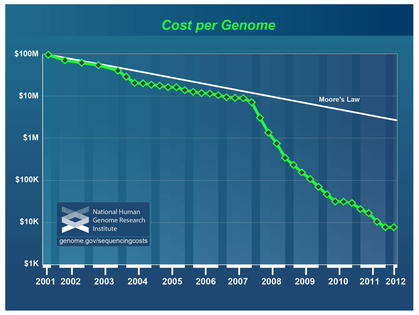
A search engine of genome sequence for papers

Heqiao Liu, Robert Edwards San Diego State University, Department of Computer Science

Background

Due to improvement of sequencing technology in recent years, the amount of genomes being sequenced is booming. At the same time, publishing papers through electronic media, especially on-line journals encourages more papers of studies about genome sequences published. The NCBI accession number and the Gi number are two major types of IDs, which are widely used to share and track DNA sequences record in three major sequence database (DDBJ/EMBL/GenBank).



[1]

Popular search engines, such as Google scholar, do well to search specified words among files on-line. However, as they are not designed to handle the specific type of search of sequence IDs, the results they returned can be irrelevant matters. At the same time, as accession numbers may update through time, a general search engine may give less results than expected. Thus, a search engine focus on this issue will be an additional useful tool for researchers in the field.

Challenges and Assumption

- **Challenges**
  - IDs in database may change through time, but IDs in a paper won't. There are solutions through NCBI web service.
  - As combinations of letters and numbers can represent variate of ID in different fields, how can the engine know such pattern represents the ID it is looking for.

Assumption

- There are certain patterns will notify the appearance of sequence IDs, in the text, near the location of the ID.

Methods

N – gram[2]

A sequence of n elements. Words in this case.

- Looking for 3-gram and 4-grams as potential patterns.
- Train the machine with positive sets of papers, to find out significant 3 or 4-grams.

Index-n ... (Index-3 Index-2 Index-1) IDxxx index+1 … index+n

Hamming Distance

How different two string of characters are.

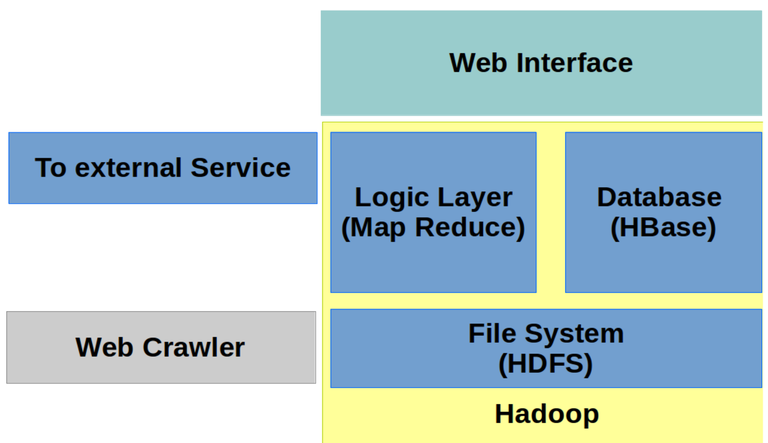
- If Two 3-grams have difference which is less than longer one's 1/3 length, this pair considered to be identical.

Parts-of-speech(POS) Tagging[3][4]

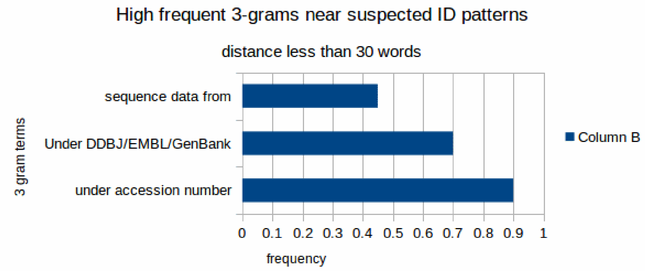
Identify what part of speech a word is.

- Noun (NN) appears near where possible ID shows up should be helpful to identify what it is.
- With a python library, NLTK, Nouns are retrieved, and high frequent nouns are considered to indicate ID near by as sequence ID.

Tech Stack



Results



Three terms above are high frequency terms, that they may transform a little in different papers, such as “under accession number”, “the accession number”, “under accession numbers” are all considered as the same group.

The POS approach mostly gathered “accession number” , and sequence. It matches 3-gram approach. However, it is notable, that with the results from two approach, it is not confident to say the suspicious ID pattern is sequence ID.

Discussion and Further work

Surrounding text of a possible ID pattern is unitary. The possible explanation of this fact is, that there is not different way to declare a sequence. The weaker assumption than the previous explanation, is that papers of the training set are edited by editors, or enforced with certain formats.

The size of training set and test set needs to be increased in the future. It is very likely single 3-gram or frequent nouns are not enough to judge a pattern. Within declare the range of the area of paper, it is likely a decision tree will be helpful in such task.

References

[1] http://www.genome.gov/sequencingcosts/

[2]Cavnar, W. B. (2009). N-gram-based text categorization.

[3]http://www.nltk.org/

[4]Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O’Reilly Media Inc.