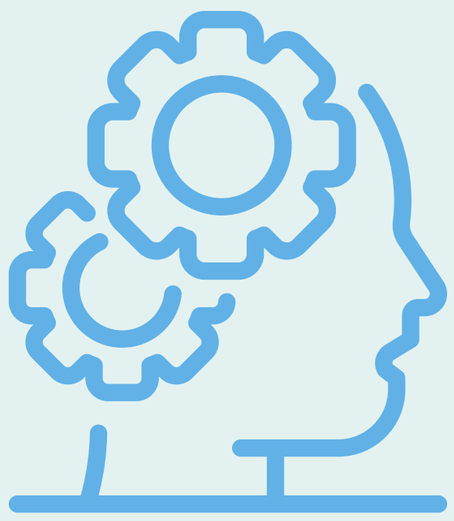
COUNT MIN SKETCH IN NLP

NLP: Natural language processing

subfield of Computer Science to process and analyze large amounts of natural language data.



Perform 2 Evaluation

intrinsic evaluations: show that low-frequent items are more prone to errors.
extrinsic evaluations: computing approximate PMI scores

Sketch is effective on large-scale language processing tasks . The Count-Min Sketch is a widely adopted structure for approximate event counting in large scale processing.

Count-Min sketch can be made duplicate-insensitive for exploiting multi-path routing in sensor networks.

Count-Min sketch to store the frequency of keys (words or word pairs) for NLP applications

NLP applications, tolerate either over-estimation or under-estimation errors.
breakdown the error into over-estimation (OE) and under-estimation (UE) errors

In NLP tasks we are interested by the low-frequency items. It solves three largescale NLP problems using small bounded memory footprint.

Count min sketch never under estimates so over estimation is considered.
Hence, to compute the over-estimation MRE, we take the average of positive values over all the items in each bucket.

Source

http://dimacs.rutgers.edu/~graham/pubs/papers/nlpsketch.pdf
https://arxiv.org/abs/1604.05492
https://aclanthology.org/W10-1503.pdf