Spelling Correction and Autocompletion for Mobile Devices

Ziang Lu

Bachelor's Thesis

Computer Science Department
University of Freiburg

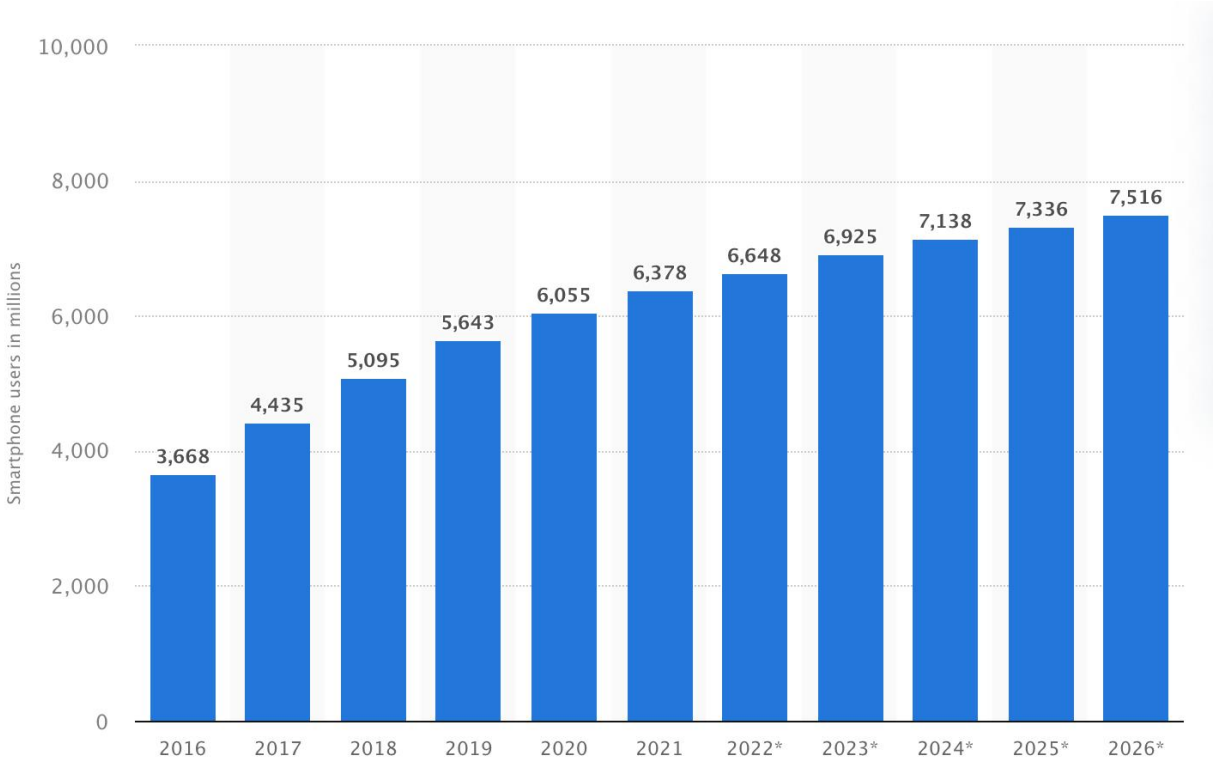
Examiner: Prof. Dr. Hannah Bast
Adviser: Matthias Hertel

UNI FREIBURG

Motivation

Number of smartphone users from 2016 to 2026

(in billions)



(Statista, 2021)

Tasks

we are going to watch a movie

- correct the next word

  current input: *Wee (expecting correction)*

  candidates to choose: `We Lee Bee`

- predict the next word

  current input: *We (expecting prediction)*

  candidates to choose: `are do were`

- complete the next word

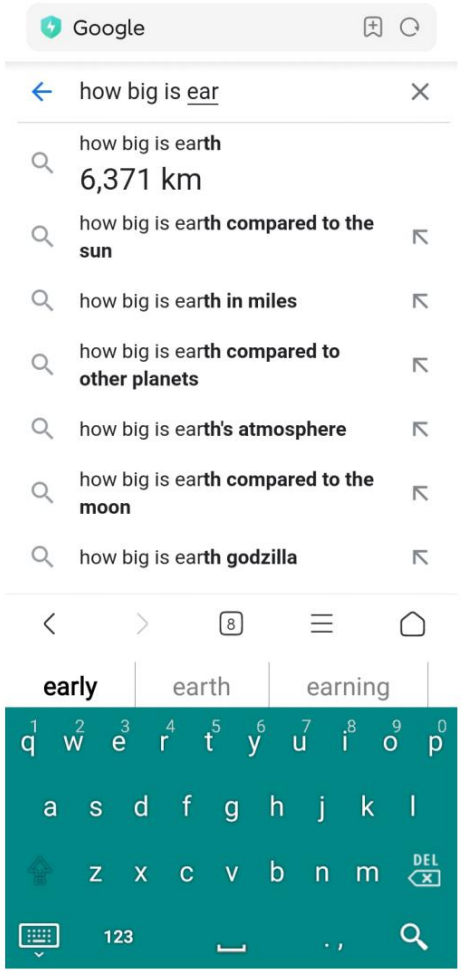
  current input: *We are g (expecting completion)*

  candidates to choose: `going gone getting`

Implementation and usage

Implementation as an Android System Keyboard

- choose candidates from vocabulary according to simalarity
- rank candidates and show the best three candidates to the user



Prefix Edit Distance

We input always from the start of a word

Edit Distance (som, something) = 6

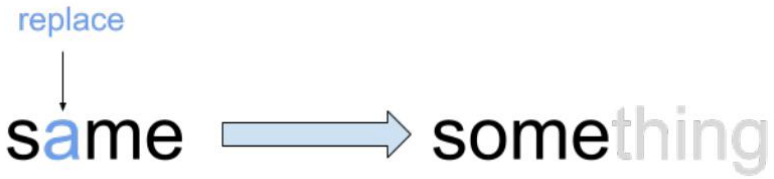
Prefix Edit Distance (PED):

$$PED(x, y) = min_{y'}(ED(x, y'))$$

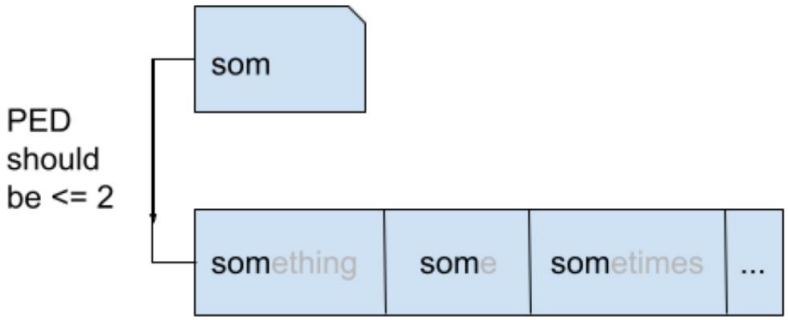
where y' is a prefix of y

PED

PED(same, something) = ?

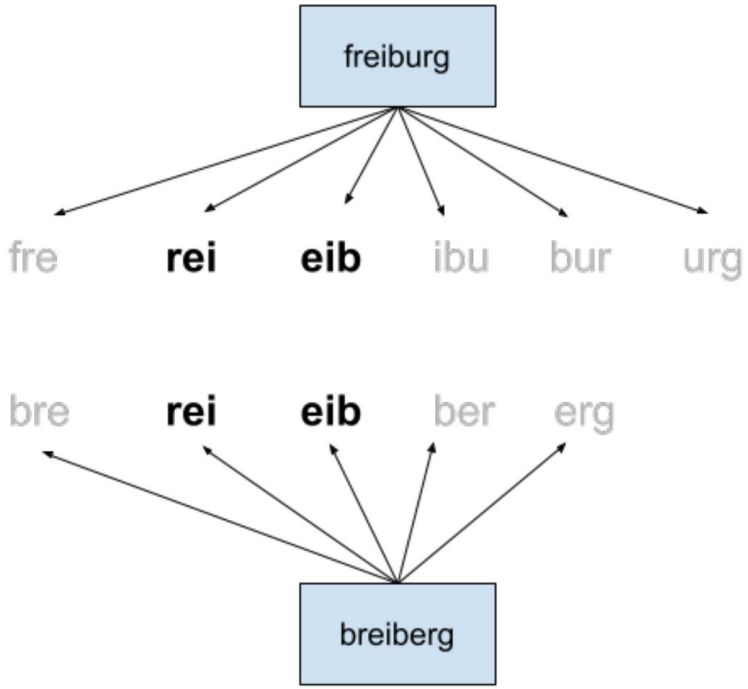


Given a threshold of **P**ED and prefix “som”



Q-gram Index

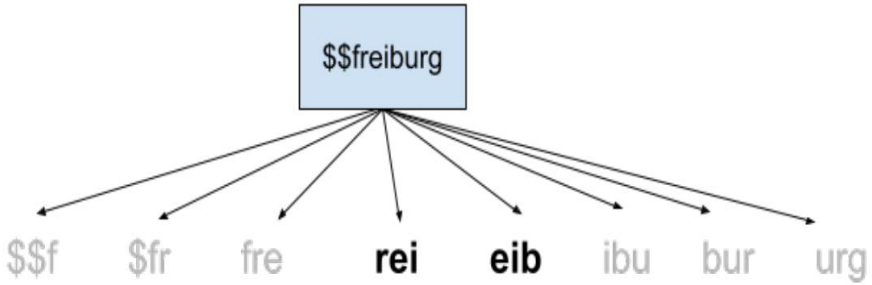
Filter impossible candidates to accelerate response



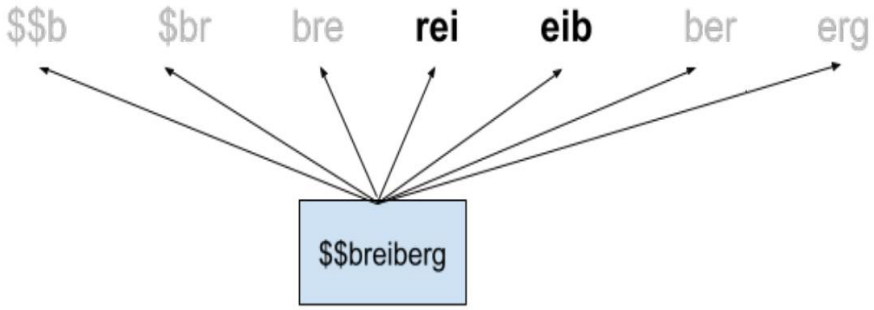
Q-gram

Some words are very short like “I”, “am”, “a”

Pad special symbols: “$$I”, “$$am” ...



$$comm(x', y') \geq |x| - q \cdot \delta$$



x' = $$freiburg
y' = $$breiberg

$$comm(x', y') = 2 \geq |8| - 3 \cdot 2.$$

N-gram model

I want to drink w___

sequence of n words

Example:

h: have you seen my

w: book

$$P(\text{book}|\text{my}) = \frac{C(\text{my book})}{C(\text{my})}$$ based on bigram

$$P(\text{book}|\text{seen my}) = \frac{C(\text{seen my book})}{C(\text{seen my})}$$ based on trigram

N-gram probability

Extend the formula to N-gram:

$$P(w_n|w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1} w_n)}{C(w_{n-N+1}^{n-1})}$$

(n represents the length of a whole sentence. w_J^I represents a sequence which consists of $w_I, w_{I+1}, w_{I+2}, ..., w_J$.)

Combination of different N-gram probabilities:

$$P(\text{N-gram}) = \sum_{i=1}^{N} P(w_n|w_{n-N+1}^{n-1}) \cdot \lambda_i$$

POS-tags

For absence of some combinations:

I	have	a	cat
PRON	VERB	DET	NOUN

Example:

“she sleeps” never appears in the corpus

candidates: sleeping **sleeps** sleep

Corpus and Vocabulary

9000 words from MIT word list + 1000 most frequent words from corpus

14 million words make up of the corpus, which was retrived from web texts.

Data storage

- avoid reading data on startup => store n-grams and q-gram index in a **database**
- load data only on demand
- lower delay of startup and low memory footprint

resume	2185	my resume	32	my	331930	null	null	null	null
resume	2185	your resume	228	your	597852	null	null	null	null
click	24403	the click	143	the	6719158	null	null	null	null
click	24403	can click	407	can	475829	you can click	246	you can	138502

Methods and test set

Testset: 5% of corpus and Enron Emails

Criterion of evaluation: number of **saved keystrokes**(%)

Example:

sentence: Let's have a drink.

length(including space): 19
keystrokes needed with keyboard: 10

$$\frac{19-10}{19} \approx 47.37\%$$

Results and analysis

Performance based on different grams

Gram	Saved keystrokes(%)
bigram	39.30%
trigram	39.97%
quadrigram	39.98%
quinquegram	39.98%

Test for the POS-Tags

Evaluate the performance with the attendance of POS-Tags

Mode	Saved keystrokes(%)
Without POS-Tags	39.31%
With POS-Tags	39.98%
+	0.67%

She sleeps very well and he walks into the forest

Test for the correction

The first letter of every word will be changed

100 sentences from Enron Emails

Keyboard	No mistake	With mistakes
Gboard	48.81%	19.84%
Nboard	45.64%	21.57%

Current problems and Future work

Problems:

- POS-tagger was not precise
- Paied little attention to more parts of a sentence

TODOs:

- customized and preciser POS-tagger
- more complicated language model => Transformer

Conclusion

- Theories and their application
- Implementation of the keyboard
- Evaluation

Thanks

Any other questions?