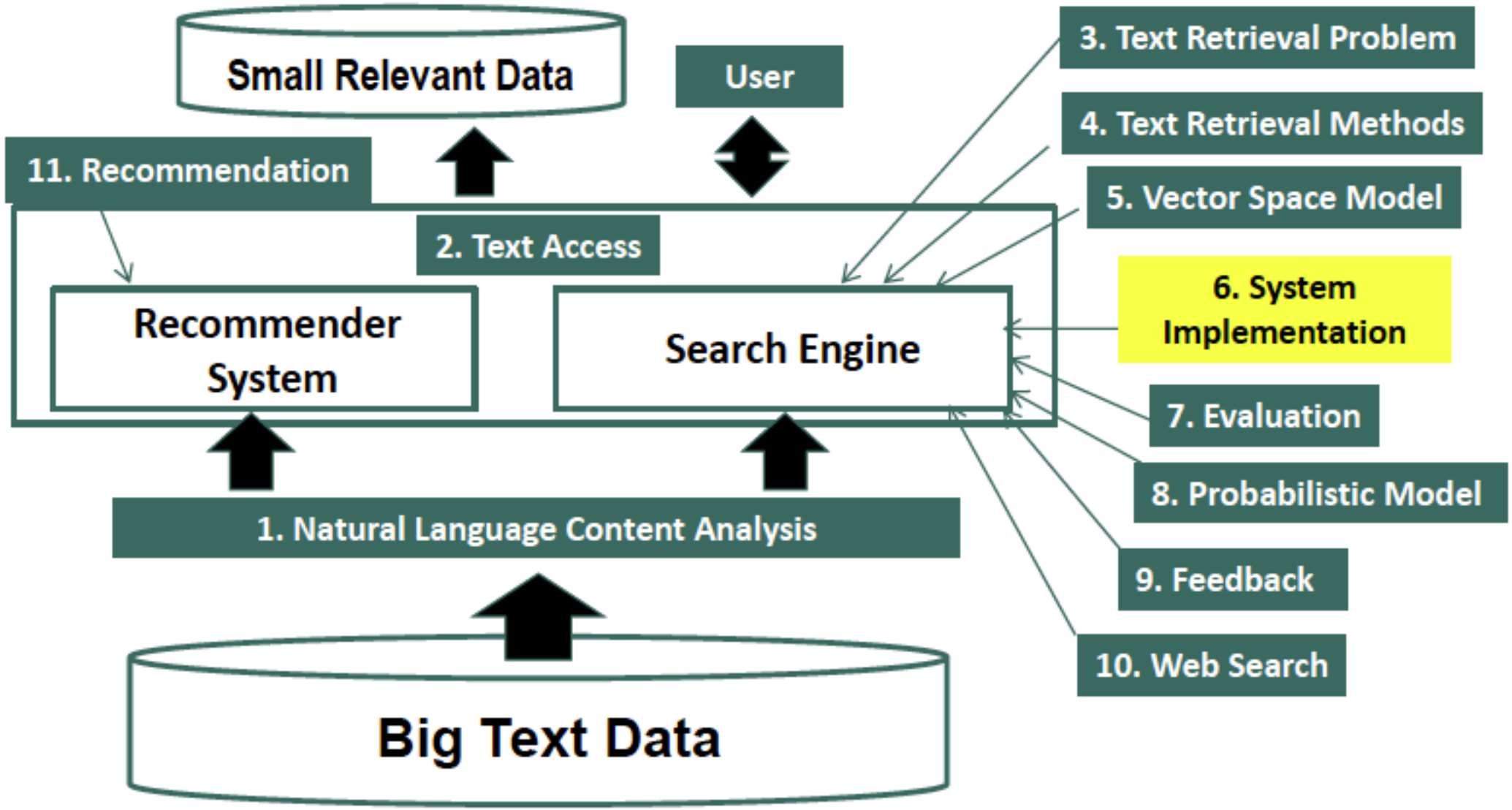
Information Retrieval

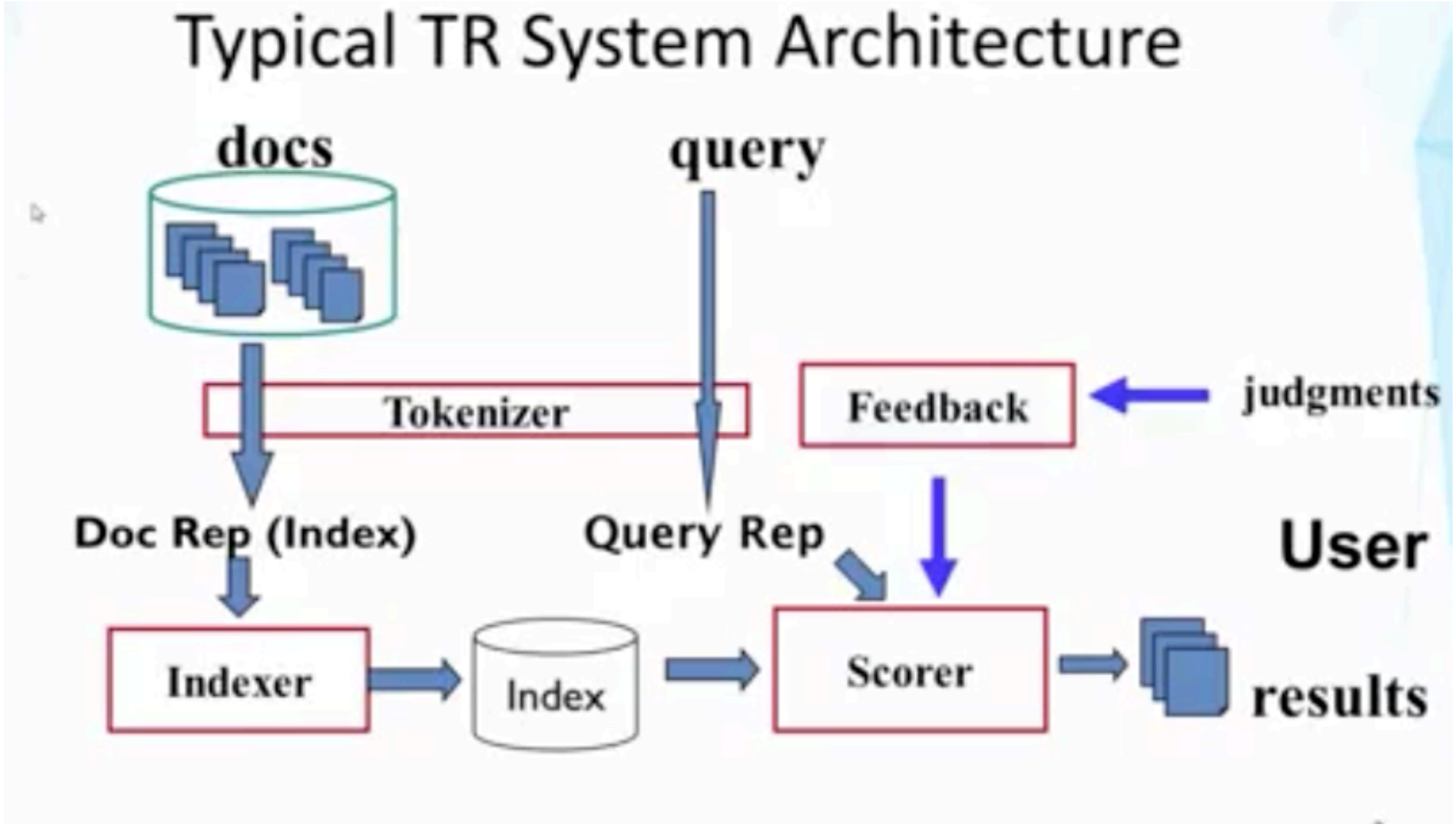
Implementation of Text Retrieval System

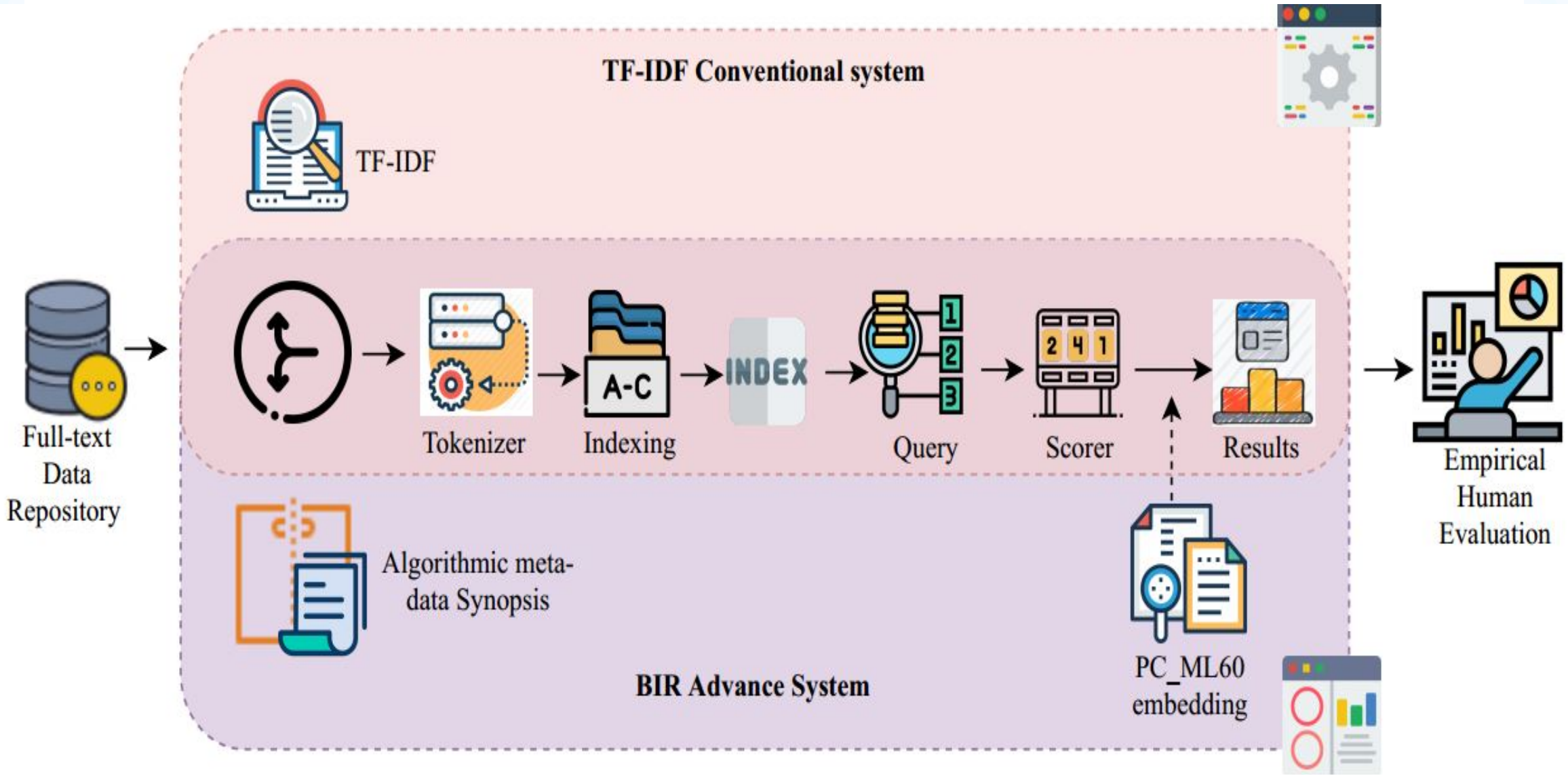
Dr. Iqra Safder

Implementation of Text Retrieval Systems

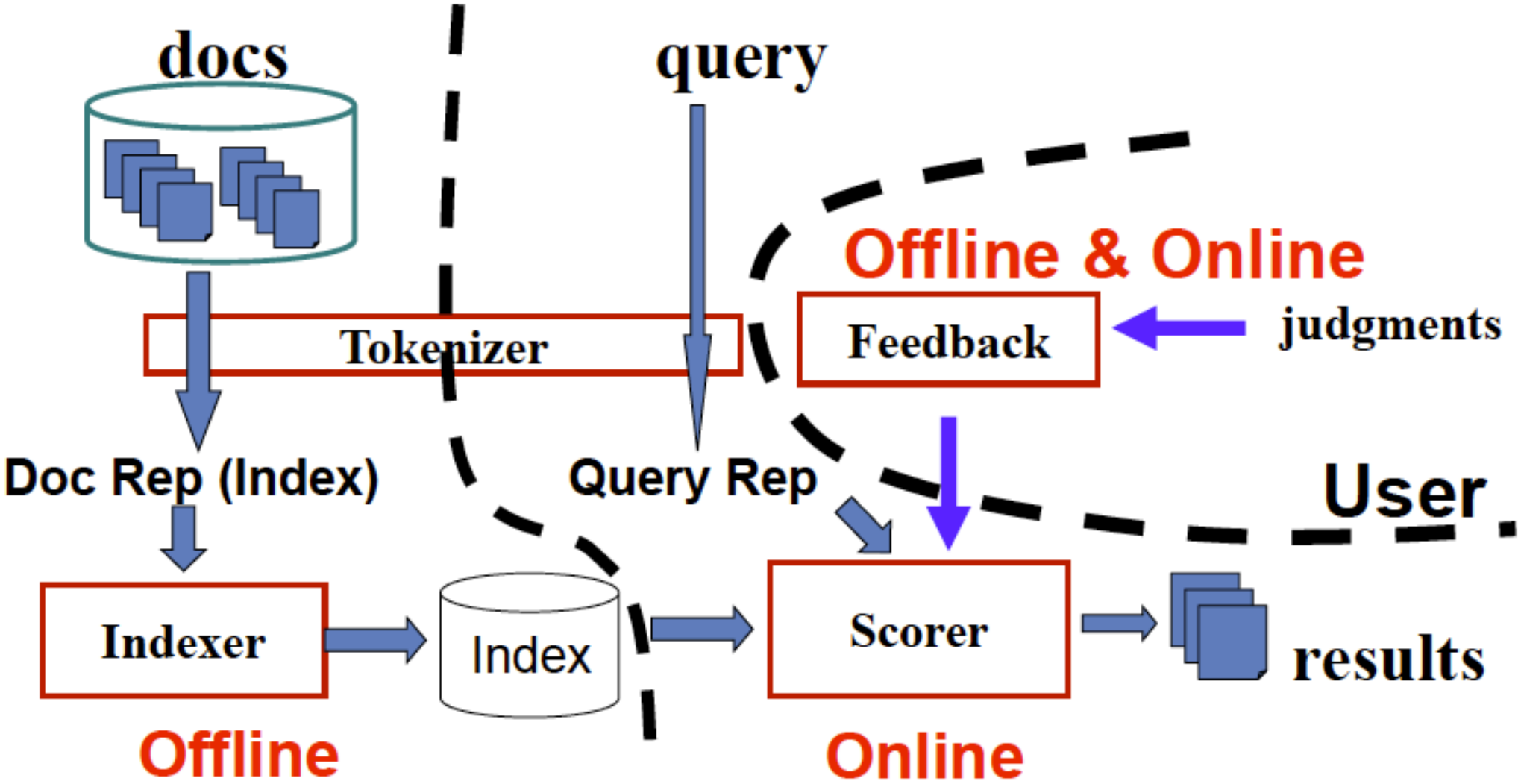


Typical TR System Architecture




TF-IDF Conventional system
TF-IDF
Full-text
Data
Repository
Tokenizer
Indexing
A-C
INDEX
Query
Scorer
Results
Empirical
Human
Evaluation
Algorithmic meta-
data Synopsis
PC_ML60
embedding
BIR Advance System

Typical TR System Architecture



Tokenization

- Normalize lexical units: Words with similar meanings should be mapped to the same indexing term
- Stemming: Mapping all inflectional forms of words to the same root form, e.g.
  - computer -> compute
  - computation -> compute
  - computing -> compute
- Some languages (e.g., Chinese) pose challenges in word segmentation

Lematization VS Stemming

Indexing

- Indexing = Convert documents to data structures that enable fast search (precomputing as much as we can)
- Inverted index is the dominating indexing method for supporting basic search algorithms
- Other indices (e.g., document index) may be needed for feedback

How you can preprocess the data so that you can quickly response to the query with just one term?

Inverted Index Example

doc 1

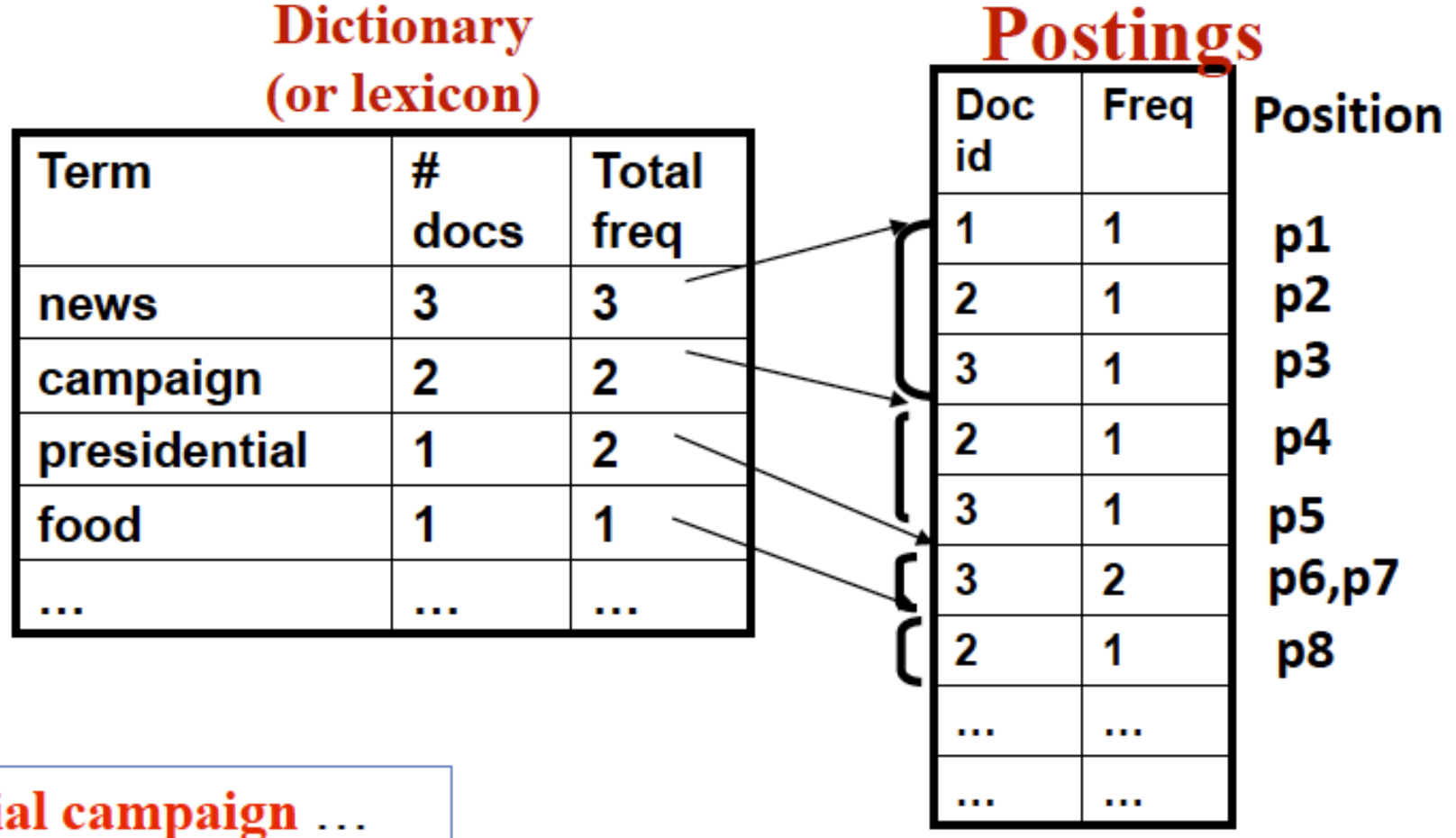
… **news about**

doc 2

… **news about** organic food **campaign**…

doc 3

… **news** of **presidential campaign** …
… **presidential** candidate …



IDF is calculated using # docs

Inverted Index for Fast Search

- Single-term query?
- Multi-term Boolean query?
  - Must match term “A” AND term “B” Conjunctive
  - Must match term “A” OR term “B” Disjunctive
- Multi-term keyword query
  - Similar to disjunctive Boolean query (“A” OR “B”)
  - Aggregate term weights
- More efficient than sequentially scanning docs (why?)

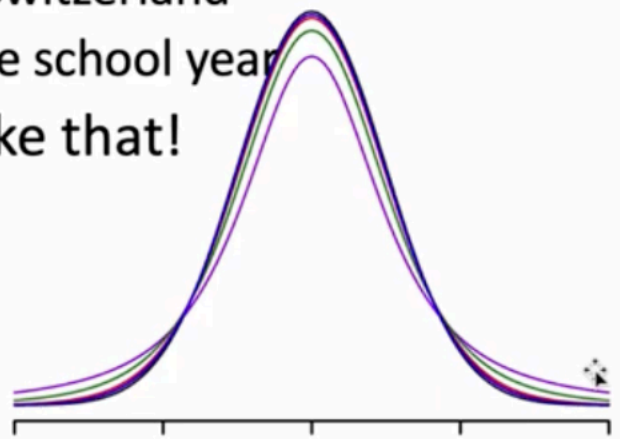
Corpus Linguistics

Do most words in a corpus occur with average frequency?

No.

G.K. Zipf

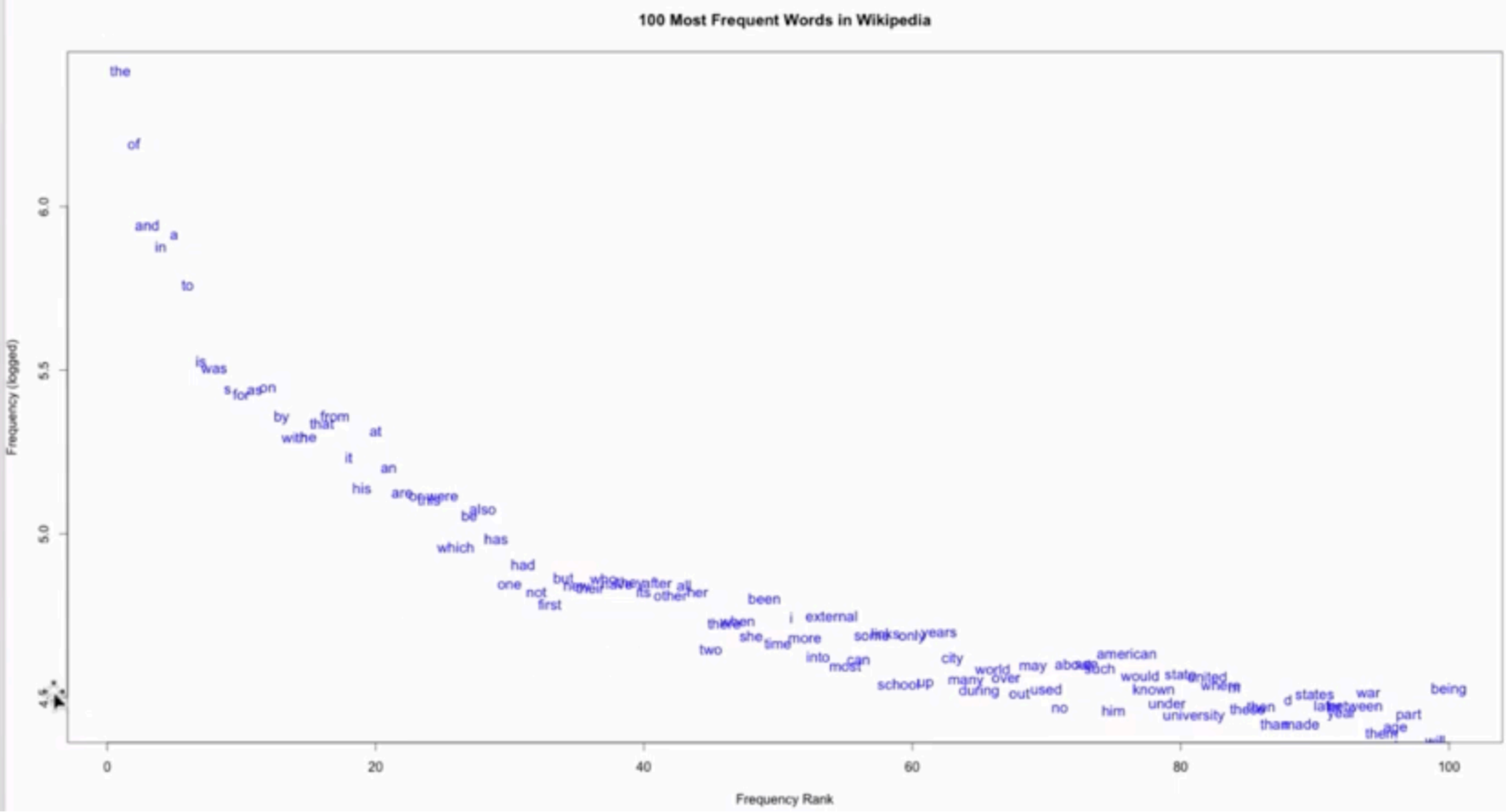
- Many phenomena in the world follow a “normal distribution”: Very few observations with extremely low values, many observations with average values, very few observations with extremely high values.
- Examples:
  - rolling three dice at a time
  - life expectancy of people in Switzerland
  - height of children in the same school year
- Word frequencies are not like that!

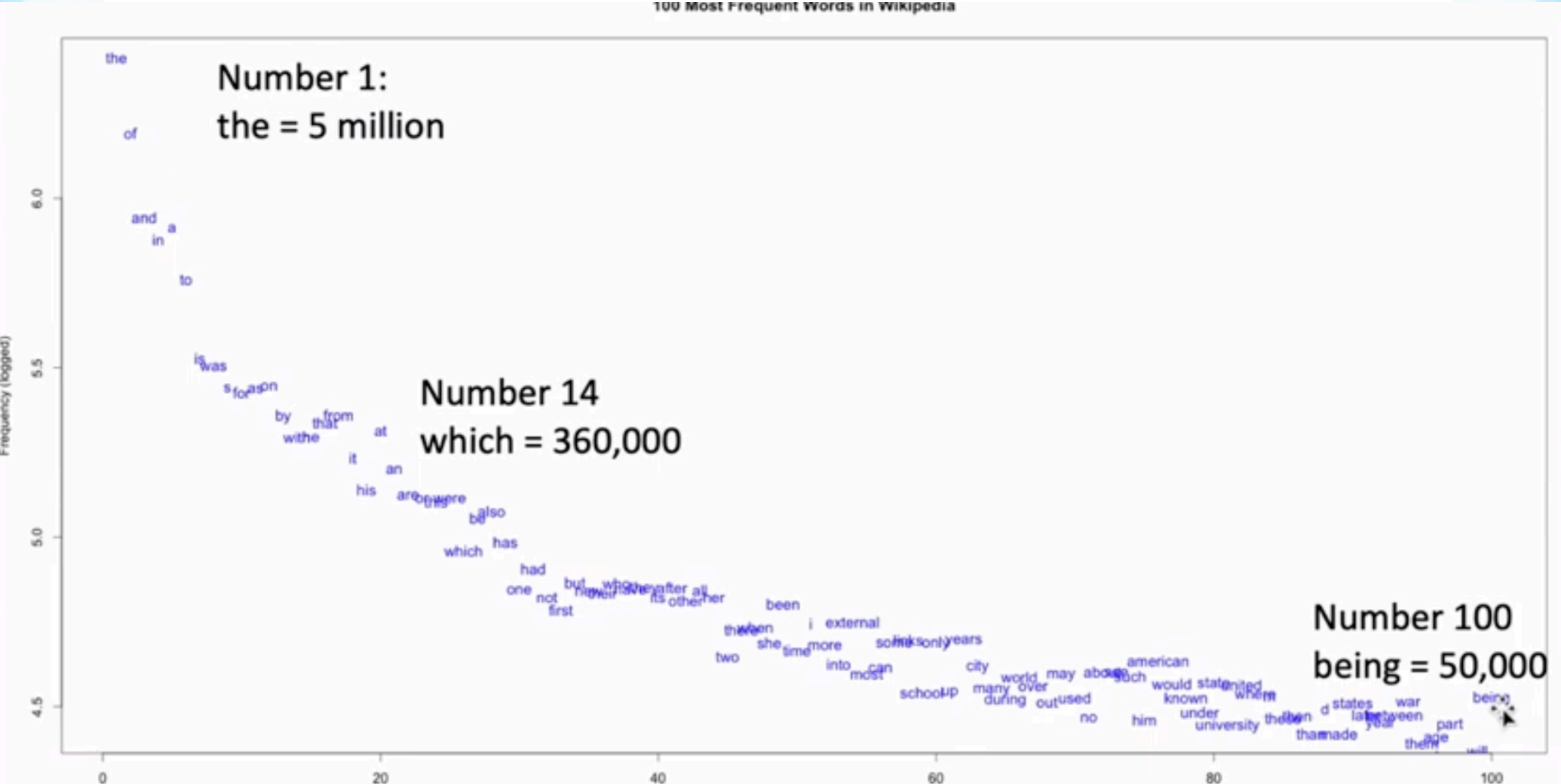


Empirical Distribution of Words

- There are stable language-independent patterns in how people use natural languages
- A few words occur very frequently; most occur rarely. E.g., in news articles,

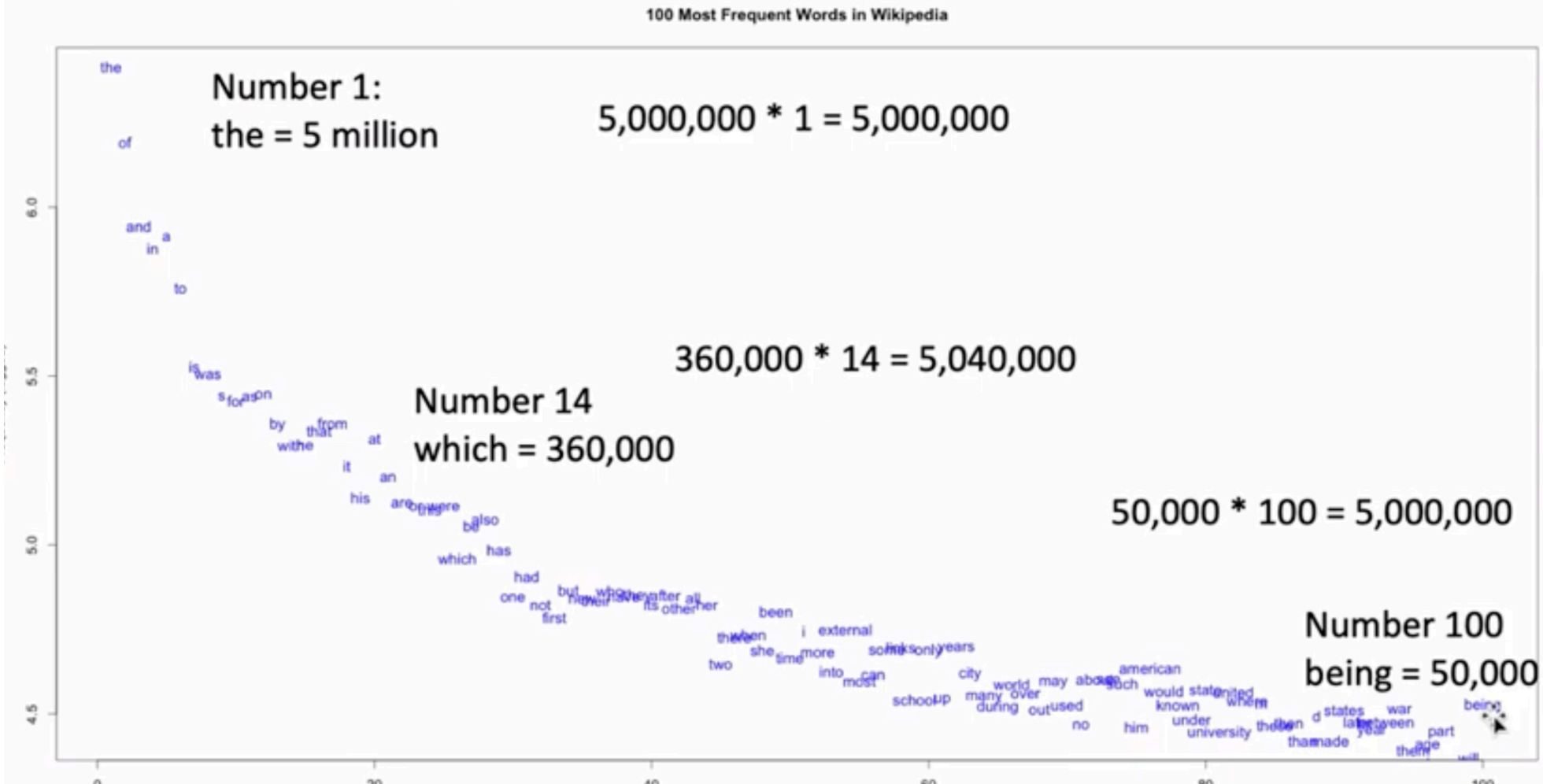
- The most frequent word in one corpus may be rare in another


100 Most Frequent Words in Wikipedia
Frequency (logged)
Frequency Rank
6.0
5.5
5.0
4.5
0
20
40
60
80
100
the
of
and
a
in
to
was
by
from
at
it
an
his
also
which
has
had
one
not
first
been
i
external
she
more
two
into
years
city
world
may
american
would
known
under
university
no
him
d
states
war
being
part
used
out
during
many
over


100 Most Frequent Words in Wikipedia
Number 1:
the = 5 million
Number 14
which = 360,000
Number 100
being = 50,000
Frequency (logged)
6.0
5.5
5.0
4.5
0
20
40
60
80
100

Zipf's Law

- The frequency of a word is inversely proportional to its frequency rank.



Zipf's law

- Some words are used very frequently:
  - grammatical words: the, of, and, to, a, in, …
- Many words are used very rarely
  - hapax legomena: words that occur only once in a corpus
- Frequent words are short, rare words tend to be longer.

20,80 % PRINCIPLE

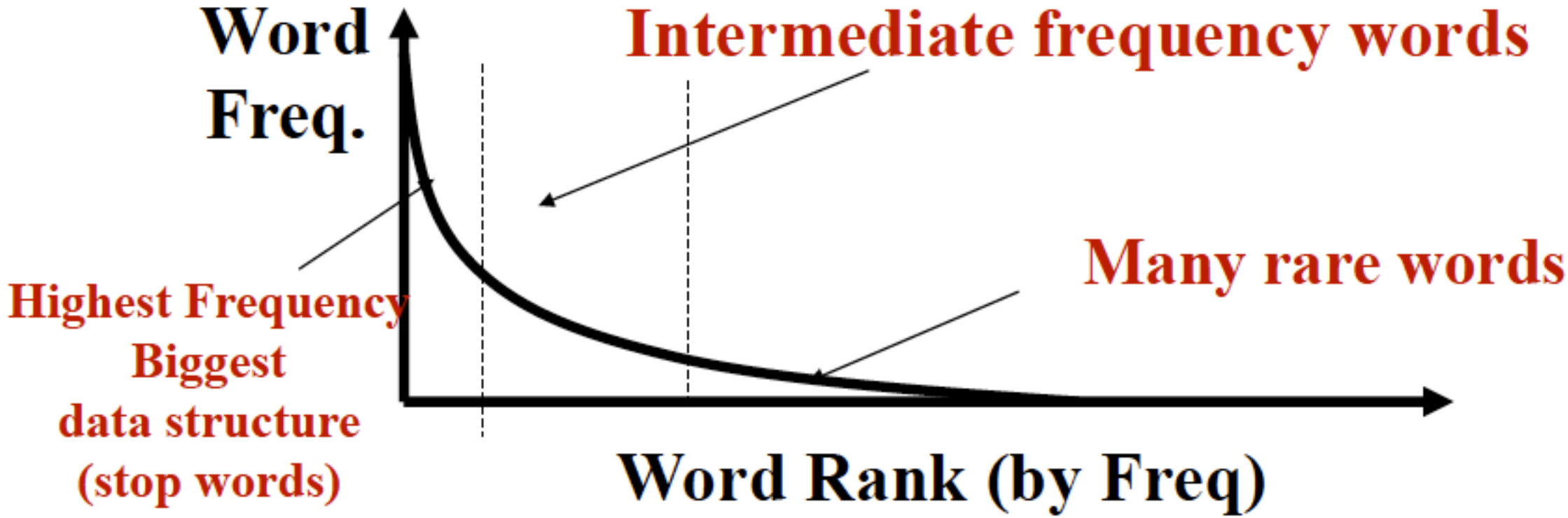
Why are word frequencies Zipfian?

- the principle of least effort:
  - If we have to use a word very often, it saves time and energy if that word is very short.
- communication efficiency and robustness:
  - Shorter words carry a higher risk of being misheard, so they do not carry a lot of information, but they convey this information efficiently.
  - Longer words carry more information and are more reliably identified, but at a higher processing cost.

Zipf's Law

- rank * frequency ≈ constant

$$F(w) = \frac{C}{r(w)^{\alpha}} \qquad \alpha \approx 1, C \approx 0.1$$



Tells you how many frequent words and rare words you are going to get into your collection of tex

Zipf's law

- Observation: frequent words and rare words
  - "of" and "the" make up 10% of all occurrences
  - hardly ever see "aardvark"
- Rank words by frequency
- Zipf's law:
  - rank of the word times its probability (frequency) is approximately a constant

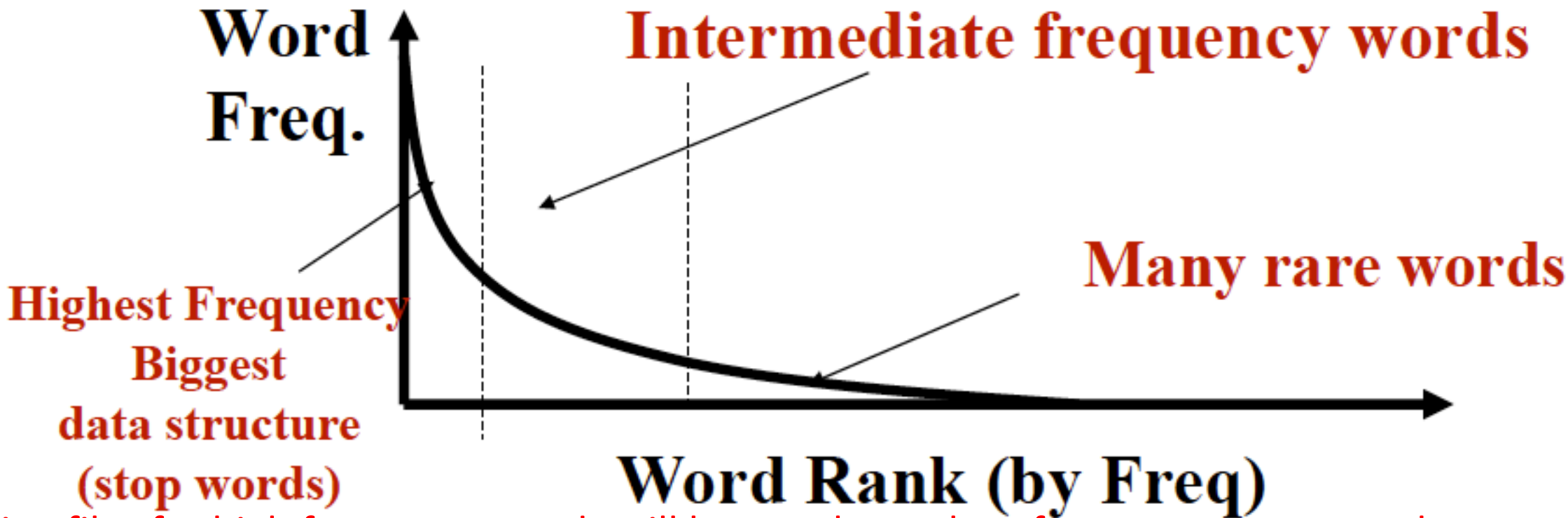
$$r \times P_r \approx const$$

Word	*Freq.*	*r*	P_r(%)	
the	2,420,778	1	6.49	6.49
of	1,045,733	2	2.80	5.6
to	968,882	3	2.60	7.8
a	892,429	4	2.39	
and	865,644	5	2.32	
in	847,825	6	2.27	
said	504,593	7	1.35	
for	363,865	8	0.98	7.8
that	347,072	9	0.93	

Zipf's Law

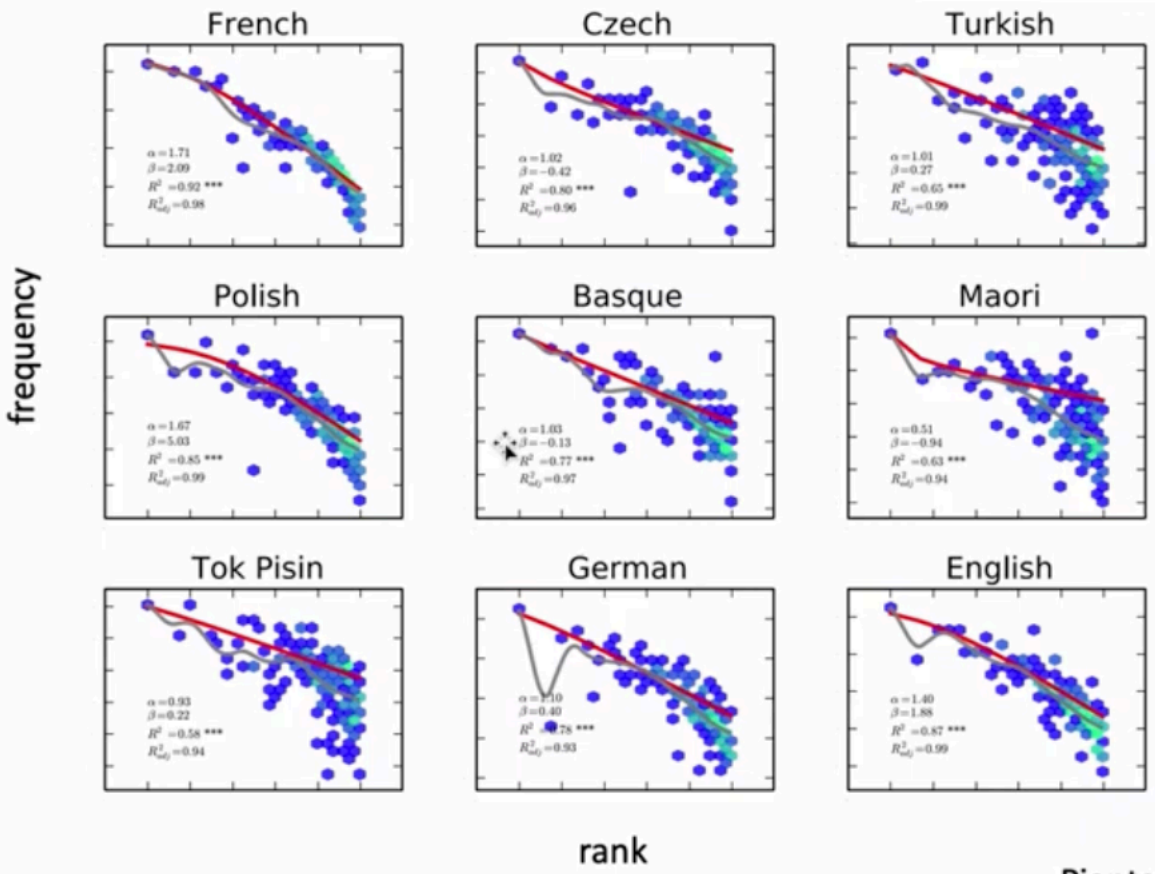
- rank * frequency ≈ constant

$$F(w) = \frac{C}{r(w)^{\alpha}} \qquad \alpha \approx 1, C \approx 0.1$$



Posting files for high frequency words will be very long, therefore we can remove these words and can save space.

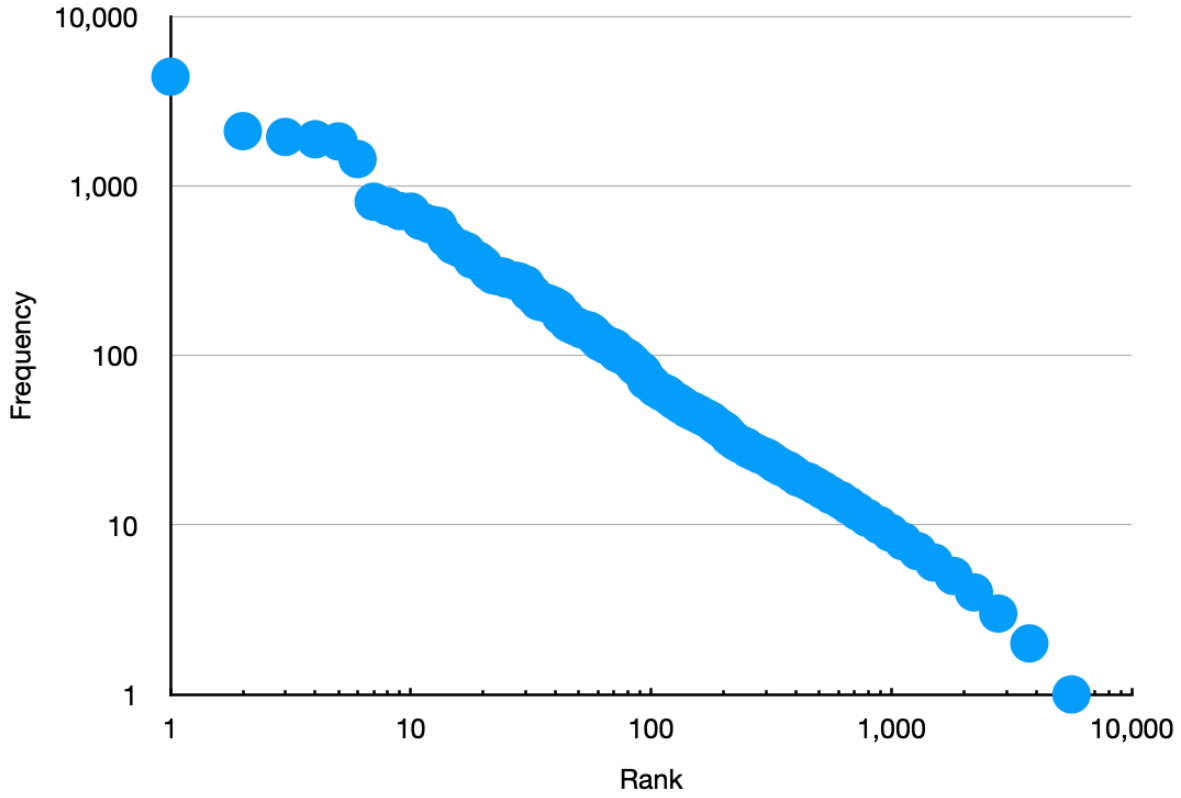
Zipf's law across languages



Piantadosi 2014

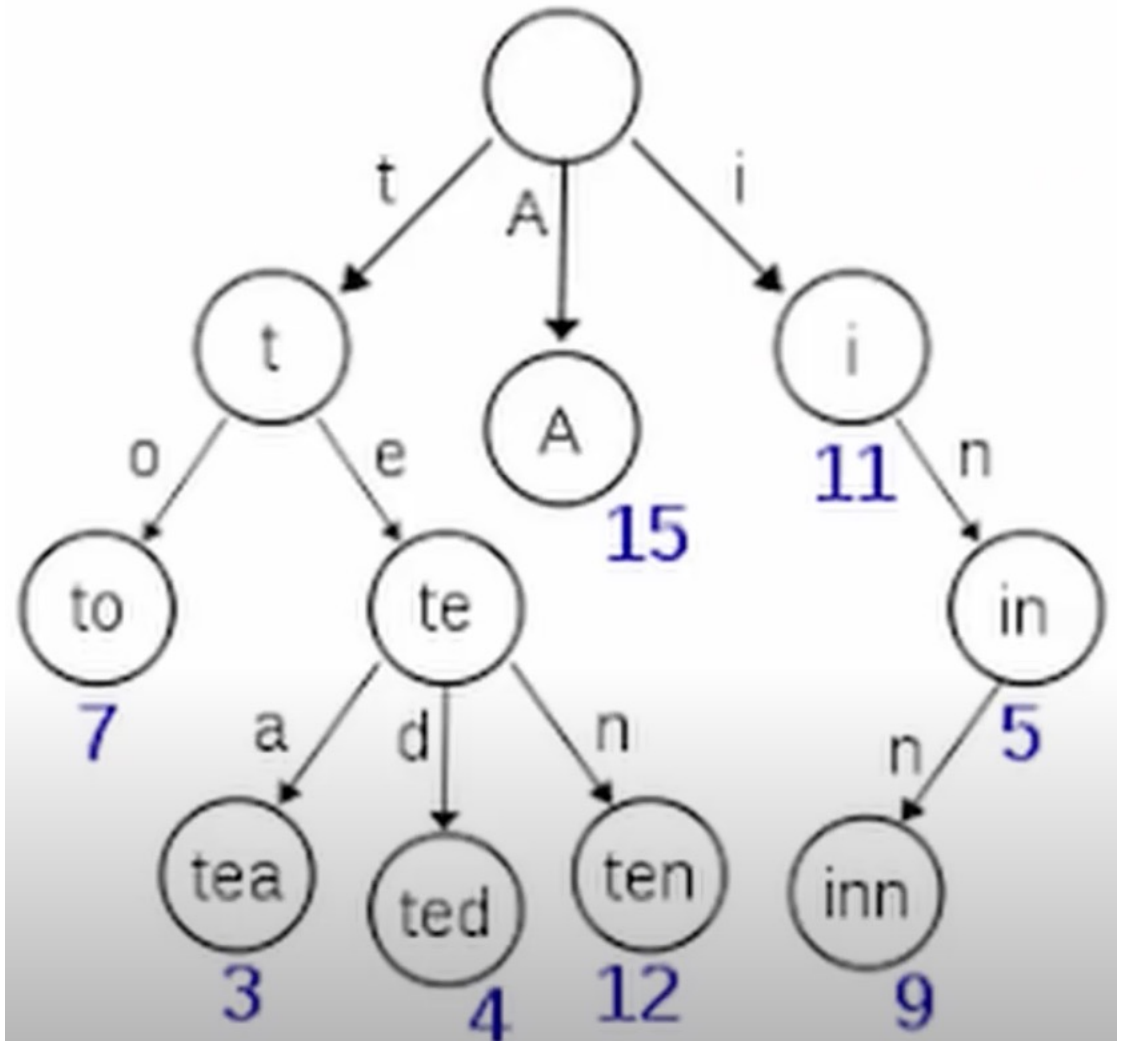
Demo

- https://www.laurenceanthony.net/software/antconc/



Data Structures for Inverted Index

- Dictionary: modest size
  - Needs fast random access
  - Preferred to be in memory
  - Hash table, B-tree, trie, …
- Postings: huge
  - Sequential access is expected
  - Can stay on disk
  - May contain docID, term freq., term pos, etc
  - Compression is desirable


t
A
i
t
A
i
o
e
11
n
15
to
te
in
7
a
d
n
n
5
tea
ted
ten
inn
3
4
12
9