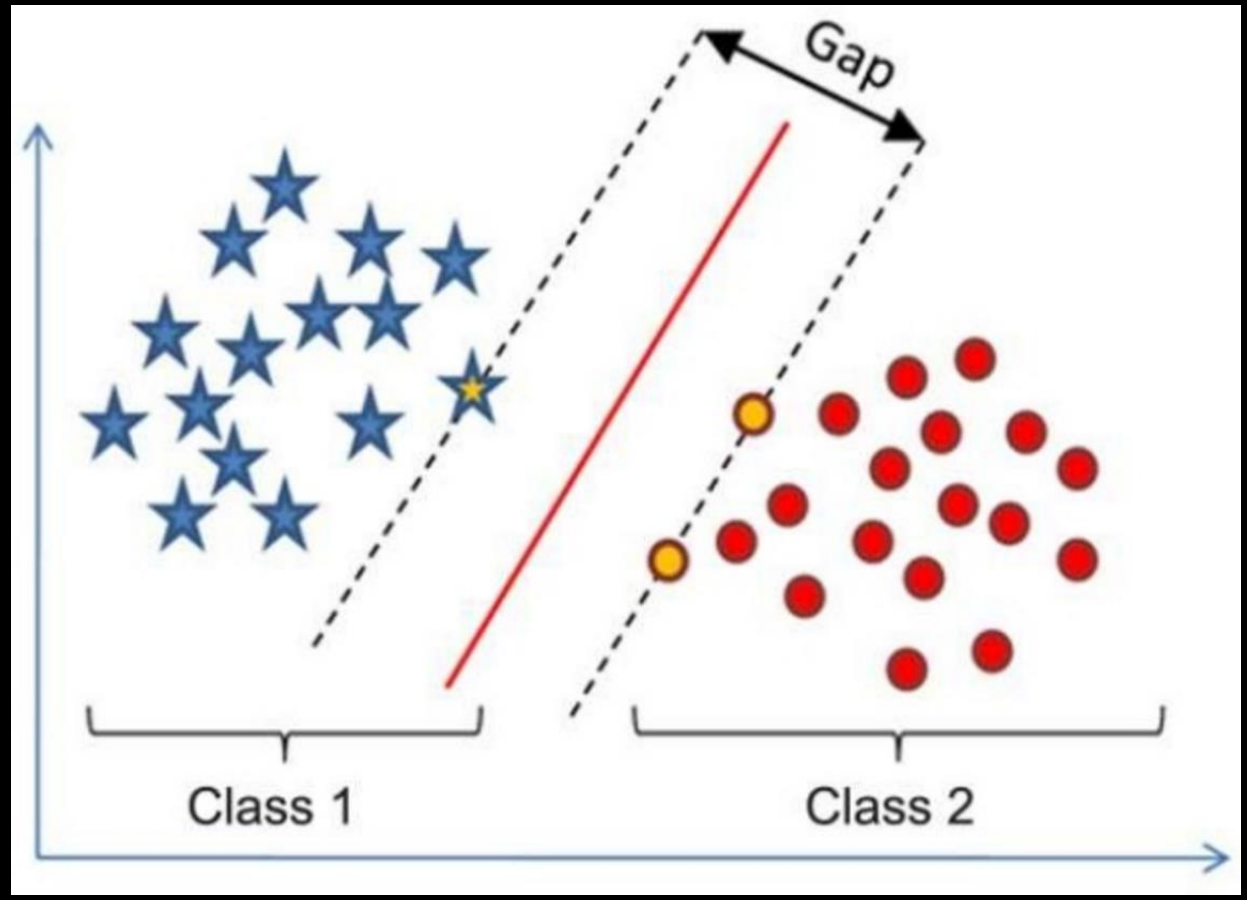
Features


TF-IDF vs Information Gain
Classification accuracy
100
90
80
70
60
50
1
1.5
2
2.5
3
3.5
4
Value of n
TF-IDF
Information gain

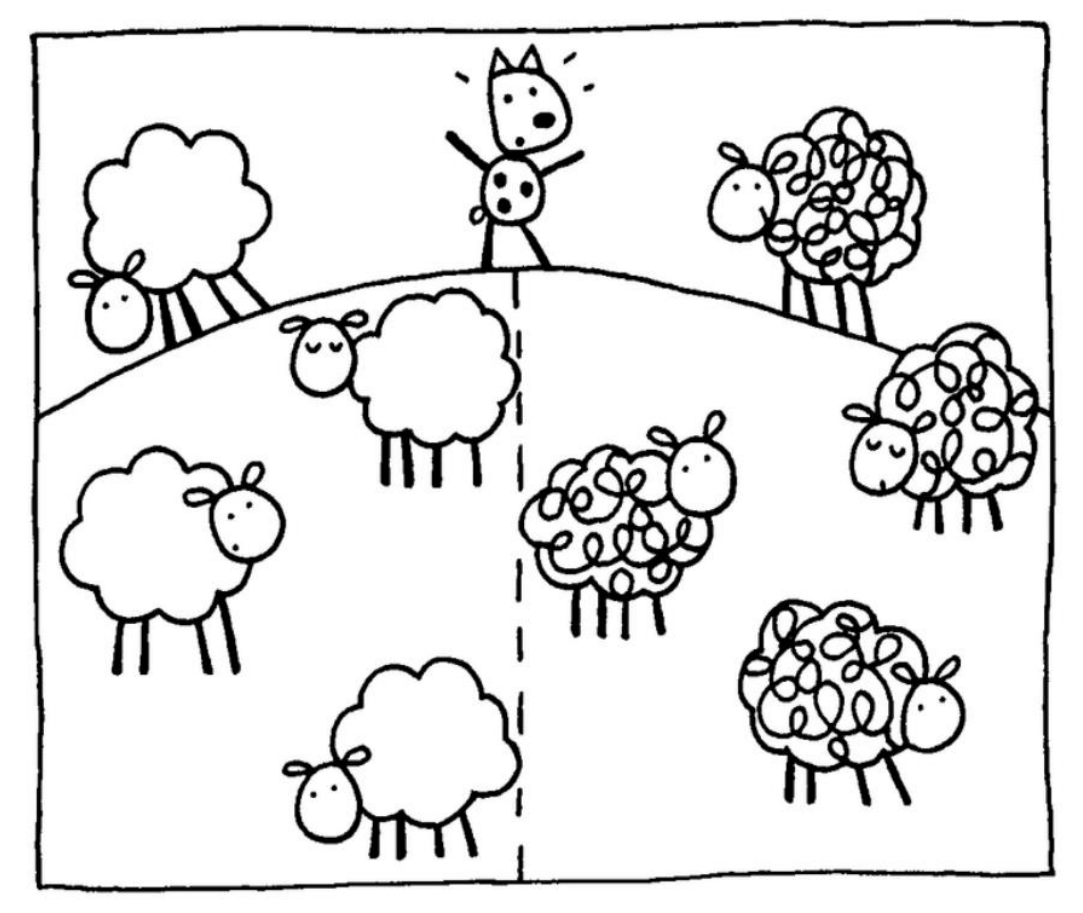
Support Vector Machine



Support Vector Machine

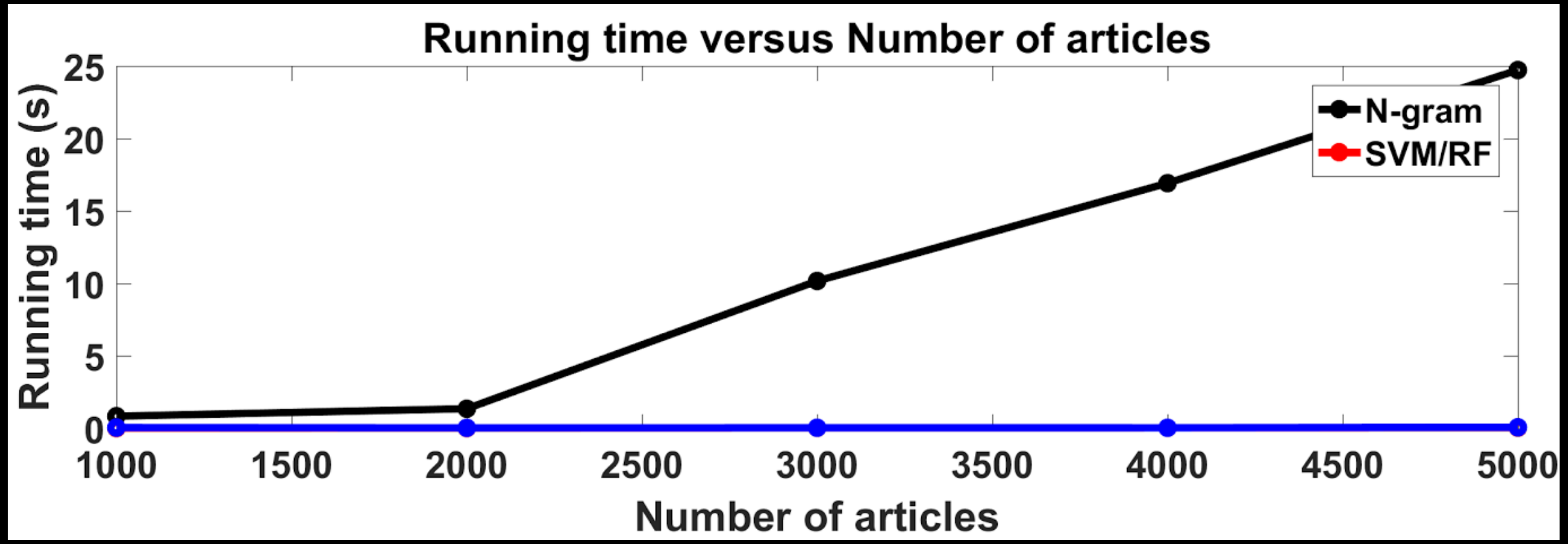


Telemetry: Classification Task

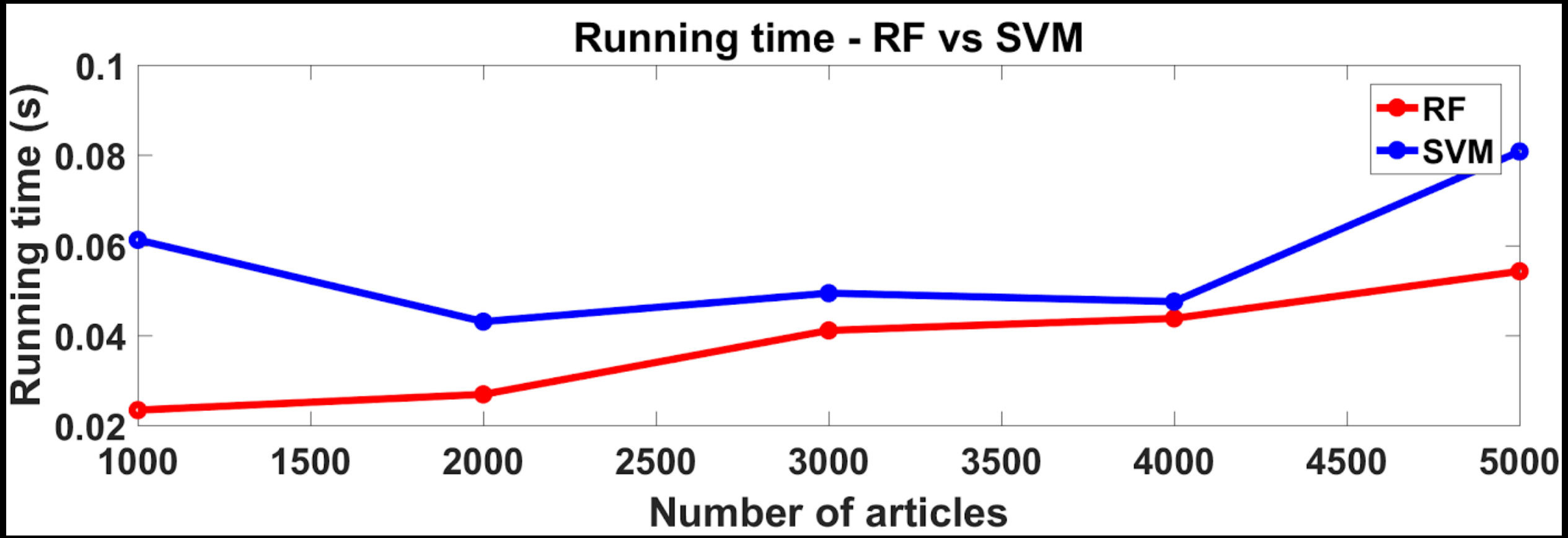


- Evaluate accuracy on manually curated data
- Negatives mined from unrelated StackExchange categories

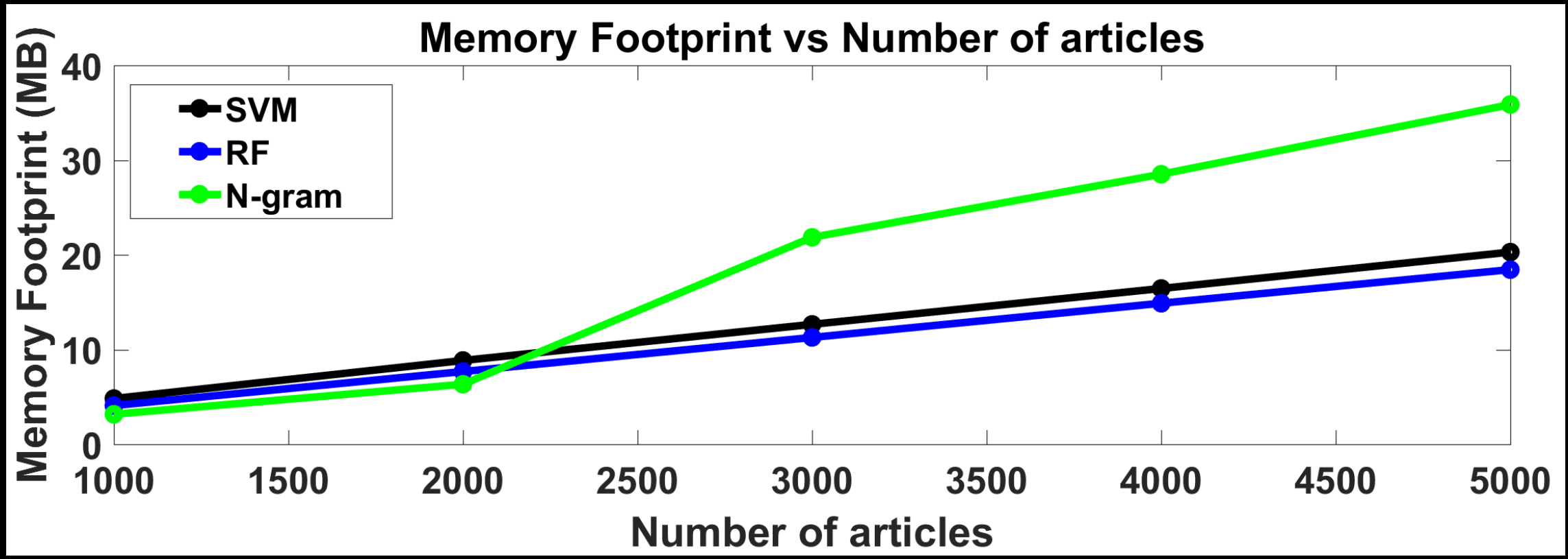
Telemetry: run time



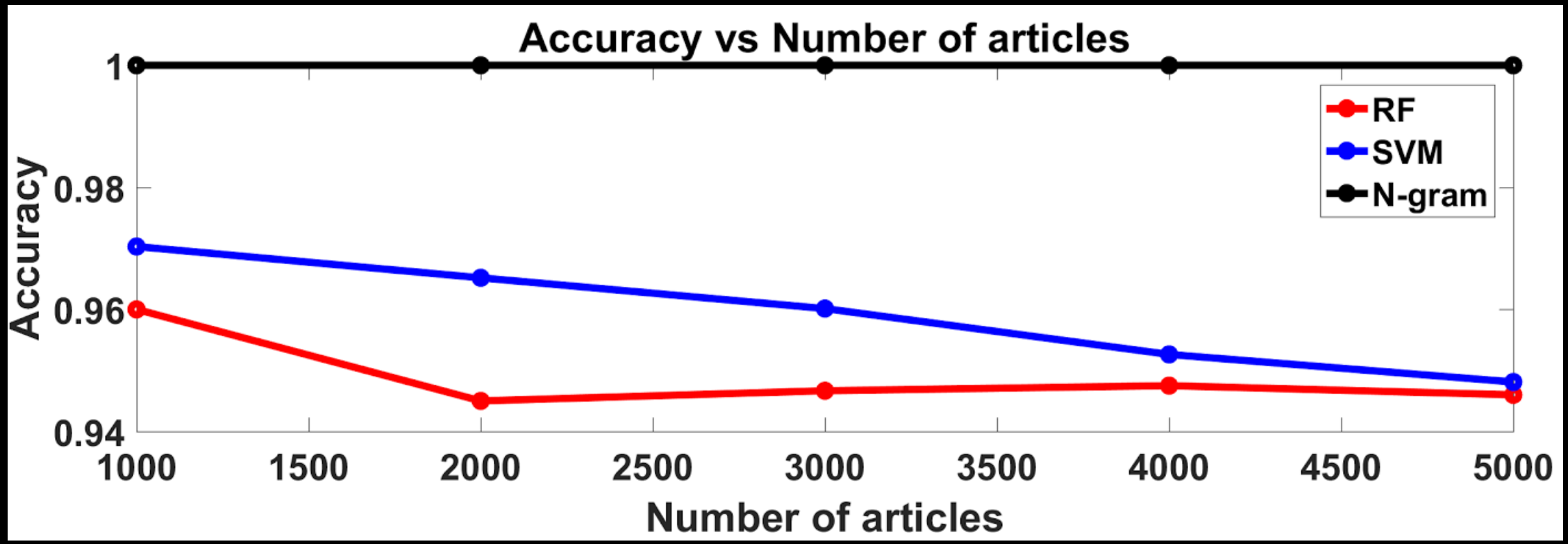
Telemetry: run time



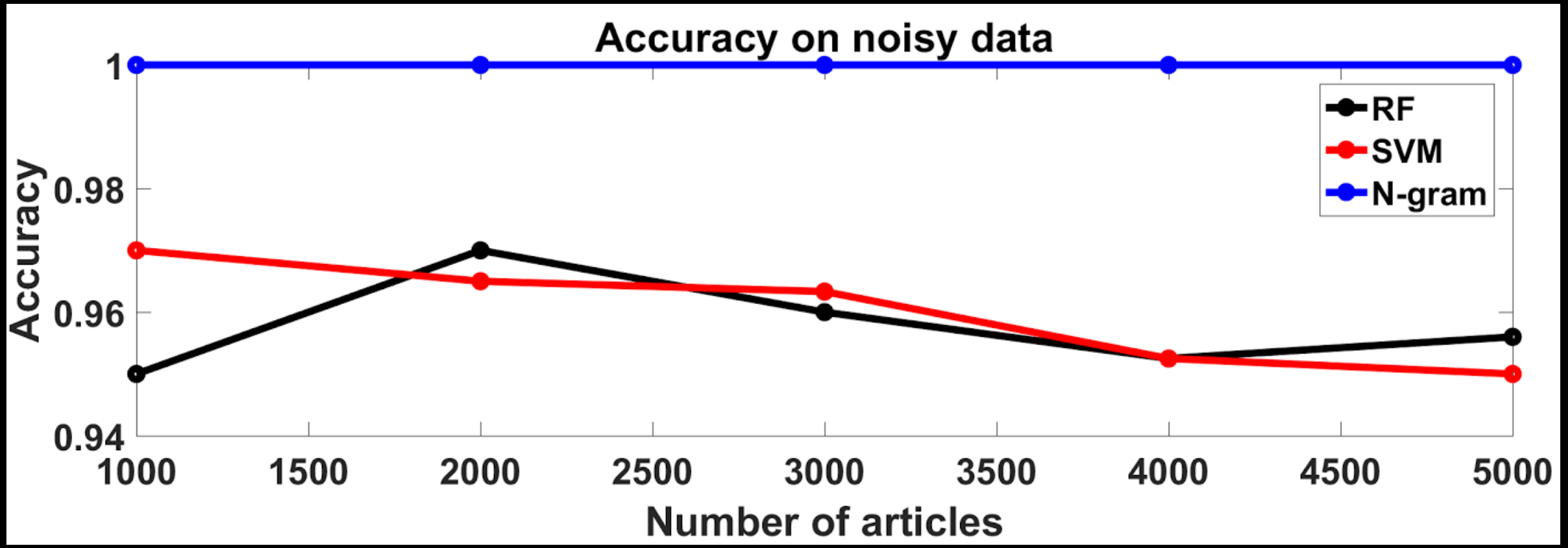
Telemetry: Memory



Telemetry: Accuracy



Telemetry: Accuracy on noisy data

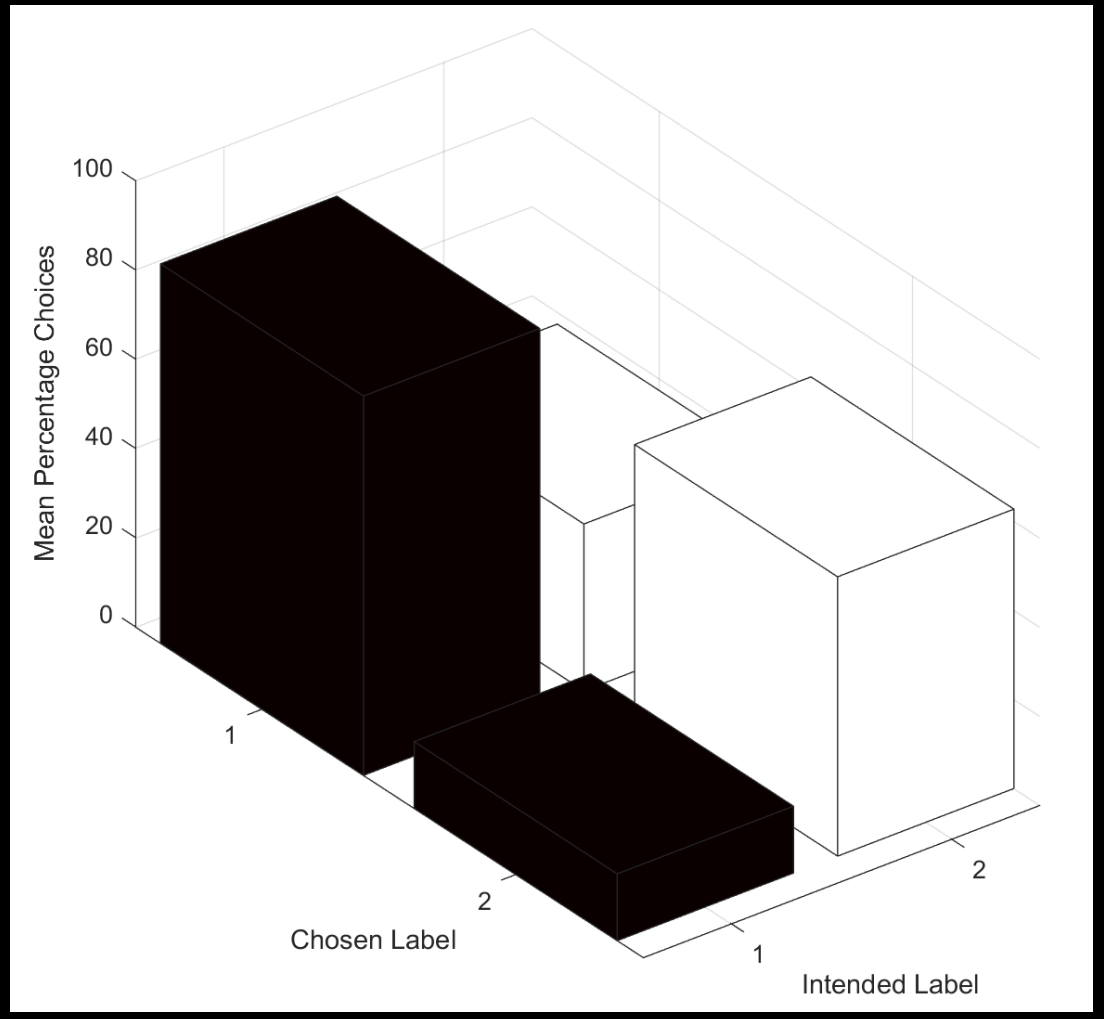


Telemetry: User responses

- User picks 20 positives from a 100 random sample
- Train classifiers on 50% of ground truth data
- Negatives mined from remainder of data set

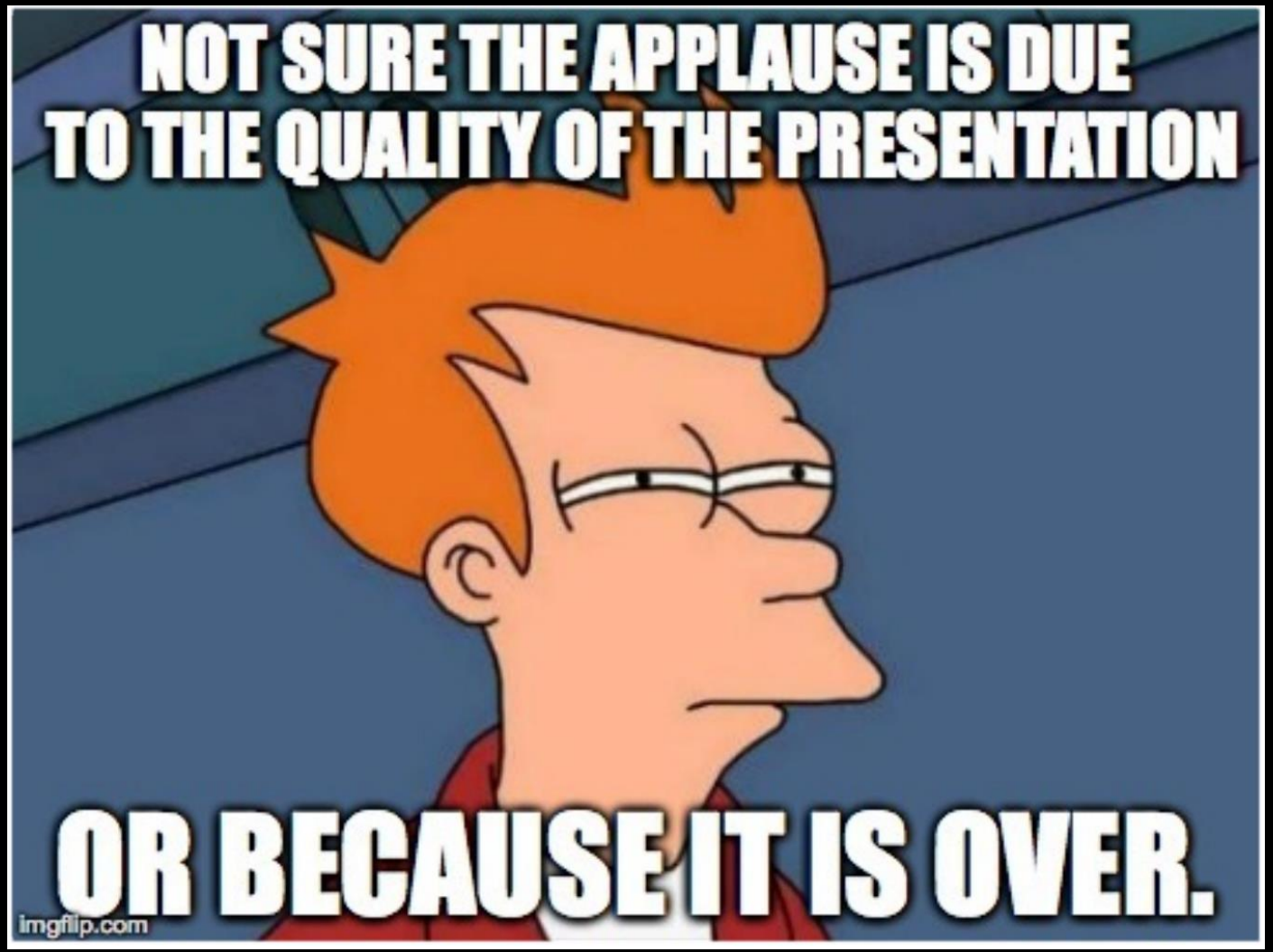
- Average accuracy, RF: 71.6%
- Average accuracy, SVM: 76.27%

Telemetry: User responses



Summary

- N-grams: computationally infeasible, high memory footprint
- RF’s: Most computationally efficient, least memory footprint
- SVM’s: Most accurate on real world data


NOT SURE THE APPLAUSE IS DUE
TO THE QUALITY OF THE PRESENTATION
OR BECAUSE IT IS OVER.
imgflip.com