**How long the program took to acquire the text characteristics?**

-> 0.9574556350708008 s

**How the program handles:**

**A. Upper and lower case words (e.g. "People", "people", "Apple", "apple")**

-> All the words are converted to lower case and then tokenized

**B. Words with dashes (e.g. "1996-97", "middle-class", "30-year", "tean-ager")**

-> Hyphen is removed from compound words so middle-class becomes middleclass

**C. Possessives (e.g. "sheriff's", "university's")**

-> Apostrophee is also removed from words so sheriff's becomes sherrifs

**D. Acronyms (e.g., "U.S.", "U.N.")**

-> Dot(.) from acronyms are removed so U.S. becomes us

-> Numbers and digits are also tokenized

**Briefly discuss your major algorithms and data structures.**

-> To save results of tokenization and stemming dictionaries are used so that the retrieval and insertion of data takes o(1)

-> Dictionary stores data as {string(token or stem):integer(frequency)}

Tokenization output:

* Number of Tokens: 132710
* Number of Unique words: 12272
* Number of words that occur only once: 6221
* 30 Frequent words:

flow: 1736

pressure: 1132

number: 964

boundary: 897

results: 885

mach: 816

theory: 775

layer: 728

method: 683

shock: 589

surface: 558

obtained: 539

given: 520

effects: 510

solution: 496

heat: 483

velocity: 481

temperature: 478

equations: 477

supersonic: 460

made: 449

ratio: 442

body: 439

wing: 430

presented: 425

found: 422

experimental: 421

laminar: 413

conditions: 411

effect: 402

* Average tokens per document: 94.79285714285714

Stemming output:

* Number of distinct stems: 10562
* Number of stems that occur only once: 5499
* 30 Frequent words:

flow: 1965

number: 1336

pressure: 1279

result: 1069

boundari: 926

effect: 917

method: 883

theori: 868

layer: 859

solution: 847

mach: 817

equation: 774

bodi: 740

wing: 710

present: 685

surface: 661

obtain: 632

shock: 614

distribution: 598

problem: 591

ratio: 588

temperature: 583

velociti: 554

case: 545

given: 520

heat: 518

test: 518

condition: 515

plate: 494

us: 479

* Average stemmed tokens per document: 94.79285714285714