Recitation 4

CS435: Introduction to Big Data

GTA: Bibek R. Shrestha
Email: cs435@cs.colostate.edu

Feb 09, 2018
Today...

- Introduction to Programming Assignment 1
N-gram

- A contiguous sequence of N items from a given sequence of text or speech

- Example:
  “We, analyze, large, datasets”

- 1-grams (aka unigram)
  We, analyze, large, datasets

- 2-grams (aka bigram)
  (__, We), (We, analyze), (analyze, large), (large, datasets), (datasets, __)
Dataset

- ~1.5 million Wikipedia articles (summarized)
- ~1 GB size
- Format:
  
  ```
  Title_of_Article-1<====>DocumentID<====>Text_of_Article-1
  NEWLINE
  NEWLINE
  Title_of_Article-2<====>DocumentID<====>Text_of_Article-2
  NEWLINE
  NEWLINE
  ```

- *Note: Text of an article starts with its title.*
Profile 1

- A complete list of unigrams in the dataset
- Must be sorted in alphabetical order (ascending)
- Eliminate duplicates.
- Use of combiner to eliminate local duplicates can be important (one of the approaches that can be taken).
Profile 2

- A list of unigrams and their frequencies within each article
- Similar to word-count program
- Each article has unique integral Document ID.
- Output should be grouped by Document ID.
- Output files will correspond to the number of reducer used.
Profile 3

- List of unigrams and their frequencies in the corpus
- List should be sorted from most frequent unigrams to least frequent ones.
- Solution to generate profile 3 can be seen as combination of profile 1 and 2.
- Here, we list unigram with its total occurrence in the complete dataset (of course, without considering per article construct as of profile 2).
Data Preprocessing

- Consider only alphabetic and numeric text.
- Convert upper cases to lower cases.
- Example (Original $\rightarrow$ processed $\rightarrow$ toLowerCase)

  - U.S.A. $\rightarrow$ USA $\rightarrow$ usa
  - west. $\rightarrow$ west $\rightarrow$ west
  - (USA) $\rightarrow$ USA $\rightarrow$ usa
  - D.C., $\rightarrow$ DC $\rightarrow$ dc
  - 9.8 $\rightarrow$ 98 $\rightarrow$ 98
Tips

- Before starting, design your system architecture.
- Make sure you understand Word-Count program.
- Start writing program by making modification in given Word-Count program.
- Test your program in standalone/local mode with given sample file.
- Verify your output in even smaller text file, if required, before moving on to larger files.
- Parsing (preprocessing) the dataset will be an important part and will require considerable effort. (This is generally true for any big data application.) Spend some time to observe the dataset provided.
In the next recitation...

- We will go through problems you are facing.
- We will try to demonstrate solutions to common problems asked in piazza and in-person.