Søktek-rankedRetrieval!!

Ranked retrival

Relevance (Document similarity, tf-idf)

Et knippe oppgaver

Uoffisiell midtveisevaluering

Live-progge oblig B?

Assignment C aid

Ranked retrieval

- Sorter matches basert på "brahet"
- Brahet = fx hvor bra et document matcher queryen
- Nyttig feature

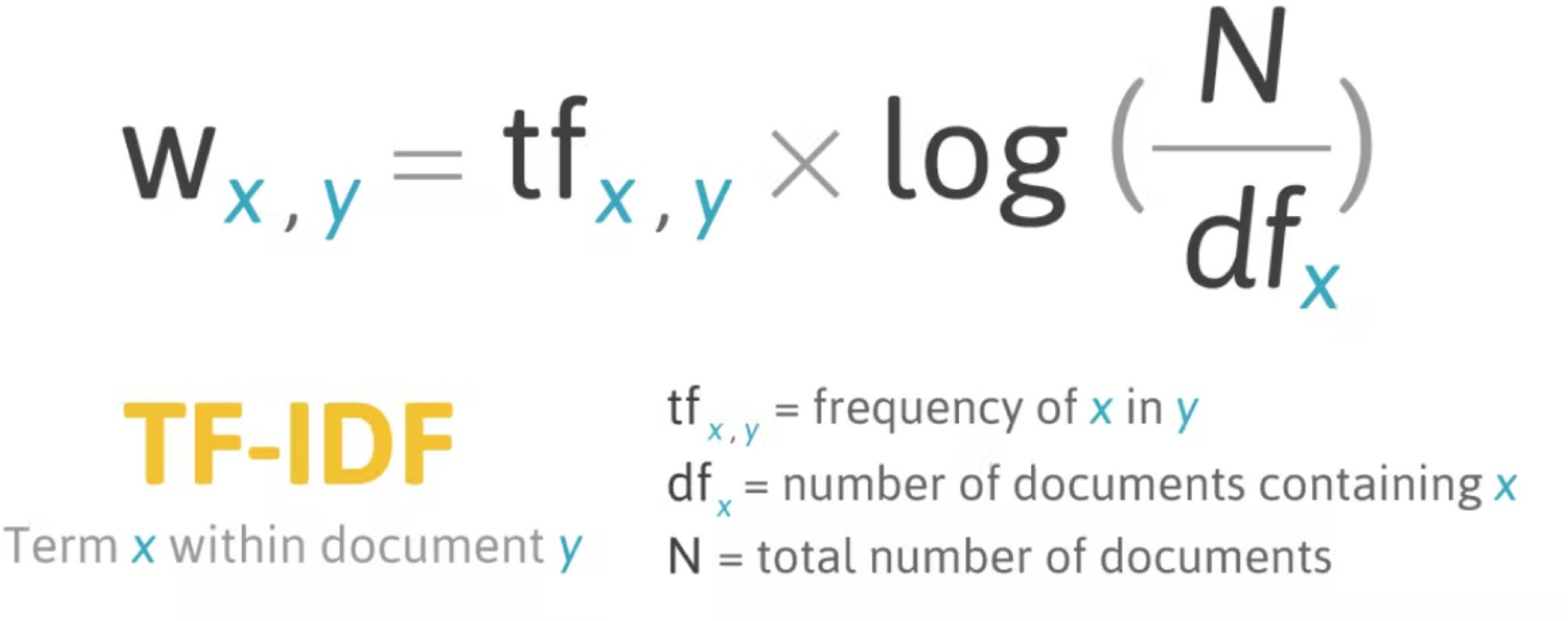
Document similarity (Relevance)

- Tf-idf
- Cosine similarity
- Bullet 3

Tf-idf

- "Term frequecy-inverted document frequency"
- NB: Streken er ikke 'minus'
- Anbefaler: https://ted-mei.medium.com/demystify-tf-idf-in-indexing-and-ranking-5c3ae88c3fa0

Demystify TF-IDF in Indexing and Ranking



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

TF-IDF

Term x within document y

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

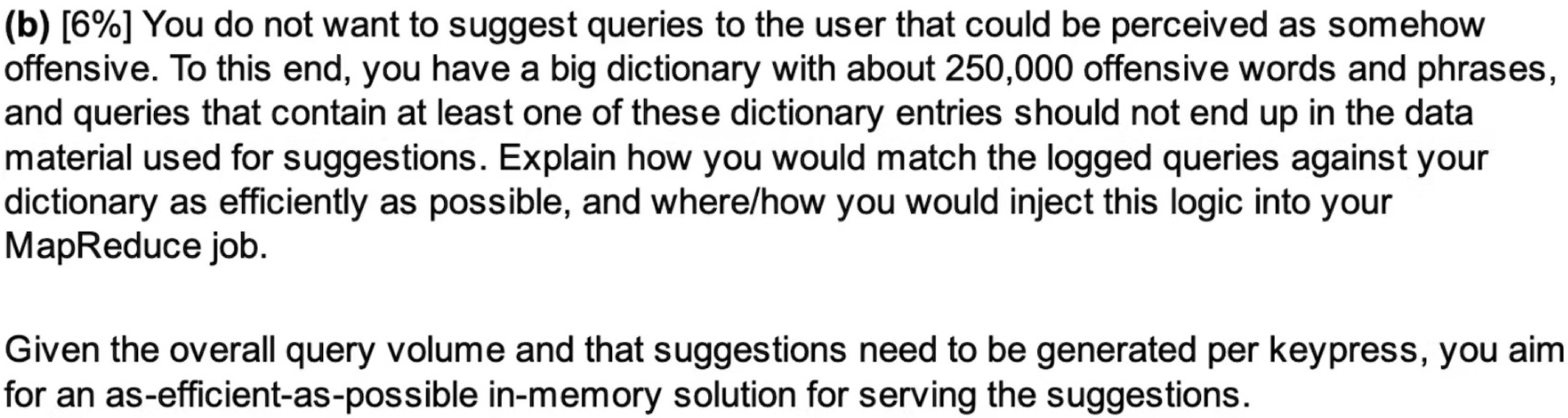
df_x = number of documents containing x

N = total number of documents

(Stjålet fra linken i forrige foil)

Diskusjosoppgave coming up

Om oblig B-pensum, spesielt papers

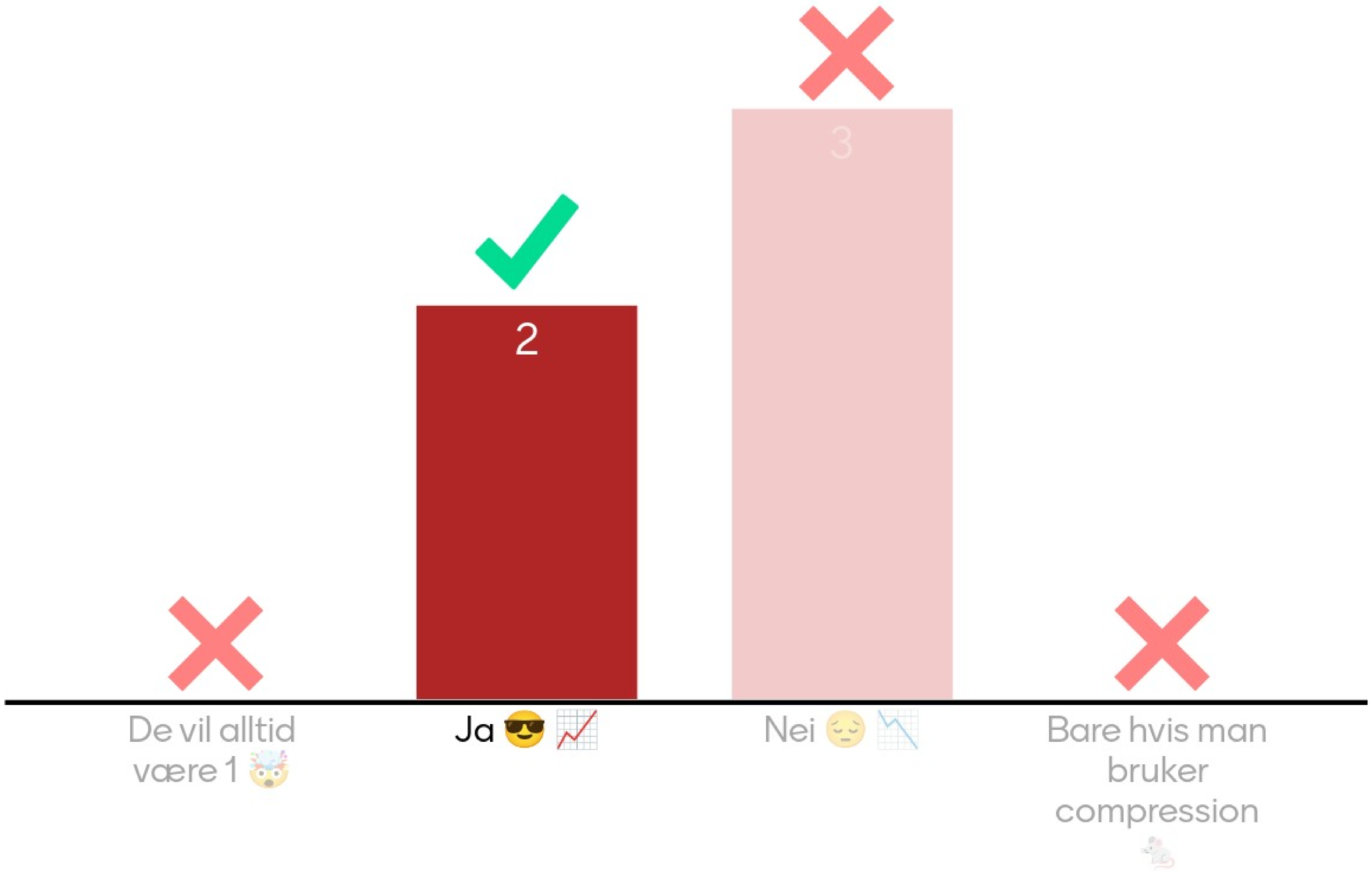


(b) [6%] You do not want to suggest queries to the user that could be perceived as somehow offensive. To this end, you have a big dictionary with about 250,000 offensive words and phrases, and queries that contain at least one of these dictionary entries should not end up in the data material used for suggestions. Explain how you would match the logged queries against your dictionary as efficiently as possible, and where/how you would inject this logic into your MapReduce job.

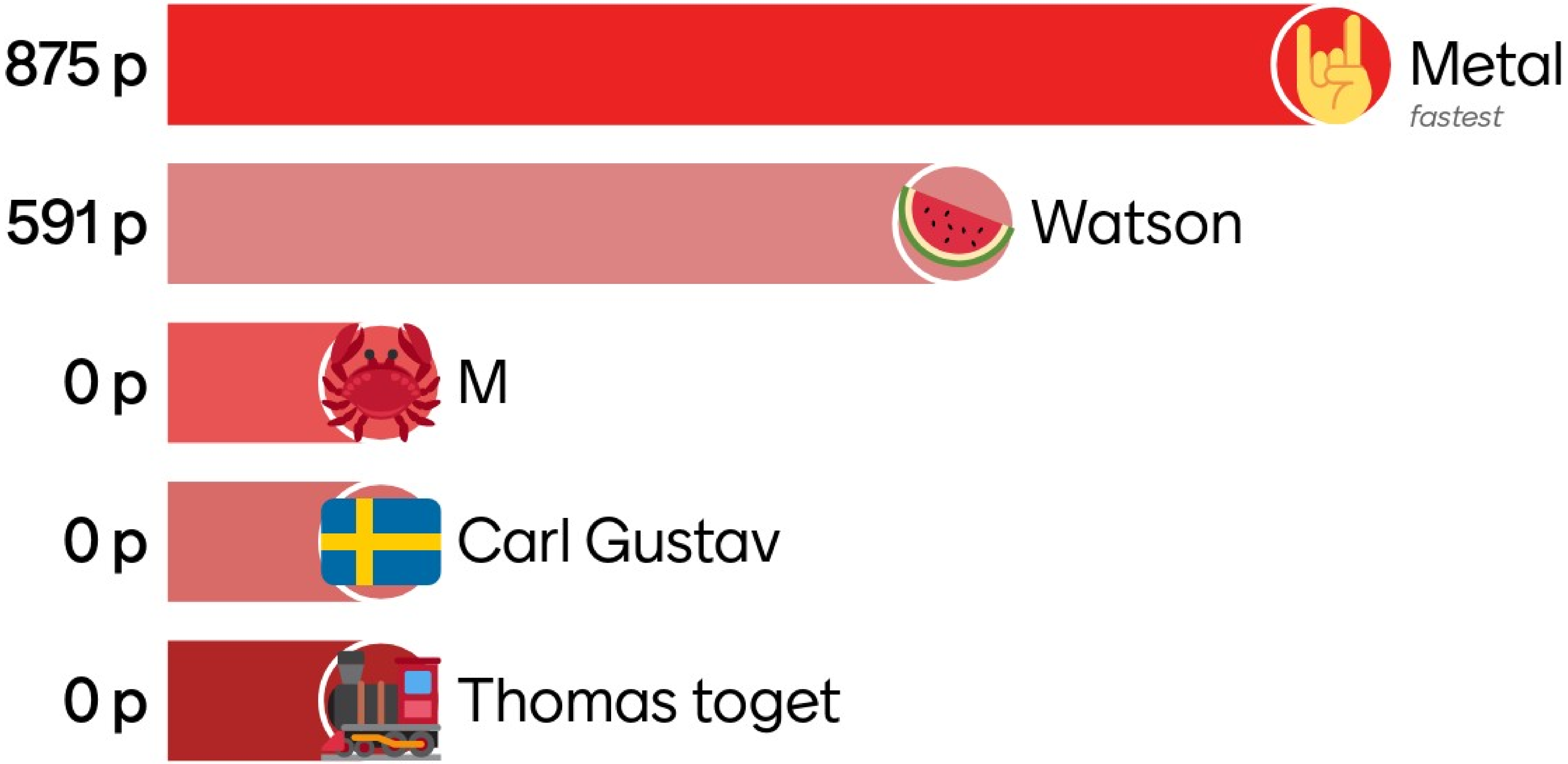
Given the overall query volume and that suggestions need to be generated per keypress, you aim for an as-efficient-as-possible in-memory solution for serving the suggestions.

Noen tanker om hvordan man kan få til dette? (Fra eksamen 2021)

Kan en tf-idf-verdi overstige 1? (Eks 2020)



Leaderboard



Om en tf-idf-verdi kan overstige 1:

- Ja!
- Log-funksjonen vokser monotont til infinity
- Kan også justifyes med et eksempel

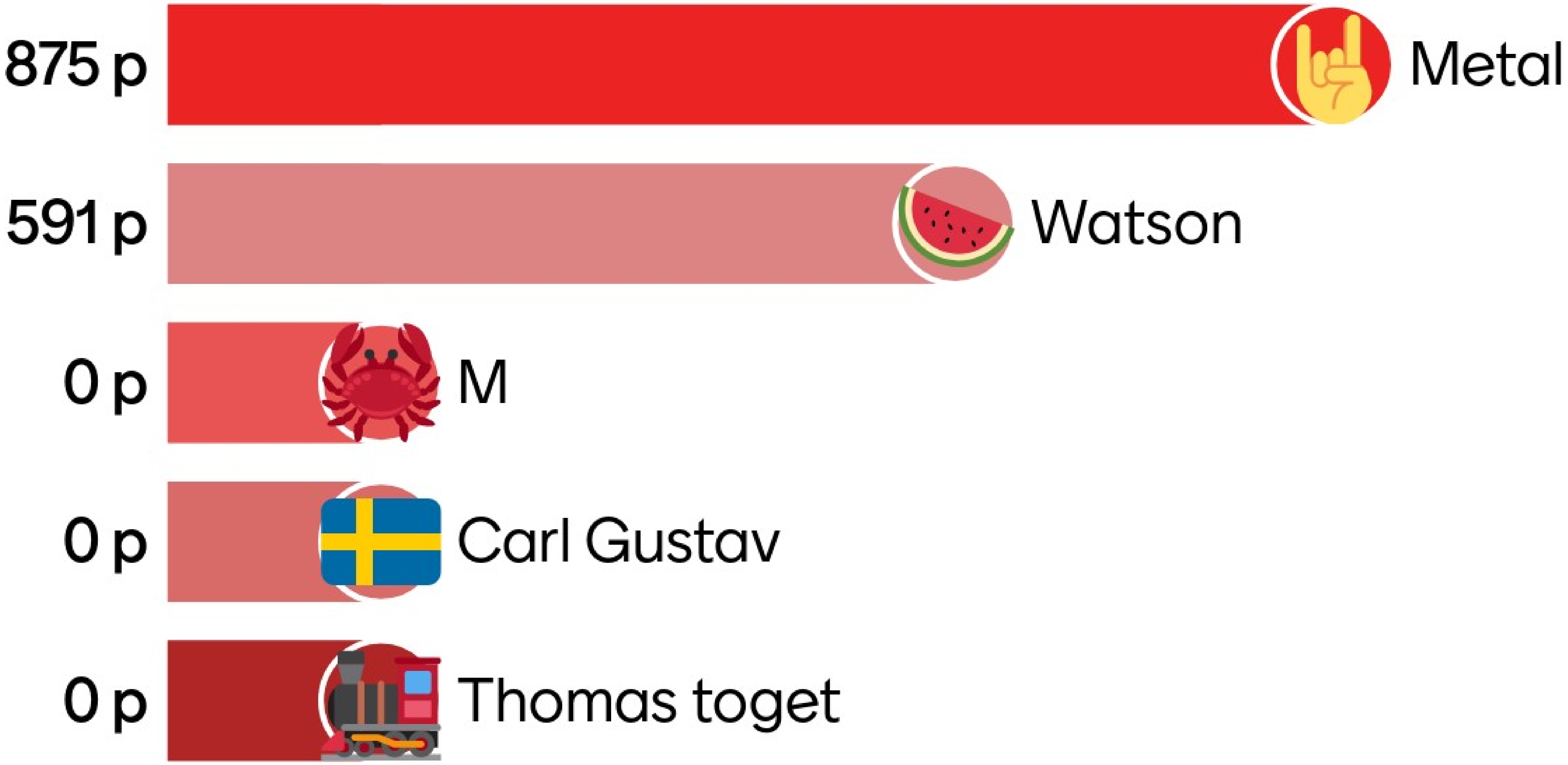
Describe the idea behind Simple9-encoding. (Eks 2018)

9 forskjellige beskrivelser for noe encoding i bitsene ✕

ㄟ□☯□☯✿□ㄏ ✕

The correct answer is: Control bits

Leaderboard



(ii) Pack multiple integers into a 32-bit word. Then, 4 control bits tell you how to interpret the remaining 28 bits: As one 28-bit number, or as two 14-bit numbers, and so on. (In all, 9 different ways to interpret and break up those 28 bits, sometimes with wasted bits.)

Uoffisiell midtveisevaluering

- Jeg vil holde bedre gruppetimer 😎
- 100% anonymt og ikke endorsed av ifi/FUI
- https://nettskjema.no/a/soketek-midterm-evaluation

Live-progge oblig B?

Kommer an på hvem som er i gruppetimen 🤪

Oblighjelp / 1-1-spm

Lurt å ha sett på oblig C før dagens forelesning